USE OF THESES

This copy is supplied for purposes of private study and research only. Passages from the thesis may not be copied or closely paraphrased without the written consent of the author.
THE MANIFEST IMAGE

THE NATURE AND STATUS OF FOLK PSYCHOLOGY

Ian Martin Ravenscroft

September 1994

A thesis submitted for the degree of Doctor of Philosophy at the Australian National University
I declare that, unless otherwise indicated, this thesis is entirely my own original work.

Ian Ravenscroft
ACKNOWLEDGMENTS

I am very pleased to have this opportunity to thank the following people for their encouragement and support.

At the Research School Of Social Sciences, ANU, I benefited greatly from Professor Frank Jackson's thoughtful supervision and from his careful reading of the many drafts and papers from which this thesis emerged. Professor Philip Pettit kindly commented on Chapters 1-4 and helped in many other ways. In addition, I had the pleasure of discussing the early chapters with Dr Galen Strawson who visited the RSSS in the winter of 1993.

Part of this thesis was completed whilst I was a visiting Research Fellow at Rutgers University, New Jersey. Thanks are due to Professor Stephen Stich for his friendship, support, and collaboration.

Finally, without the love and encouragement of Tamara Zutlevics, this thesis would never have been completed. To her I give my heartfelt thanks.
ABSTRACT

This thesis is concerned with the nature and status of folk psychology. At the outset a distinction is made between the theory theory (internal) and the theory theory (external). The former posits a largely tacit, internally represented, theory of human psychology which facilitates the prediction and explanation of behaviour. The latter claims that our everyday talk about mental states implicitly constitutes a theory of the mind. Both the largely tacit, internally represented, theory of human psychology and the theory of mind implicit in our everyday talk about mental states have been labeled 'folk psychology'. To avoid confusion, I have called the theory posited by the theory theory (internal), folk psychology (internal), and the theory posited by the theory theory (external), folk psychology (external).

The theory theory (internal) is not the only existent theory of our capacity to predict behaviour. So-called off-line simulation theory also seeks to account for that capacity. In Chapter 2 I sketch off-line simulation theory and defend the theory theory (internal) against it.

In Chapters 3-5 the focus shifts to the nature and status of folk psychology (external). I defend a commonsense functionalist analysis of the states posited by folk psychology (external), and argue that Fodor's asymmetric dependency theory of content provides the correct account of the semantic properties of (external) folk psychological beliefs and desires. A variety of objections to functionalism exist in the literature. Chapter 5 is devoted to drawing the fangs of some common objections to functionalism, including the qualia problem and the difficulties raised by Ned Block and John Searle.

Chapters 6 and 7 are devoted to the eliminativism issue. Arguments on both sides of the debate are examined and largely found wanting. Most pro- and anti-eliminativist arguments have focussed on the posits of folk psychology (external). I briefly consider some of the issues surrounding eliminativism and folk psychology (internal). Chapter 7 is devoted to eliminativist concerns about intentional non-naturalism.

Finally, in Chapter 8 I consider whether folk psychology (external) might form the basis of a scientific investigation of the mind. Arguments to the effect that it will not are rejected, and an extended example of scientific research which rest heavily on folk psychology (external) is described.
# CONTENTS

*Acknowledgments*  
*Abstract*  

1. **A Tale Of Two Theories**  
1.1 Introduction  
1.2 Children's acquisition of the capacity to predict and explain behaviour  
1.3 Folk mechanics  
1.4 Modelling folk psychology (external)  
1.5 Is folk psychology (external) a theory?  
1.6 Overview  

2. **Off-line Simulation Theory**  
2.1 Introduction  
2.2 What is off-line simulation theory?  
2.3 Defending the theory theory (internal)  
2.4 Concluding remarks  

3. **Analyzing Belief**  
3.1 Introduction  
3.2 Philosophical behaviourism  
3.3 A normative approach  
3.4 Commonsense functionalism  

4. **The Problem Of Content**  
4.1 Introduction  
4.2 The resemblance theory  
4.3 The indicator theory  
4.4 The teleological theory  
4.5 Fodor's theory  
4.6 Concluding remarks
<table>
<thead>
<tr>
<th>5</th>
<th>Defending Functionalism</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>The Blockhead</td>
<td>98</td>
</tr>
<tr>
<td>5.3</td>
<td>The Chinese room</td>
<td>103</td>
</tr>
<tr>
<td>5.4</td>
<td>Troubles with qualia</td>
<td>108</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>Eliminativism</th>
<th>119</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>119</td>
</tr>
<tr>
<td>6.2</td>
<td>Arguing for eliminativism</td>
<td>122</td>
</tr>
<tr>
<td>6.3</td>
<td>Arguing against eliminativism</td>
<td>132</td>
</tr>
<tr>
<td>6.4</td>
<td>Eliminativism and folk psychology (internal)</td>
<td>136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7</th>
<th>Naturalism And Irrealism</th>
<th>139</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>139</td>
</tr>
<tr>
<td>7.2</td>
<td>Global supervenience</td>
<td>140</td>
</tr>
<tr>
<td>7.3</td>
<td>Arguing for irrealism</td>
<td>145</td>
</tr>
<tr>
<td>7.4</td>
<td>Concluding remarks</td>
<td>149</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8</th>
<th>The Scientific Image</th>
<th>151</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>151</td>
</tr>
<tr>
<td>8.2</td>
<td>Folk psychology (external) as science (I)</td>
<td>152</td>
</tr>
<tr>
<td>8.3</td>
<td>Folk psychology (external) as science (II)</td>
<td>157</td>
</tr>
<tr>
<td>8.4</td>
<td>Concluding remarks</td>
<td>165</td>
</tr>
</tbody>
</table>

Bibliography | 167 |
1.1 Introduction. We have a remarkable capacity to predict our own and others’ behaviour. This capacity is so commonplace as to escape notice, so let me dwell on it for a moment. I predict that when you have read the last word on this page you will turn the leaf; that if your phone were to ring now you would answer it; that many of my colleagues will gather downstairs for morning tea at about ten thirty; and that many more people will be found in the Calypso coffee lounge a little after one o’clock this afternoon.

The mundane subject matter of my predictions obscures their extraordinary power. I am an enormous, highly structured array of cells of diverse descriptions. Imagine trying to predict my lunchtime behaviour on the basis of my cellular make up. Indeed, I am a vast collection of subatomic particles of various kinds. Imagine trying to determine what I will do at lunchtime on that basis. From the perspectives of biology and physics our ability to predict each other’s behaviour is quite extraordinary, all the more so when we observe the wide range of circumstances – many of them novel – under which predictions are successfully made.

We also seek to explain human behaviour. We typically do so by attributing one or more of a wide range of mental states to the person or persons in question. The attributed states include beliefs and desires as well as emotions and sensations. Of course we cannot straightforwardly claim for our capacity to explain behaviour the success we claim for our capacity to predict it. Behaviour is typically publicly observable, so our success at predicting it is easily judged. The states attributed in everyday explanations of behaviour, on the other hand, are internal and hence not readily observed. The success of everyday explanations of behaviour is thus a moot point.

We also attribute beliefs and desires to one another in circumstances that are not prima facie cases of explanation. That is, we sometimes "interpret" ourselves and others as holding certain beliefs and desires. It is an open question whether the processes involved in interpretation differ from those involved in explanation. After all, both involve the attribution of mental states on the basis of information about the

---

1 This Section is based in part on Stich & Ravenscroft 1994.
person's history, behaviour, and environment. In what follows I will often, for the sake of brevity, use 'explanation' in a way that embraces interpretation.

How do we do it? How do we predict, and offer explanations of, behaviour? How do we interpret ourselves and others? One of the main themes of this work is that in making such predictions and offering such explanations, each of us deploys a largely tacit, internally represented, theory of human psychology. That claim was first put forward in the 1950's by the psychologist Fritz Heider, and has subsequently come to be known as the theory theory.² Explaining human cognitive capacities by positing internal "knowledge structures" (including theories) is the modus operandi of contemporary cognitive science. Thus, famously, Chomsky sought to explain our capacity to generate and interpret grammatical sentences by positing an internally represented grammar.³ More recently, the strategy of positing largely tacit, internally represented, knowledge structures has been applied to our capacity to recognize objects visually; to solve mathematical problems; and to predict the movements of middle-sized physical objects.⁴ It is hardly surprising that this strategy has been applied to the capacity to predict and explain behaviour.

The theory theory is aptly named. It seeks to account for our capacity to predict and explain behaviour by positing a further theory — a largely tacit, internally represented, theory of human behaviour. In keeping with contemporary usage I will call the posited theory folk psychology. The theory theory thus seeks to account for our capacity to predict and explain behaviour by positing folk psychology. The expression 'theory theory' has, though, another use. According to the alternative sense of 'theory theory', our everyday talk about mental states implicitly constitutes a theory of mind. This idea can be traced back to Wilfred Sellars and Paul Feyerabend, and is most commonly associated with the work of David Lewis.⁵ The theory implicit in our commonsense platitudes about mental states is also called 'folk psychology'. So we have two quite separate senses of 'theory theory' in play, as well as two quite separate senses of 'folk psychology'. According to one sense, the theory theory is an account of our capacity to predict and explain behaviour. Call that usage the 'theory theory (internal)'. According to the other sense, the theory theory is the claim that our everyday talk about mental states implicitly constitutes a theory of mind. Call that usage the 'theory theory (external)'. Nothing much hangs on this choice of nomenclature. I chose 'internal' because the theory theory (internal)

---

² See Heider 1958.
³ See Chomsky 1965.
⁵ See Sellars 1956; Feyerabend 1963; Lewis 1972. Note that none of these authors actually uses the term theory theory which was introduced by Adam Morton. See Morton 1980.
posits an *internally* represented theory of human psychology, and 'external' because commonsense platitudes are *external* to their utterers.

We can similarly distinguish between the two senses of 'folk psychology' in play. Let us call the theory of human psychology posited by the theory theory (internal), 'folk psychology (internal)', and the theory of mind posited by the theory theory (external), 'folk psychology (external)'. Thus, the theory theory (internal) seeks to account for our capacity to predict and explain behaviour by positing folk psychology (internal). The theory theory (external), on the other hand, claims that our everyday talk about mental states implicitly constitutes folk psychology (external).

What is the relationship between folk psychology (external) and folk psychology (internal)? We can begin to approach this question by thinking about a parallel issue in linguistics. A controversy exists about just what a grammar is supposed to be. On the one hand, a grammar might be regarded as nothing more than a systematization of speakers' linguistic judgements. Considerations of simplicity are appealed to in selecting between otherwise equally satisfactory systematizations. On this view there may be no unique grammar for a language. After all, it may turn out that a variety of equally simple systematizations of speakers' grammaticality judgements can be formulated. On the other hand, a grammar might be taken to be part of the cognitive apparatus which generates grammaticality judgements. On this view, a grammar is an internally represented theory of the language in question. Call the sense of 'grammar' according to which a grammar is a systematization of grammaticality judgements, 'grammar (external)', and call the sense of 'grammar' according to which a grammar is part of the cognitive apparatus which generates grammaticality judgements, 'grammar (internal)'. So our question is: What is the relationship between grammar (external) and grammar (internal)?

An obvious suggestion is that grammar (internal) just is grammar (external). But, as we have noted, there may be more than one equally satisfactory systematization of speakers' grammaticality judgements. With which one should we identify grammar (internal)? Moreover, we cannot safely assume that the internally represented theory of language which drives our grammaticality judgements is the (or one of the) simplest systematizations of our grammaticality judgements, that is, there is no reason to expect our cognitive systems to reflect the virtue of simplicity. As François Jacob has remarked, evolution is a tinkerer. Richard Dawkins offers

---

7 See Chomsky 1965; Chomsky & Katz 1974.
8 See Jacob 1977.
the following elegant example of unexpected complexity in an evolved system.⁹ The recurrent laryngeal nerve controls the movement of the larynx. Considerations of simplicity strongly suggest that the laryngeal nerve should not pass from the brain to the larynx via the aorta. But it does. And that is a long detour in a giraffe. Similarly, grammar (internal) may be decidedly unsimple. Speaking metaphorically, it too may detour past the aorta on its way to the larynx.

Finally, grammar (internal) may differ from grammar (external) in that whilst the latter is likely to take the form of a set of declarative sentences, the former may not. Grammar (internal) may well be just a list of procedures or rules – a "program" if you will. On this view, sentences are parsed, grammatical utterances generated, and grammaticality judgements produced, by executing a program. It might be argued that there could exist a unique, well-motivated, mapping of the program onto a set of declarative sentences, and thus that the question 'Is grammar (internal) identical to grammar (external)?', is still well-conceived. However, there does seem to exist a class of connectionist networks which do not admit of unique, well-motivated, mappings onto sets of declarative sentences.¹⁰ If the cognitive mechanism underlying our capacity to generate grammaticality judgements is properly modelled by such a network, the question 'Is grammar (internal) identical to grammar (external)?' will have no clear meaning.

Let us return to the question of the relationship between folk psychology (internal) and folk psychology (external). An obvious suggestion is that folk psychology (internal) just is folk psychology (external). There are, though, a variety of reasons for regarding this response as implausible. First, following the externalist grammarians, we might identify folk psychology (external) with the simplest systematization of our everyday talk about mental states. But once again there is no guarantee that there is a unique systematization of such talk. And if there is no unique systematization, folk psychology (internal) cannot be straightforwardly identified with folk psychology (external). Moreover, there is no reason to assume that folk psychology (internal) is the (or one of the) simplest systematizations of folk psychology (external). As we saw above, biological systems are not known for exhibiting the virtue of simplicity. Finally, again echoing the case of grammar, folk psychology (internal) may not take the form of a set of declarative sentences. It may be a set of rules or procedures – what I called in the case of grammar (internal) a 'program'. If there exists a unique and well-motivated mapping of the program onto a set of declarative sentences then the question 'Is folk psychology (internal) identical to folk psychology (external)?' may still be regarded as well-conceived.

¹⁰ I appeal to this class of networks cautiously as their existence is a moot point.
But if the cognitive mechanism underpinning the prediction of behaviour is a
collectionist network which does not permit a unique and well-motivated mapping
onto a set of declarative sentences, then there will be no clear sense in which folk
psychology (internal) and folk psychology (external) are or are not identical.

There is a further consideration which tells against the idea that folk
psychology (internal) is identical to folk psychology (external). It seems likely that
in making predictions of behaviour we at least sometimes rely on cues and
inferences of which we are not consciously aware. Perhaps, for example, when I
predict that S is about to order a second cappuccino, my prediction is based in part
upon my having unconsciously observed a certain glint in her eye. The example is
frivolous but the underlying idea not at all far-fetched. After all, when forming
judgements about the depth of my visual field, I rely on a number of cues of which I
am not generally conscious: texture gradient, stereopsis, and overlap, to name a
few. If it is true, as seems likely, that in forming predictions about each other's
behaviour we rely to some extent on unconscious cues, folk psychology (internal) is
unlikely to be identical to folk psychology (external). After all, if the relevant cues
are only recognized unconsciously, then they are unlikely to be referred to in the
assembled commonsense platitudes that implicitly constitute folk psychology
(external).

If the relation between folk psychology (external) and folk psychology
(internal) is not one of identity, what is it? I suggest that folk psychology (external)
is, to a large extent, causally dependent on folk psychology (internal), and thus that
folk psychology (external) is an important (though defeasible) source of information
about folk psychology (internal). Grammarians of an internalist bent are interested,
needless to say, in grammaticality judgements and systematizations of those
judgements. Such judgements and systematizations are a major source of
information about grammar (internal). They are not, though, the only source of
information. Internalist grammarians will also pay close attention to data about
children's acquisition of natural languages. Similarly, cognitive scientists interested
in accounting for our capacity to predict and explain people's behaviour will play
close attention to folk psychology (external). But other sources of information are
available to them, including studies of children's development of the ability to
predict and explain behaviour. I sketch the main outlines of this development in the
next Section.

This thesis is about folk psychology - in both senses. Determining the nature
of folk psychology - internal or external - is a major empirical challenge. At present
very little is known about folk psychology (external), and even less about folk

11 This example is drawn from Fodor 1981.
psychology (internal). As mentioned above, one important source of information about folk psychology (internal) which does presently exist is an extensive – and expanding – literature on children's acquisition of the ability to predict and explain behaviour. I review that literature briefly in Section 1.2. Then, in Section 1.3, I offer a brief detour into folk theories of mechanics. This is not merely sight-seeing. Studies of folk mechanics provides genuine insight into the nature of internally represented theories and the difficulties of modelling them. In Section 1.4 I review the available sources of evidence about the nature of folk psychology (external), and offer a tentative sketch of the theory of mind implicit in our everyday talk about mental states. Section 1.5 argues that folk psychology (external) is indeed worthy of the label 'theory', and a brief overview of the rest of the thesis is presented in Section 1.6.

1.2 Children's Acquisition Of The Capacity To Predict And Explain Behaviour. I remarked at the beginning of the previous Section that we have a remarkable capacity to predict and explain the behaviour of ourselves and others. When do children acquire this capacity and does it exhibit – as cognitive faculties so often do – a characteristic pattern of development? In this Section I will briefly review the literature on children's acquisition of the capacity to predict and explain behaviour. Note that, if the theory theory (internal) is true, then acquiring the capacity to predict and explain behaviour is tantamount to acquiring and learning to deploy folk psychology (internal). However, until more has been said in favour of the theory theory (internal), I prefer to talk about the child's acquisition of the capacity to predict and explain behaviour rather than the child's acquisition of folk psychology (internal).

Even two year olds show some capacity to predict behaviour. They can typically predict, for example, that if John wants a cookie and sees one in the cookie jar then he will reach for the cookie jar. Children of this age also make extensive use of (external) folk psychological terms for desiring and perceiving. They do not, though, use (external) folk psychological expressions for believing. Moreover, two year olds fail so-called 'standard belief tasks'. A typical standard belief task is as follows.

---

12 As the expression 'explain' is often used with "success grammar", I should perhaps say that we have a remarkable capacity to predict and putatively explain behaviour. I will, however, generally avoid this clumsy locution.

13 See Ch. 2 for a defense of the theory theory (internal).


15 See Bretherton & Beeghly 1982.
Sam wants to find his puppy. His puppy might be hiding in the garage or under the porch. But Sam thinks his puppy is under the porch. Where will Sam look for his puppy: in the garage or under the porch?\textsuperscript{16}

That is, two year old children cannot accurately predict behaviour on the basis of information about beliefs and desires.

By the age of three, (external) folk psychological expressions for believing and related processes – 'thinking', 'knowing', 'remembering', etc – have become common in the child's spontaneous speech.\textsuperscript{17} Moreover, three year olds can typically perform well on standard belief tasks. For example, 77\% of three year old subjects performed correctly on the belief task just given.\textsuperscript{18} This suggests that three year olds have at least some capacity for forming accurate predictions of behaviour on the basis of information about beliefs and desires. They have difficulty, however, forming accurate predictions about behaviour driven by beliefs which they are in a position to judge false. Thus, most three year olds fail so-called 'false beliefs tasks' typified by the following experiment.

The child is shown a scenario involving two dolls, Sally and Anne. The first doll, Sally, hides her marble in a box, then goes away for a walk. While she is away, Anne transfers Sally's marble to a basket and hides it there. Sally then returns wanting her marble. The child is asked some control questions to make sure he or she has followed the events. Then the child is asked, 'Where will Sally look for her marble?'\textsuperscript{19}

Obviously, the right answer is that Sally will look for her marble where she believes it to be – in the box – not where it actually is – in the basket. That is, Sally's behaviour is driven by a false belief about the location of her marble. Normal three year olds, however, typically fail this task, predicting that Sally will look where her marble actually is.\textsuperscript{20} That is, it seems that three year olds have trouble predicting the behaviour of people whose beliefs they are in a position to judge false. Normal four year olds, however, typically have no difficulty predicting where Sally will look for her marble; that is, they can readily predict the behaviour of people with false beliefs.\textsuperscript{21}

\textsuperscript{16} Wellman 1990: 64.
\textsuperscript{17} See Shatz et al 1983.
\textsuperscript{18} See Wellman 1990: 64.
\textsuperscript{19} Leslie 1988: 19.
\textsuperscript{20} See Perner et al 1987.
\textsuperscript{21} See Wimmer & Perner 1983; Baron-Cohen et al 1985.
So far we have focussed on children's ability to form accurate predictions of behaviour on the basis of beliefs, desires, and perceptions, and we have seen that three year olds have at least a limited grasp of these states. In addition, three year olds also seem to have some understanding of dreams and images. They do not, though, avail themselves of such states when predicting and explaining behaviour.22 Four and five year olds, in contrast, seem not only to understand dreams and images, but also to base behavioural predictions and explanations upon them.

Finally, what of affect? Can children predict the sorts of emotions people are likely to experience under various circumstances? Wellman has explored four year olds' abilities to predict whether a person would be happy or unhappy if they did or did not get what they wanted.23 The subjects were read two simple stories. In both stories a character wants something (say orange juice), but only in one story does the character get what she wants. After each story the subjects were asked whether or not the character was happy. 94% of four year olds predicted that someone who wants orange juice and gets it will be happy, whilst someone who wants orange juice and does not get it will be unhappy.

In sum, by the age of four or five, children have become quite expert at predicting and explaining behaviour, and at attributing a variety of emotional states to others. At earlier ages children have some capacity to predict and explain behaviour, but that capacity is limited. Very young children (around two) can merely predict that people will move towards objects that they desire and can perceive. Three year olds gain the capacity to handle straightforward belief tasks, but are incapable of dealing with situation that involve false belief. Finally, four and five year olds can make accurate predictions in cases which involve either false beliefs or dreams and images. They can also predict the emotions people will experience under a range of circumstances.

There is an obvious analogy to be drawn between children's acquisition of the capacity to predict and explain behaviour and their acquisition of the capacity to speak and understand a natural language. The latter capacity is widely taken to be largely innate, with the linguistic experience of the child setting various pre-existing parameters.24 Three arguments are typically advanced in favour of the innateness claim in the case of natural languages.25 (1) The child's exposure to natural language is too limited to allow the child to acquire a natural language by generalizing from the available cases. (2) There apparently exists a pan-cultural 'universal grammar' common to all natural languages. The existence of such a grammar is readily

---

22 See Wellman & Estes 1986.
24 See Chomsky 1965.
accounted for by the advocate of the innateness hypothesis but is a considerable embarrassment to her opponent. (3) A social species whose members pretty much automatically acquired the capacity for detailed communication would be at a considerable selective advantage. We therefore have an evolutionary argument in favour of innateness.

It is tempting to offer a parallel series of arguments in the case of the capacity to predict and explain behaviour. (1') If the capacity to predict and explain behaviour is not largely innate, then presumably it is learnt. Obviously it is not explicitly taught, and therefore we must conclude that if the capacity is not innate then it is developed from exposure to behaviour. But the child’s exposure to behaviour is too limited to allow the child to acquire the capacity to predict and explain behaviour by generalizing from the available cases. (2') It seems that the capacity to predict and explain behaviour is pan-cultural.26 Once again, this is readily explained by the advocate of the innateness hypothesis but is an embarrassment to her opponent. (3') A social species whose members pretty much automatically acquired the capacity to predict and explain each other's behaviour would be at a considerable selective advantage. We therefore have an evolutionary argument in favour of innateness.

It is very difficult to assess the strengths of arguments (1') to (3'). Argument (1') rests on the claim that the child’s exposure to behaviour is too limited to determine the capacity to predict and explain behaviour. However, without an estimate of how much behaviour the child actually is exposed to, together with an estimate of how much behaviour the child would have to be exposed to, we cannot determine the plausibility of argument (1'). Argument (2') rests on the premiss that the capacity to predict and explain behaviour is pan-cultural. There is some evidence to support that claim, but a lot more is required. Finally, argument (3') appeals to the selective advantage of an innate capacity to predict and explain behaviour. Such arguments must, though, be treated with considerable caution. We must consider the cost of having such an innate capacity, not just the benefits, yet no estimate of costs is on offer.

In balance, whilst the innateness hypothesis is important, as yet we lack sufficient evidence to support it and thus it remains largely speculative. In the next Section I turn to our capacity to predict and explain the movements of middle-sized physical objects.

---

26 Most research on children’s acquisition of the capacity to predict and explain behaviour has focussed on Western children (Austria, Britain, Canada, and the United States). But see Alvis and Harris’ study of Baka children (Avis & Harris 1991).
1.3 Folk Mechanics. In our daily activities we interact almost ceaselessly with middle-sized physical objects. Often these objects are moving, and we show a remarkable facility to predict their movements and to offer explanations of why they moved thus and so. Cognitive scientists have sought to account for our capacity to predict and explain the movements of middle-sized physical objects by positing an internally represented theory – folk mechanics. Let me outline a few typical experiments.

McCloskey and his co-workers presented 135 college students taking an introductory psychology course with a diagram like the one in figure 1.1, below. The subjects were then given the following instructions.

The diagram shows a side view of a cliff. The top of the cliff is frictionless (in other words, perfectly smooth). A metal ball is sliding along the top of the cliff at a constant speed of 50 miles per hour. Draw the path the ball follows after it goes over the edge of the cliff. Ignore air resistance.27

![figure 1.1]

The correct response is given in figure 1.2, below. Many subjects, however, responded incorrectly. The most common incorrect responses are given in figures 1.3 and 1.4, also below.

What is intriguing is that the incorrect responses can be explained if we attribute to the subjects an impetus theory similar to that advanced by Philoponus (circa 6th century) and developed by Jean Buridan (circa 14th century).28 Impetus theory was developed in response to the following puzzle: What keeps an object...

---

28 For a discussion of impetus mechanics see Dijksterhuis 1961. Champagne et al 1980 have argued that folk mechanics is Aristotelian rather than medieval. For a rebuttal see McCloskey 1983: 318.
moving after the motivating force has ceased to act upon it? A ball, for example, continues to move for sometime after it has left my hand. Impetus theory responds to this puzzle by postulating a force, *impetus*, which is impressed upon the object by the motivating force, and which continues to propel the object after the motivating force has ceased. Different versions of the theory disagreed about the fate of the impetus force. All agreed that it dissipated due to such factors as air resistance, but some thought that it would dissipate eventually even in the absence of air resistance.

Consider again figure 1.3. In that figure the ball, once it has gone over the cliff, follows a curved path for some distance and then falls straight down. This is the path predicted by a simple impetus theory. According to such a theory, at the point the ball goes over the edge it has two forces acting upon it: its own impetus directed horizontally to the right, and gravity directed straight down. Under the combined influence of these two forces the ball follows a curved path right and down. The right-directed impetus force, though, dissipates over time, eventually becoming zero. At that point the ball falls straight down.

Note that the subjects were explicitly told to ignore air resistance. Even so, the impetus of the ball directed to the right dissipates according to figure 1.3. That suggests that the impetus theory held by many subjects postulates the dissipation of impetus over time even in the absence of such factors as air resistance; that is, that impetus has, as it were, a natural decay rate.
What about figure 1.4? In that diagram the ball is depicted as continuing horizontally and to the right for some distance before following a curving path downwards. After following the curving path for a while, the ball falls straight down. Such a trajectory is explained by a variant on the simple impetus theory sketched above. If we assume that gravity cannot act upon the ball until its horizontal impetus has dropped below a certain point, we obtain the result sketched in figure 1.4.

The special version of the impetus theory which predicts the trajectory sketched in figure 1.4 admits of a limiting case: gravity cannot act upon the ball until its horizontal impetus has fallen to zero. The limiting case theory predicts that the ball will have the trajectory given in figure 1.5, below. One of McCloskey's subject offered just such a response.

![figure 1.5](image)

Buridan and other impetus theorists distinguished between pushing or pulling on the one hand and being carried on the other. It is only the former, they thought, that impressed an impetus upon an object; an object that is merely being *carried* acquires no impetus. In an intriguing experiment McCloskey and his co-workers demonstrated that at least some subjects held a similar view. The subjects were presented with a diagram similar to figure 1.6, below. Figure 1.6 again depicts a cliff viewed side-on. This time the metal ball is propelled across the frictionless upper surface by a conveyor belt. The ball is held to the conveyor belt by an electromagnet. When the ball reaches the edge of the cliff the electromagnet is switched off, releasing the ball. The conveyor belt continues to move to the right at the same speed as before. What happens to the ball?

---

Some subjects drew a trajectory for the ball which was essentially the same as figure 1.3, above. Other subjects, however, predicted that the ball would fall straight down immediately after release. This suggests that the latter subjects drew a fundamental distinction between being pushed or pulled and being carried. Objects that are carried gain no impetus and thus fall straight down upon release. Similar results were obtained when subjects were asked to predict the trajectory of a bomb that had been 'carried' by a plane.30

Let me sum up. Our everyday capacity to predict and explain the movements of middle-sized material objects is to be accounted for by positing a tacit theory of mechanics. I have called that theory 'folk mechanics' and I have identified it with the impetus theory of Buridan and others. According to folk mechanics, objects continue to move after a motivating force has ceased because the motivating force impresses upon them an impetus. The studies discussed suggest that folk mechanics is committed to the impetus force dissipating even in the absence of air resistance and similar factors. They also suggest that a number of competing refinements and elaborations of the impetus theory exist within the general category of folk mechanics. These refinements and elaborations can be distinguished along at least two dimensions. On the one hand, subjects diverged about how gravity affects an object moving under the influence of impetus. And on the other, subjects diverged as to whether an impetus can be impressed upon an object by being carried.

These results are very interesting and highly suggestive. But note the difficulties involved in interpreting them. *Exactly what* is internally represented? A set of declarative sentences of the sort that a modern physicist might write down if she was trying to state, say, Buridan's mechanics? A set of declarative sentences plus

---

a set of rules which facilitate using the former under typical conditions? Just a set of rules or procedures? Clearly, the data provided by McCloskey and others is compatible with each of these options, and it is not obvious how these competing hypotheses are to be sorted out.

1.4 Modelling Folk Psychology (External). According to the theory theory (external), our everyday talk about mental states implicitly constitutes a theory of the mind – folk psychology (external). In what follows I offer a model of folk psychology (external). Obviously, modelling folk psychology (external) is an empirical task. However, as stressed in Section 1.1, very little careful empirical research has been undertaken in this area. Indeed, the available sources of evidence are embarrassingly scant. It follows that the model I offer below is necessarily somewhat limited. I begin by reviewing the available sources of evidence.

First source. We can take ourselves as working folk psychologists and consider our own talk about mental states. This strategy is, needless to say, fairly limited. There are obvious problems associated with being both subject and investigator. In particular, there is a real risk of investigating one's own beliefs about folk psychology (external) rather than investigating folk psychology (external) itself. In spite of this obvious difficulty, we often have little choice but to rely on our intuitions about what we would say in certain circumstances, for, as we shall see, there is a considerable paucity of other evidence.

Second source. The risk of investigating beliefs about, or theories of, folk psychology (external), rather than investigating folk psychology (external) itself, is also present in the second source of data – explicit questioning. Roy D'Andrade explicitly questioned a number of subjects about (external) folk psychological states and the relationships between them. Below is a small sample of D'Andrade's questions (page references are to D'Andrade 1987). They make it obvious, I think, that D'Andrade inadvertently encouraged his subjects to generate and consult a model of folk psychology (external), and yet we have no reason at all to accept the accuracy of their modelling. (D'Andrade's approach can be compared with attempting to develop a grammar of some natural language, not by eliciting subjects' grammaticality judgements, but by having them consult the grammar they learnt at school.)

(1) Could you think something was true, believe it, but not know that you believed it? (p. 128)

---

31 See D'Andrade 1987.
(2) When someone does something, do they usually have an intention in mind? (p. 131)

(3) Can one just do something for no reason at all — nothing intended or wanted? (p. 133)

(4) Sometimes someone says they did not know something at first, but then they figured it out. What do they do when they 'figure out' something? (p.135)

(5) How come people have the ability to control themselves, at least some of the time? (p.136)

So much for sources of evidence. As remarked earlier, they are embarrassingly scant. I will make two more prefatory remarks before getting down to business. First, let me emphasize that our task here is purely descriptive. At this point our aim is to describe folk psychology (external), not theorize about or extend it. Thus it is no use drawing on contemporary theories of cognition. We want to model the (external) folk psychological conception of the mind, not the conception found in, say, computational psychology. Second, much traditional philosophy of mind has taken the (external) folk conception pretty much for granted — taken it, as it were, as an a priori given — and then sought to theorize about (external) folk psychological states in various ways. Think, example, of the vast philosophical literature on akrasia. Whatever its merits, such philosophical theorizing is irrelevant to the task at hand. At this point we want to describe folk psychology, not theorize about its posits.

The (external) folk psychological conception of the mind is essentially Humean. Hume recognized that much of our behaviour is caused by the interplay of our beliefs and desires. As is standard, I will use the label action for that subset of behaviour which is (putatively) explained by folk psychology (external) in terms of beliefs and desires. Note, though, that folk psychology (external) at least at times concerns itself with behaviour which is not action. Thus folk psychology (external) offers a truism to the effect that if you touch a hot stove you will cry out (ceteris paribus), but it does not seek to explain such crying out as due to beliefs and desires. The set of actions and the set of behaviours of interest to folk psychology (external) are therefore not co-extensive. Nevertheless, folk psychology (external) is overwhelmingly concerned with action.

---

32 See Hume 1739, especially Book II, Part III, Section III.
Folk psychology (external), then, posits internal states—principally but not exclusively beliefs and desires—which are causally responsible for a certain class of behaviour. Folk psychology (external) is, moreover, thoroughly realist about beliefs and desires. They are the real causes of action; they are not postulates of a merely instrumental calculus. Beliefs and desires are, moreover, semantically evaluable. Thus beliefs are either true or false; desires satisfied or not. Beliefs and desires have semantic properties because in some sense they are representations, or relations to representations. This is reflected in our usual way of expressing beliefs and desires. Typically, we express beliefs and desires using sentences involving that clauses: S believes that p; S desires that q. I will have more to say in Chapters 3 and 4 about the (external) folk psychological concept of belief. For the moment I will simply remark, in a rather question begging fashion, that beliefs represent how the agent takes the world to be; desires how she wants the world to be.

I take the Humean conception of action to be the core of folk psychology (external). Notice its predictive and explanatory power. I predict that someone who wants a cappuccino and believes that they are available at the Calypso coffee lounge will proceed there. Alternatively, I explain someone's proceeding to the Calypso coffee lounge by citing their desire for a cappuccino and their belief that they are available there. Of course, these predictions and explanations are plausible only if we append suitable ceteris paribus clauses. Someone who desires a cappuccino and believes that they are available at the Calypso coffee lounge will proceed there only if they have the time, if they can afford the price of a coffee, if they prefer the Calypso over alternative venues, and so forth.

So far I have relied entirely upon the (external) folk psychological expressions 'belief' and 'desire'. A large number of synonyms and near-synonyms of these terms exist. I will not pause to consider the exact relations between the various synonyms and near-synonyms of 'belief' and 'desire'. There exist, moreover, a large number of folk psychological expressions for mental states which are not obviously synonyms or near-synonyms of 'belief' or 'desire'. To what extent these expressions can be identified with elaborate concatenations of beliefs and desires is an open question.

Note also that we often shun 'belief', 'desire', and their synonyms and near-synonyms altogether, preferring to "intentionalize" a non-psychological verb. This is particularly the case when offering explanations of behaviour. From my office window I see a woman walking across the carpark. I explain her action by saying that she is going to the library. 'Going' is not standardly a psychological verb, but in this case it has been intentionalized. It conveys that the woman is an intentional agent.

33 Famously, Dennett has offered an instrumentalist intentional psychology (Dennett 1981, 1987). But, as we shall see in Section 3.3, Dennett's conception of beliefs and desires does not map well onto the (external) folk psychological conception of those states.
who wants to go to the library and who believes that she can do so by crossing the
carpark.

A further important point to note about (external) folk psychological beliefs
and desires is that they need not be conscious. Folk psychology (external) recognizes
the existence of unconscious beliefs and desires - beliefs and desires which we cannot
'bring to the surface', or can do so only with difficulty. It follows that the theory
theory (internal) is (external) folk psychologically plausible. Folk psychology
(external) is comfortable with the idea that we possess bodies of largely tacit
knowledge. It could have been otherwise: there is no guarantee that the theory
theory (internal) be compatible with folk psychology (external).

As a model of folk psychology (external), the Humean conception is rather
thin. We can begin to fill in some of the details by considering the sources of belief
and desire recognized by folk psychology (external). Obviously, folk psychology
(external) recognizes that beliefs arise from perception: I see a black cat before me
and I come to believe that there is a black cat before me. We also monitor our own
internal environment: I feel hungry and come to believe that I am hungry. Further,
folk psychology (external) recognizes that we obtain new beliefs from old ones by
processes of reasoning. But folk psychology (external) is under no illusions about the
quality of our reasoning processes. It recognizes that we are not always rational - that
sometimes we reason very poorly indeed. As we will see in Chapter 3, this is an
important feature of folk psychology (external).

Where do desires come from? It is plausible that mental state talk varies at
least a little from person to person, and thus that the theory of mind implicit in my
mental state talk is slightly at odds with the theory implicit in yours. Indeed, each
person's mental state talk may constitute not a single folk theory of the mind but a set
of overlapping and to some extent competing theories. It follows that some questions
about (external) folk psychological states may not have straightforward and
unambiguous answers. These considerations apply to the question just posed: Where
do desires come from? Obviously, desires are in some sense related to perception, but
folk psychology (external) is undecided between two possible ways in which this
might be so. Say that I see a piece of chocolate cake. Uncontroversially, seeing the
cake before me causes me to believe, ceteris paribus, that there is a cake before me.
Also uncontroversially, sometimes when I see a piece of cake I want it. But does
seeing the cake directly cause me to want a piece of cake, or does the desire arise only
indirectly via my belief that there is a piece of cake before me? Folk psychology
(external) has no unambiguous answer to this question.

Note also that folk psychology (external) recognizes the existence of purely
instrumental desires. Say that I want a cappuccino but I realize that I do not have
enough money on me to buy one. I therefore form the instrumental desire of going to
the bank in order to get the money to buy a cappuccino, thereby fulfilling my initial desire. It is tempting to push such chains of instrumental desires back a long way. Thus it is tempting to see my desire for a cappuccino as arising from a desire to satisfy my craving for caffeine, and my desire to satisfy my craving for caffeine can in turn be seen as due to my desire to be comfortable. Folk psychology (external), though, is unwilling to take this process very far: the claim that I went for a cappuccino because I desired to be comfortable sounds more than a little strained. So-called 'drive reduction theory' – the claim that human behaviour is best understood in terms of the reduction or appeasement of various deep-seated drives – is therefore at least partly at odds with folk psychology (external). Folk psychology (external) is happy to explain my desire to go to the bank by citing my desire for a cappuccino. It does not, though, posit deep-seated drives for comfort, pleasure, or what-have-you.

So far we have seen two sources of desires. According to folk psychology (external), sometimes we want things simply because we have seen them, and sometimes we want things as means to obtaining something else. But not all desires can be thus accounted for. Sometimes we simply want things or, less colloquially, we will them. The will is visualized as an inner and deeply significant self with its own mental states. In particular, the posited homunculus has its own desires. On this view, when I say that I 'just want' something, I am expressing the desires of my inner homunculus.

We have seen a variety of sources of beliefs and desires. They commit folk psychology to positing the existence of three functionally-defined modules: a reasoner, a perceptual apparatus, and the will. The cognitive architecture we have obtained so far is sketched below in figure 1.8. Beliefs and desires are "stored" in "boxes". The perceptual apparatus and the reasoner are represented by circles. Arrows mark the flow of mental states from boxes to circles and vice versa. I have included a representation of the will as a processor which generates a special class of desires. Recall that earlier I distinguished action as that subset of behaviour (putatively) caused by beliefs and desires. Folk psychology (external) allows that some behaviour is not action. Figure 1.8 includes a representation of a causal pathway linking perception to behaviour but not proceeding via belief and desire.

A great deal has been left out of this model of the (external) folk psychological account of the production of behaviour. We have not considered, for example, if folk psychology (external) allows that beliefs and desires can directly cause action rather than via the reasoner. Nor have we considered the role of plans in the causation of action, and the difference between planning and intending.

---

34 Drive reduction theory originated in the work of C.L. Hull. See Hull 1943.
35 Wellman has observed the tendency of folk psychology (external) to posit an homunculus with its own mental states. See Wellman 1990: 269-270.
Moreover, I have said nothing at all about modelling what we might call the affective aspects of folk psychology (external). I will say just a little about affect now.

The affective side of folk psychology (external) posits a large range of states loosely known as 'feelings' or 'emotions'. How do such states fit into the model of folk psychology (external) developed thus far and sketched in figure 1.8? One approach to the emotions is to see them as dispositional properties of the cognitive system as a whole. Thus Wellman describes emotions as 'coloring' our thoughts.36 Similarly, Aaron Sloman has remarked that 'we need posit no special subsystem to account for emotions since mechanisms underlying intelligence suffice'.37 In this sense emotions are similar to character traits. The character trait of intelligence, for example, is the disposition to correctly and easily infer new beliefs from old ones; selfishness the disposition to form self-serving desires; vanity the disposition to readily believe that you possess in large measure various desirable properties. What distinguishes emotions such as cheerfulness and anger from character traits such as

---

36 See Wellman 1990: 112-113.
intelligence and vanity is the relative durability of character traits. A character trait is a dispositional property of a person's cognitive system which endures for a considerable period of a person's life. An emotion is far more transient, typically only lasting a few minutes or hours.

What is the source of emotions such as cheerfulness and anger? At least to some extent they arise due to discrepancies between the world and our beliefs and desires. Surprise is an emotion we experience when the world turns out not to be as we believed it to be; happiness an emotion we experience when the world turns out to be as we desired it to be. As we have seen, children as young as four can correctly predict that someone who wants orange juice and gets it will be happy, whilst someone who wants orange juice but does not get it will be unhappy.\(^{38}\) So it seems that quite young children are familiar with at least some aspects of the (external) folk psychology of emotion.

What we have elaborated so far is a sketch of what might be called 'external' folk psychological cognitive architecture: we have drawn up a rough plan of the mind as conceived by folk psychology (external). But folk psychology (external) posits more than a cognitive architecture. It offers a vast number of truisms connecting environmental circumstances to mental states, mental states to other mental states, and mental states to behaviour. Such platitudes can be seen as filling out the boxes, circles, and arrows of figure 1.8. Thus folk psychology (external) holds that if you want a coke and believe that you can get one from the machine downstairs then, \textit{ceteris paribus}, you will go downstairs. This truism provides details about the causal relations between beliefs and desires – details missing from figure 1.8. In particular, it tells us that a certain belief/desire pair (believing that there is a coke machine downstairs and desiring a coke) yields, \textit{ceteris paribus}, a certain action (going downstairs).

The model of folk psychology (external) just sketched leaves many details unspecified. Nevertheless, enough has been said for the time being. In Chapters 3 and 4 I undertake a much more detailed investigation of the (external) folk psychological concept of belief. In the next Section I rebut the claim that folk psychology (external) is unworthy of the label 'theory'.

1.5 Is Folk Psychology (External) A Theory? The theory theory (external) claims that our everyday talk about mental states implicitly constitutes a theory of the mind. It is therefore open to attack on the grounds that folk psychology (external) does not deserve the label 'theory'.\(^{39}\) An obvious way to press such an attack is to take

\(^{38}\) See Section 1.2, above.

\(^{39}\) Adam Morton contrasts theories with what he calls \textit{schemes} and insists that folk psychology (external) be assimilated to the latter rather than the former (Morton 1980, Ch. 1). He is, though,
scientific theories as paradigmatic of theoryhood, and then demonstrate that folk
psychology (external) is removed from the paradigm. I will consider that argument
strategy later. For the moment let me try to get a little clearer on what we might
mean by 'theory'.

The literature analyzing the notion of theoryhood is vast. We can locate
within it three major approaches. (1) A syntactic account favoured by the logical
positivists. On this view, a theory is a set of sentence identified as the theorems of
an axiomatic system. (2) A semantic account given in terms of an analogy with
Tarskian model theory. (3) An historical approach which stresses the importance of
scientific paradigms and problem solving.

How does folk psychology (external) fare as a theory when theories are
understood syntactically? Not well. Folk psychology (external) as modelled in the
previous Section is probably not axiomatizable in any non-trivial sense. But then,
many contemporary scientific theories are not axiomatizable either. Biological
theories are a case in point: much of biology is not susceptible to axiomatization.
The problem here seems to be a positivist emphasis on physics. It is true that many
physical theories can take the form of axiomatic systems. Once we broaden our gaze
beyond physics, though, the temptation to insist on a syntactic account of theoryhood
passes.

The syntactic approach overemphasizes the linguistic formulation of theories.
Suppe has argued that theories are extralinguistic entities and therefore cannot be
identified with sets of theorems as the syntactic approach demands. Whilst not to
be identified with sets of sentences, theories can nevertheless be described or
characterized by any number of linguistic formulations. This immediately suggests
that a theory is a model for its linguistic formulations, in much the same sense that a
Tarskian model can be a model for a set of sentences in, say, the first order calculus.
This approach offers the additional advantage of not demanding an axiomatizable
formulation of the theory.

On the semantic view as presented by Suppe and van Fraassen, a theory with
$n$ dimensions is modelled as an $n$-dimensional phase space. Theories with laws of
succession (think of the laws of Newtonian mechanics) locate vectors in the phase

---

unclear on this point. Thus on p. 14 he remarks that 'Common-sense psychology is rather like an
implicit higher-order theory, I think'.

40 See Carnap 1950; Hempel 1965; Reichenbach 1938.
42 See Kuhn 1970; Laudan 1977.
43 Folk psychology (external) could be trivially axiomatized by taking as axioms the set of (external)
folk psychological truism.
44 See Lloyd 1988 for a discussion of evolutionary theory's resistance to axiomatization.
space; theories with laws of coexistence (think of Boyle's Law) locate volumes. Folk psychology (external) can be similarly modelled as a phase space, with one dimension for each possible sensory input, mental state, and behavioural output. Truisms such as 'If you touch a hot stove you will cry out' are then treated as laws of succession, plotting vectors in the phase space.

Finally, what of the historical account of theoryhood favoured by Kuhn and Laudan? The central notion in this approach to theoryhood is that of a paradigm. A paradigm is a conceptual scheme shared by workers within a field or discipline. It establishes what counts as a research problem and - crucially - what counts as solving it. Newtonian mechanics, for example, functioned as a paradigm for physicists for over two hundred years. Within that paradigm calculating the orbit of Halley’s comet constituted a research problem, and solutions to that problem were strongly constrained by the demand that the calculation be treated as an application of Newton's laws of motion.

Whilst the notion of a paradigm is notoriously vague, it does seem that folk psychology (external) functions in some sense as a paradigm. Folk psychology (external) both establishes "research" problems and constrains solutions to those problems. A typical problem might be, say, explaining why someone went to the Calypso coffee lounge at one o'clock, and solutions to that problem are constrained in that they must appeal only to the person's system of mental states and processes, and to properties of that system, as recognized by folk psychology (external). But the vagueness of the notion of a paradigm is problematic. The notion of a paradigm is no clearer than, and to some extent dependent upon, the notion of a theory. Appealing to paradigms in the analysis of theoryhood is therefore unilluminating.

We have located grounds, then, for ruling that folk psychology (external) is a theory. Significantly, folk psychology (external) counts as a theory on the most persuasive contemporary account of theoryhood: the semantic account. Nevertheless, I think that there is more to be said about theoryhood and the status of folk psychology (external). In what follows I develop an account of theoryhood inspired in part by Wellman. My approach turns on contemporary empirical research into everyday concepts. I begin by very briefly reviewing that research.

Research by cognitive psychologists into the nature of everyday or folk concepts has revealed that they bear little resemblance to quasi-formal definitions

46 The examples given are deterministic. The approach can be extended to statistical laws. See Suppe 1977: 226-228.
47 Folk psychology (external) also offers laws of coexistence: some combinations of mental states are ruled out.
48 Masterman has located no less than twenty one senses of the term 'paradigm' in Kuhn's own writings. See Masterman 1970.
49 See Wellman 1990, Ch. 5. Note that his account is compatible with the semantic conception endorsed above.
given in terms of short lists of necessary and jointly sufficient conditions. They consist, rather, of a prototype or exemplar. An example falls under a concept if it is sufficiently similar to the relevant exemplar or prototype along various dimensions. Thus consider the concept *bird*. The exemplar for *bird* appears to be (at least for northern hemisphere Europeans) something like the robin. Dimensions of variance might include being feathered, being able to fly, having a beak, being able to sing, and so forth. The sparrow is close to the robin on all these dimensions; the penguin is distant on some and close on others.

*Theoryhood* is a concept. It has a slightly technical air to it — it is somewhat less “everyday” than the concept *bird* — but it is nevertheless unlikely to be characterizable by a short list of necessary and jointly sufficient conditions. Rather, it is probable that there exists certain core examples of theoryhood and a number of dimensions along which particular cases may vary. Core examples might include, say, Newton’s mechanics and Darwin’s theory of natural selection. What of the dimensions along which variance may occur? I suggest that they include at least the following. 1. Ontological commitment — the extent to which, realistically construed, the candidate theory is committed to the existence of various entities and properties. 2. Interdefinability of terms — the extent to which the theoretical terms of the candidate are interdefinable. 3. Explanatory power — the extent to which the candidate can explain and/or predict events within its scope. 4. Structuredness — the extent to which the candidate admits of a highly structured linguistic formulation.

Features (1) to (4) allow us to recognize a continuum of what we might call *data structures*. At one end of the continuum are mere lists. A list largely lacks the features given above. It makes no claims about the ontological status of its elements; its elements are not necessarily interdefinable and the list itself in no way constrains or demands interdefinition; it has little or no explanatory power; and it is largely unstructured.

Features (1) to (4) allow us to recognize a continuum of what we might call *data structures*. At one end of the continuum are mere lists. A list largely lacks the features given above. It makes no claims about the ontological status of its elements; its elements are not necessarily interdefinable and the list itself in no way constrains or demands interdefinition; it has little or no explanatory power; and it is largely unstructured. 55 At the other end of the continuum lie full-blown scientific theories.

50 See Rosch 1973; Smith & Medlin 1981. Much of this research was inspired by Wittgenstein’s remarks on family resemblance. See Wittgenstein 1975: 17.
51 Fodor has claimed that no concept has ever been successfully assimilated to a traditional short list of necessary and jointly sufficient conditions (Fodor 1981). I am inclined to think that this is an overstatement. Nevertheless, the onus of proof has certainly passed to the traditionalist.
52 I am indebted to Wellman at this point. His discussion of theoryhood draws attention to the importance of (1) to (3) below. See Wellman 1990, Chs 1 & 5.
53 The rider ‘realistically construed’ is important. It allows for the existence of purely instrumental theories. Thus whilst Ptolemaic astronomy is treated purely instrumentally by modern navigators and thus lacks ontological commitment, it would have ontological commitment if it were realistically construed.
54 Note that, unlike Wellman, I do not restrict explanation to causal explanation. Whilst much explanation is causal, other acceptable forms of explanation exist: explanations in terms of composition or supervenience. Thus I might remark that the rock sank because it was composed of silicon atoms, or that the fitness of an animal in an environment supervenes on various physical properties of the animal and the environment. These remarks constitute non-causal explanations.
55 By a ‘list’ I mean an unordered n-tuple. An ordered n-tuple exhibits at least some structure.
Think of Newtonian mechanics. Newtonian mechanics exhibits to a large extent features (1) to (4). Realistically construed it is committed to the existence of space, time, mass, force, etc; its terms are extensively interdefined (force is defined in terms of mass and acceleration; gravity is a force; etc); it offers a very powerful explanatory framework; and it admits of a highly structured linguistic representation.56

Where on this continuum does folk psychology (external) lie? I think that pretty clearly folk psychology (external) lies well towards the scientific theory end of the continuum. Consider features (1) to (4). Folk psychology (external), realistically construed, carries considerable ontological commitment. It posits the existence of an entire system of mental states and processes, plus various properties of that system. Moreover, its terms are extensively interdefined: action in terms of belief and desire; cheerfulness in terms of a disposition to be happy; crying out in terms of pain; etc. Further, folk psychology (external) has, as we have seen, considerable predictive and explanatory power. It is only on the last feature – structuredness – that folk psychology (external) loses some ground. It does not admit of a highly structured linguistic presentation. Rather, it consists to a very large extent of a loose set of platitudes or truisms that resist axiomatization and other forms of structuring.

I think it is clear that folk psychology (external) lies well towards the scientific theory end of the continuum – close enough to deserve the label 'theory'. It is not, though, a scientific theory. But to insist that folk psychology (external) is not a theory because it is not a scientific theory is to unacceptably curtail the use of the term 'theory'. Folk psychology (external) exhibits, I have urged, sufficient similarity to the core exemplars of theoryhood to warrant the application of the term 'theory'. That other theories are still closer to the core exemplars is irrelevant.

1.6 Overview. I began this Chapter by drawing attention to our remarkable capacity to predict and explain action. The theory theory (internal) seeks to explain that capacity by positing an internally represented theory of human psychology: folk psychology (internal). As we saw in Section 1.1, explaining cognitive capacities by positing an internally represented knowledge structure is the modus operandi of contemporary cognitive science. There is therefore a certain presumption in favour of the theory theory (internal). There exists, though, a rival account of our ability to predict and explain action: off-line simulation theory. In Chapter 2 I elaborate and attempt to rebut that position.

56 Indeed, it admits of an axiomatization. Thus my approach concedes just a little to the syntactic approach to theoryhood discussed above. Note, though, that structuredness is neither a necessary nor a sufficient condition for theoryhood on my account. It is merely one of several typical features of theories.
In Section 1.4 I made an attempt to model folk psychology (external), focussing particularly on (external) folk psychological cognitive architecture. According to the model offered, folk psychology (external) is essentially Humean: behaviour is largely seen as caused by the interplay of beliefs and desires. The concepts of belief and desire thus lie at the very heart of (external) folk psychological theorizing. What are beliefs and desires? Or, more properly, what is the (external) folk psychological conception of beliefs and desires? In Chapter 3 I examine various analyses of belief. The authors of these analyses have not always seen themselves as seeking analyses of an (external) folk psychological concept, but it is imperative that we bear in mind that our aim is to get clear on the meaning of a theoretical term introduced by a folk theory.

The analysis of belief I endorse in Chapter 3 has its origins in the work of David Armstrong and David Lewis. Armstrong and Lewis' approach, moreover, allows us to see in what sense our everyday talk about mental states might be said to implicitly constitute a theory of mind. Chapter 3 therefore functions in addition as a defense of the theory theory (external).

One feature of (external) folk psychological beliefs and desires not discussed in Chapter 3 is their semantic properties. Beliefs are true or false; desires satisfied or not. In Chapter 4 I review existing accounts of the contentfulness of (external) folk psychological propositional attitudes and endorse Jerry Fodor's asymmetric dependency theory of content.

The Armstrong/Lewis account of belief is a functionalist one. Both Ned Block and John Searle have authored well-known attacks on functionalism. It is not clear that either Block or Searle saw themselves as attacking an approach to analyzing (external) folk psychological concepts; nevertheless, their attack can be thus construed. In Chapter 5 I outline their arguments and show that they do not constitute a serious problem for the analysis of (external) folk psychological concepts which I endorse. In addition, functionalism has been attacked on the grounds that it is incapable of accounting for qualia. I attempt to rebuff the qualia objection in Section 5.4.

By the end of Chapter 5 we should be fairly clear about the general nature of folk psychology (external). Notoriously, eliminativists think that folk psychology is a radically false theory. Indeed, according to eliminativists, folk psychology is such a bad theory that its central posits – beliefs and desires – do not exist. Just as we have distinguished two senses of 'folk psychology', so must we distinguish two senses of

58 See Fodor 1987; 1990c.
60 See for example P.M. Churchland 1981; Stich 1983.
'eliminativism': eliminativism about the posits of folk psychology (internal) and eliminativism about the posits of folk psychology (external). Most of the eliminativist literature has focussed on the posits of folk psychology (external). In Chapter 6 I survey and reject a variety of eliminativist arguments. In addition, I consider two anti-eliminativist strategies and argue that, at least for the present, it is reasonable to accept that folk psychology (external) is true simply because of its considerable predictive success.

Chapter 7 takes up a highly specialized eliminativist concern ignored in Chapter 6. Jerry Fodor, Fred Dretske, and others have worried that intentional states such as beliefs and desires may not be naturalizable. This is a source of concern because they claim that non-naturalism entails irrealism: if intentional states cannot be naturalized then they do not exist at the actual world. We can only begin to address Fodor and Dretske's worries, however, if we have at hand an account of naturalism. I endorse an account of naturalism developed by David Lewis and Frank Jackson. On that account, though, intentional non-naturalism does not entail intentional irrealism. An argument does exist, however, from intentional non-naturalism, plus another premiss, to intentional irrealism. I consider that argument in some detail and conclude by making some general remarks about naturalism.

Chapters 6 and 7 are, as we have seen, largely focussed on eliminativist concerns. I do not endorse any of the eliminativist arguments surveyed, nor do I have new ones of my own to put in their place. The relationship between folk psychology - in either sense - and mature cognitive science is therefore wide open.

In the 1960's Wilfred Sellars theorized about the relationship between what he called the manifest and scientific images. The manifest image is, roughly, the world as it is presented to us; the scientific image the world as science depicts it to be. The two are in tension: the manifest image of my desk's top is one of a smooth continuous surface; the scientific image one of a complex structure of particles in empty space. A similar tension exists between the manifest image of ourselves and the image revealed by neuroscience. I see myself in (external) folk psychological terms - as an agent holding various beliefs and driven by various desires - but the neurobiologist views my behaviour as resulting from interactions between vast arrays of cells of various specialized descriptions.

---

62 See Lewis 1983a; Jackson forthcoming.
63 See Sellars 1963.
64 Strictly speaking, my identification of folk psychology (external) with (part of) the manifest image is inaccurate since, according to Sellars, the manifest image posited no unobservable entities.
In Chapter 8 I argue, *contra* eliminativism, that the manifest and scientific images need not be in tension. An intentional psychology based on folk psychology (external) will indeed find its place amongst the cognitive sciences.
2.1 Introduction. I began the previous Chapter by drawing attention to our capacity to predict, and offer explanations of, our own and others' behaviour. The theory theory (internal) seeks to account for that capacity by positing an internally represented theory of human psychology – folk psychology (internal). How is the theory theory (internal) to be defended? We should not expect to find convincing a priori arguments for the theory theory (internal). It is, after all, an empirical hypothesis, and empirical hypotheses are not provable a priori. What we might hope to find, rather, is an argument to the best explanation; that is, an argument of the form: all things considered, the theory theory (internal) is the best available explanation of the capacity in question.

We have already began to assemble such an argument. As stated in Section 1.1, positing internal knowledge structures is the modus operandi of contemporary cognitive science. Explanations of many cognitive capacities have been offered that appeal to internally represented theories, or to other types of knowledge representations. The theory theory (internal) is therefore parsimonious in that it deploys an explanatory strategy already widely used, rather than developing an entirely new strategy to handle just the capacity to predict and explain behaviour. Moreover, the theory theory (internal) has been successfully exploited in the explanation of various features of cognitive development. More on this theme below.

Defending the theory theory (internal) by a strategy of argument to the best explanation, though, demands that we examine rival accounts of our capacity to predict and explain behaviour. At present only one rival exists: off-line simulation theory. In Section 2.2 I sketch off-line simulation theory. Then, in Section 2.3, I consider various arguments for and against off-line simulation theory, focussing on those arguments I find most persuasive. Other arguments exist but will be largely ignored. I close the Chapter by briefly mentioning an intriguing proposal which

---

1 These arguments are well stated in Stich & Nichols 1992, forthcoming.
links the capacity to predict and explain behaviour with autism, and by quickly drawing attention to a claim made by certain off-line simulation theorists about the status of the eliminativism debate.

2.2 What Is Off-line Simulation Theory? Off-line simulation theory posits a cognitive architecture closely related to that of folk psychology (external).² Figure 2.1, below, offers a simplified version of (external) folk psychological cognitive architecture.

² What if folk psychology (external) is false and the cognitive architecture it posits mistaken? The extent to which off-line simulation theory is compatible with alternative cognitive architectures is an interesting question which I will not pursue.
Note that according to the theory theory (internal), the prediction and explanation of action involves processes that are *information rich*. The theory theory (internal) claims that a very extensive theory of human psychology is stored in the belief box and utilized in predicting and explaining behaviour. (See figure 2.2, below.) Simulation theory, on the other hand, calls upon processes which are *information poor*. Rather than positing an internally represented theory of human psychology, it exploits the fact that each of us operates according to the principles of human psychology when we ourselves make decisions. Figure 2.3, below, illustrates how the trick is turned. Say that I believe that S both wants a cappuccino and thinks
that they are available at the Calypso coffee lounge. Now if I assume that S's reasoner works pretty much as mine does, I could predict what S would do if I could figure out what I would do if I had her beliefs and desires. According to off-line simulation theory, figuring out what I would do if I had S's beliefs and desires involves two new additions to the cognitive architecture presented in figure 2.1. First, a device is required which generates what we might call "pretend" beliefs and desires and feeds them into the reasoner. This device cannot just consist of the belief and desire boxes and related machinery. For, in the case we are considering, the beliefs and desires in question are not my beliefs and desires but S's. I am just pretending to have them. Second, a mechanism is required for taking the output of
the reasoner 'off-line'. If I acted upon the output of my reasoner in the usual fashion, I would end up acting upon S's beliefs and desires: I would go to the Calypso coffee lounge. Rather than acting upon S's beliefs and desires, I want the output fed into my belief box as a prediction.

So far so good. We have a rival account of prediction – an account which does not require the existence of a tacit theory of human psychology stored in the belief box. But we are assuming that the predictor has beliefs about the subject's beliefs and desires. In the case above, we assumed that I believed both that S wanted a cappuccino and that S thought that they were available at the Calypso coffee lounge. The theory theory (internal) has an account of how we arrive at beliefs about other people's beliefs and desires: we exploit folk psychology (internal) and whatever information we can glean from observation and background knowledge. Thus in S's case we might exploit our knowledge of her caffeine addiction, our belief that she knows her way around campus, and so forth. According to off-line simulation theory, on the other hand, we explain action by generating and testing belief/desire pairs. Thus, say that I want to explain S's going to the Calypso coffee lounge. My "pretend" belief/desire generator constructs belief/desire pairs and feeds them into my reasoner. The output is taken off-line and fed into my belief box. The result is then compared with S's actual behaviour. This process is continued until a belief/desire pair is located that causes my reasoner to generate an output which, were it on-line, would result in action identical to S's. The belief/desire pair she wants a cappuccino and thinks that they are available at the Calypso, generates the correct output and so constitutes the explanation of S's action.

Now this is a bit quick, for any number of belief/desire pairs exist which might explain S's going to the coffee lounge. For example, S might desire a sandwich and believe that the Calypso makes the best sandwiches on campus. Or S might desire a really cold coke and believe that the Calypso's cokes are the coldest on campus. And so forth. To some extent the theory theory (internal) faces the same problem: there may be insufficient information about the circumstances in which the action took place to yield a definitive explanation. But for the off-line simulation theorist the problem is particularly acute, for the processes posited by off-line simulation theory are, as already noted, information poor: according to off-line simulation theory, we lack a rich body of (internal) folk psychological lore to draw on when deciding between competing belief/desire pairs. This has led simulation theorists to exploit a principle of humanity, much in the spirit of Richard Grandy's
Davidsonian proposal, to help sort out competing belief/desire pairs. If I assume that you are pretty much like I am, I can assume that you believe and desire pretty much what I would believe and desire if I were in your shoes. The set of belief/desire pairs to be generated and tested is thus reduced.

2.3 Defending The Theory Theory (Internal). Whether defending or attacking off-line simulation theory, it is important to get clear just what is at stake. Unfortunately, many arguments advanced in favour of off-line simulation theory are badly marred by terminological confusion and equivocation. Two points are therefore worth stressing. First, I am using the 'theory theory (internal)' to refer to any theory that seeks to account for our capacity to predict and explain action by positing an internally represented theory of human psychology, irrespective of the way in which the posited theory is represented. The expression 'the theory theory (internal)' is not to be understood as assuming a mental sentence account of mental representation – perhaps the internally represented theory is represented in a distributed fashion in a connectionist network. It follows that off-line simulation theory cannot be defended simply by attacking the mental sentence account of mental representation. It is not enough to show that the mental sentence account of mental representation is false. Rather, it must be shown both that the mental sentence account of mental representation is false and that the theory theory (internal) is committed to it. No argument demonstrating (as opposed to merely stipulating) that the theory theory (internal) is committed to the mental sentence account of mental representation is forthcoming from the off-line simulation camp.

A second confusion turns on the use of 'simulation'. The theory theory (internal) is compatible with the existence of various mental processes that might be called 'simulations'. Here is an example. How many windows does your residence have? Most people, when trying to answer this question, imagine themselves walking from room to room, counting up windows as they go. Now the process of imagining yourself moving from room to room surely counts as a simulation in some loose sense of the term: you are mentally simulating the action of walking around your house. But obviously what you are doing here is accessing an internal representation of your residence; you are not running an off-line simulation of your

---


4 For a thorough listing of confusions and equivocations see Stich & Nichols 1992, forthcoming.

5 This example is from Stich & Nichols 1992: 55.
residence. How could you? You are not, after all, appropriately similar to the house you live in. The point is that it is only the existence of a very special class of simulations – off-line simulations – that supports an anti-theory theory (internal) position. Off-line simulation theory cannot be defended, as Goldman has tried to do, simply by drawing attention to the various cognitive activities that might loosely be called 'simulations'.

A variety of other arguments have been offered in favour of off-line simulation theory. These arguments do not turn on terminological confusions or equivocations, but are nevertheless unconvincing. I will give just four examples.

**First argument.** If we really do possess an internally represented theory of human psychology, we ought to be able to articulate it. Nobody has succeeded in fully articulating folk psychology (internal). Therefore we do not possess such a theory and so the theory theory (internal) is false. As the only other game in town is off-line simulation theory, off-line simulation theory should be accepted. This argument is clearly spurious. For, according to the theory theory (internal), the theory of human psychology underpinning our capacity to predict and (putatively) explain action is *largely tacit*. Without an independent argument to the effect that positing largely tacit knowledge representations is problematic, the argument under consideration simply begs the question against the theory theory (internal). Moreover, the general claim that if we possess an internally represented theory we should be able to articulate it is implausible. After all, according to current orthodoxy in linguistics, our internally represented grammar is largely tacit in just the same way as folk psychology (internal) is held to be. So a consequence of the argument under consideration is that the current orthodoxy in linguistics is false. I for one am not in a hurry to embrace that conclusion.

**Second argument.** The theory theory (internal) is committed to the existence of an internally represented human psychology which is, presumably, both complex and subtle. But quite young children are capable of predicting and explaining human behaviour. It is implausible that quite young children have mastered a theory of human psychology which is both complex and subtle. Therefore the theory theory (internal) is false and hence (given the standard assumption) off-line simulation theory is true. No argument is offered to defend the claim that young children cannot master complex and subtle theories. Indeed, there is independent evidence that they can. For again consider the case of natural languages. Quite young

---

children learn how to speak and understand natural languages. The orthodox position in linguistics is that they do so by deploying an internally represented grammar. But the grammars of natural languages are extraordinarily complex and subtle. Therefore quite young children can master at least one complex and subtle knowledge representation. If they can master the grammar of their native tongue, why deny that they can master folk psychology (internal)?

**Third argument.** The states and processes posited by the theory theory (internal) are information rich. Those posited by off-line simulation theory, on the other hand, are information poor. Considerations of parsimony therefore favour off-line simulation theory over the theory theory (internal). But, as Stich and Nichols have pointed out, this argument is a *non sequitur.* 8 It is true that the theory theory (internal) makes heavy demands on information – it requires an entire tacit theory of human psychology – but it makes relatively few demands on cognitive architecture. Off-line simulation theory, on the other hand, makes few demands on information, but it makes relatively heavy demands on cognitive architecture: in addition to the standard (external) folk psychological cognitive architecture it posits a "pretend" belief/desire generator and a mechanism for taking the reasoner off-line. So parsimony considerations are inconclusive: they favour neither off-line simulation theory nor the theory theory (internal).

**Fourth argument.** Greg Currie has suggested that the theory theory (internal) can be attacked by a modification of Frank Jackson's knowledge argument against naturalistic accounts of qualia. 9 Jackson invites us to consider Mary who has spent her entire life in a black and white room. Mary has a complete knowledge of herself and her world described as physical systems; that is, she has a total grasp of completed physics. Nevertheless, there is something which she does not know. For when she is released from her monochromatic prison she will learn what it is like to see something that is, say, red. Since by hypothesis Mary knows all there is to know about the physical world, what Mary learns when she leaves the black and white room is not physical knowledge. Therefore qualia are not physical.

This is not the moment to attempt a detailed evaluation of the knowledge argument. Rather, let us see how an analogous argument against the theory theory (internal) might be constructed. Say that the theory theory (internal) is true but that

---

9 See Currie forthcoming: fn 9. For Jackson's knowledge argument see Jackson 1982, 1986. I discuss Jackson's knowledge argument in Section 5.4, below. There also exists an analogy between Currie's anti-theory theory (internal) argument and some aspects of Searle's discussion of his Chinese room example. (See Searle 1980a.) I discuss Searle's argument in Section 5.3, below.
Mike, by some unfortunate accident, never developed folk psychology (internal). In other words, Mike's belief box does not contain folk psychology (internal). In all other respects, though, Mike is normal; in particular, his perceptual and reasoning processors are functioning normally. Fortunately, cognitive scientists have managed to write down all the generalizations about the causal relations between sensory inputs, mental states, and behavioural outputs which make up folk psychology (internal). Mike commits all the generalizations to memory. In doing so, will he have acquired the capacity to predict and explain behaviour with the same fluency as ordinary folk? Currie urges that he will not. If Currie is right then it seems that the theory theory (internal) is wrong. Since Mike now has a representation of folk psychology (internal) in his belief box, and since his perceptual and reasoning processors are, by hypothesis, working normally, he has everything that, according to the (internal) theory theorist, he requires for the fluent prediction and explanation of behaviour. Therefore, if he cannot fluently predict and explain behaviour, the theory theory (internal) is wrong.

One way to attack Currie's argument would be to deny the crucial premiss that Mike, having memorized all the relevant generalizations, would not be able to predict and explain behaviour with the fluency of ordinary folk. Currie offers no evidence in support of his claim. I must admit, though, that I share his intuition: I seriously doubt that Mike would cope with ordinary social circumstances at all well.

A second way to attack Currie's argument is to recall the observation made in Chapter I: internally represented theories may not take the form of a set of declarative sentences. It follows that there simply may not be a set of generalizations which Mike can memorize. However, for the present discussion I will simply assume that folk psychology (internal) can in fact take a form such that Mike could commit it to memory.

I think that we can begin to see what is wrong with Currie's argument if we once again draw on a comparison between folk psychology (internal) and grammar (internal). Say that Mike had failed to acquire not folk psychology (internal) but grammar (internal). Fortunately, linguists have succeeded in writing down a complete specification of grammar (internal) – perhaps in the form of phrase structures; perhaps in some other form. Mike commits the complete specification of grammar (internal), plus a large lexicon, to memory. Would we expect him to be able to speak and understand a natural language with the fluency of ordinary folk? Again there exists a strong intuition that Mike would not gain anything like normal fluency. By analogy with the previous argument it seems to follow that our capacity
to speak and understand natural languages is not underpinned by an internally represented theory. But that is absurd. There is simply no other account of our capacity to speak and understand natural languages available.\(^\text{10}\)

The anti-theory theory (internal) argument under consideration understands the metaphor of a belief box too crudely. In particular, it fails to notice that the theory theory (internal) is compatible with folk psychology (internal) normally being stored in a special, dedicated, module which facilitates its fluent deployment in the prediction and explanation of behaviour. Thus, whilst it is true that the set of generalizations which constitute folk psychology (internal) could be stored elsewhere – perhaps in a general purpose information storage and retrieval module – deploying the generalizations thus stored might be slow and cumbersome. In other words, whilst Mike could memorize the generalizations which we are assuming constitute folk psychology (internal), memorizing those generalizations will not necessarily lead to their being stored in the same way that they are stored in a normal person – a way which, moreover, facilitates their fluent deployment. Obviously, similar remarks apply to the case in which Mike memorizes grammar (internal). In normal people grammar (internal) is stored in a special, dedicated, module which facilitates its fluent deployment. Merely memorizing grammar (internal) in the fashion Mike is supposed to have done does not guarantee that it will be stored in the appropriate module.

I now want to consider four arguments against off-line simulation theory. (1) An argument from developmental psychology.\(^\text{11}\) (2) An argument from cognitive penetrability.\(^\text{12}\) (3) An argument to the effect that off-line simulation theory must tacitly assume a rich corpus of folk psychological law. This argument is a development of an argument of Paul Churchland's.\(^\text{13}\) (4) An argument to the effect that off-line simulation theory ignores a significant domain of (external) folk psychological prediction and explanation. All four arguments against off-line simulation theory have the same general structure. They proceed by trying to locate an aspect of our cognitive life which can readily be explained by the theory theory (internal) but not by off-line simulation theory. As we might expect, none of

\(^\text{10}\) It cannot be stressed enough that internally represented theories need not take the form of sentences in the language of thought. Thus, whilst it could turn out that grammar (internal) is represented in a widely distributed connectionist network, that would not constitute an objection to my claim that positing grammar (internal) is the only available account of our capacity to speak and understand natural languages.


\(^\text{13}\) See P.M. Churchland 1989: 119.
these arguments is utterly convincing. In each case off-line simulation theory can be defended by *ad hoc* modifications – by, as it were, piling on the epicycles. Parsimony considerations, however, then begin to weigh heavily in favour of the theory theory (internal).

In Section 1.2 we examined children's development of the capacity to predict and explain behaviour. A variety of (internal) theory theorists have argued that the theory theory (internal) smoothly accounts for the general course of development sketched in Section 1.2.14 In particular, Gopnik and Wellman have argued that the development of the child's capacity to predict and explain action can be seen in terms of a series of theory changes.15 Thus, the two year old possesses a crude theory of human psychology and uses it with some success to predict behaviour. However, the two year old's theory of mind only quantifies over states akin to desires and perceptions; it does not quantify over any states which can be likened to beliefs. By three, a theory change has occurred, with the child's theory of mind now quantifying over states akin to beliefs as well as over states akin to desires and perceptions. Further theory changes subsequently occur, with the child positing belief states which can be true or false, as well as positing a variety of other states including dreams and images.

Gopnik and Wellman are quite correct when they claim that the theory theory (internal) smoothly accounts for the general course of the child's development of the capacity to predict and explain behaviour. However, with only a little ingenuity the off-line simulation theorist can offer a rival account of the general course of development. After all, if the (internal) theory theorist can avail herself of a long process of theory change, the off-line simulation theorist can avail *herself* of a long process of development of the mechanism responsible for off-line simulation. Any charge that the off-line simulation theorist's account is *ad hoc* is likely to apply, *mutatis mutandis*, to the (external) theory theorist.

However, whilst the off-line simulation theorist should be able to account for the *general* course of development of the capacity to predict and explain behaviour – the course that takes us from the two year old's failure at standard belief tasks to the four year old's success at false belief tasks – the *details* of that development provide the off-line simulation theorist with considerable obstacles.

---

14 See for example Gopnik & Wellman 1992; Wellman 1990.
Wimmer, Hogrefe and Sodian have explored children's ability to predict what beliefs others will have. In a typical experiment the subject was shown that a container held only choconuts. A choconut was then transferred from the container to a bag. The subject could not see the transfer but was told that a single item was taken from the container and placed in the bag. The subject was then asked what was in the bag. The majority of four year olds correctly inferred that the bag contained a choconut. However, most four year olds were unable to predict that a second child, who had seen the contents of the container and had been told that a single item was transferred to the bag, would know what was in the bag.

The theory theory (internal) has no trouble at all explaining this result. The four year old can make an accurate assessment of the contents of the bag because she is capable of forming a belief about the contents of the container on the basis of seeing inside it; is capable of forming a belief about the transfer of an item from the container to the bag on the basis of a verbal report; and is capable of inferring from those two beliefs to the contents of the bag. But, the advocate of the theory theory (internal) can argue, she is not capable of determining another person's informational state on the basis of her beliefs about what they have seen and been told, because she has not yet developed the relevant components of folk psychology (internal). In particular, she has not yet developed a theory about informational access under various inferential conditions: she does not yet know that other people can generate new beliefs from old ones. It follows that whilst she is able to respond correctly to the question about the bag's contents, she is unable to predict what the other child will say is in the bag.

But of course we cannot defend the theory theory (internal) by simply showing that it can handle this case; we must also show that off-line simulation theory cannot handle the data presented by Wimmer et al. And indeed it does seem that off-line simulation theory is in trouble. For does not the subject have everything required for a successful simulation? She knows that the other child has seen the contents of the container and she knows that the other child has been told about the transfer. Furthermore, there is ample evidence from other studies that four year olds can correctly attribute beliefs to another when they have seen the other perceive or be

---

17 Doubt has been cast on some aspects of Wimmer, Hogrefe and Sodian's experimental set up. (See Wellman 1990: 186.) Stich and Nichols have conceded this point (Stich & Nichols forthcoming). However, there is ample experimental evidence to support Stich and Nichols' premise. See Gopnik & Wellman 1992.
18 Indeed, she may not even know that she herself has this capacity. Her own inferential practices may as yet be entirely transparent to her.
told something. Finally, the child's reasoning apparatus does not lack the relevant inferential capacity: she herself inferred what was in the bag from her belief about the contents of the container and her belief about the transfer. So it seems that off-line simulation theory predicts that the subject will correctly determine what the second child will say. Since the subject does not correctly determine what the second child will say, off-line simulation theory is false.

Now, as we might expect, this argument is not absolutely persuasive. For we can conceive of variants of off-line simulation theory as it has been presented that would yield the right result. Thus the off-line simulation theorist could claim that the four year old's off-line simulator is working well enough to permit attributions of belief to another when the other has been seen to have perceptual or verbal access to the relevant information, but not well enough to handle the case presented by Wimmer et al. These sort of responses, though, look more than a bit ad hoc. I think therefore it is fair to say that the theory theory (internal) handles more comfortably the data from developmental psychology than does off-line simulation theory.

The argument we have just been considering locates a feature of cognitive life that is readily explained by the theory theory (internal) but not by off-line simulation theory. The same general structure is apparent in the argument from cognitive penetrability. The theory theory (internal) permits a divergence between what we would predict ourselves to do and what we would in fact do. For according to the theory theory (internal), our predictions are driven by an internally represented theory of human psychology, and there is no guarantee of that theory's accuracy and completeness. In other words, according to the theory theory (internal), our predictions are cognitively penetrable: our tacit theory of human psychology influences (or 'penetrates') our predictions.19 It follows that what I think I would do in some particular circumstance and what I would in fact do in those circumstances may come apart. According to off-line simulation theory, on the other hand, prediction is not cognitively penetrable. What I predict I would do and what I would in fact do should be essentially the same. After all, if off-line simulation theory is correct, the mechanism that generates my prediction and the mechanism that generates my action are one and the same: my very own reasoner.

We can therefore distinguish between the theory theory (internal) and off-line simulation theory by determining the extent to which the prediction of action is cognitively penetrable: if it is cognitively penetrated to a marked degree, off-line

19 Stich and Nichols adopted the expression cognitively penetrable from Pylyshyn 1984.
simulation theory is false. Stich and Nichols have offered several examples of predictions of action which are cognitively penetrated.\(^{20}\) I will sketch my favourite.\(^{21}\)

At your local supermarket you approach what appears to be a market research initiative. A number of items are displayed on a table. You, the consumer, are invited to select the item of your choice. In return you are expected to say a few words about why you chose the item you did in fact choose. After looking over the items available you make your choice and give your reason - colour, quality, size, what have you. In fact the whole set up is bogus and the items are identical. What emerges is that people tend to choose items from the right hand side of the table. But - and this is crucial - subjects almost universally deny that position had any bearing on their choice. So it seems that people's predictions of their own actions can diverge from the actions they in fact perform: we would not predict that position has an effect on choice when in fact it does.

The theory theory (internal) has no trouble explaining this result: folk psychology (internal) is silent about the effect of position on choice. Guided in our predictions by an incorrect - or at least unduly circumscribed - theory of psychology, we go wrong. Off-line simulation theory, though, is in trouble. For the mechanism that subserves prediction is, according to off-line simulation theory, one and the same mechanism that generates action. Our predictions about our choice of item and our actual choice should therefore strongly coincide. But they do not - they markedly diverge.\(^{22}\)

Before moving on to the third argument against simulation theory, I would like to make a brief comment about the argument from cognitive penetrability. I said above that the cognitive penetrability of prediction offers a way of distinguishing between the theory theory (internal) and off-line simulation theory. And so it does. Thus Stich and Nichols remark that if suitable experiments establish the cognitive penetrability of prediction, 'then the simulation theory should be abandoned; if they come out the other way, then it is the theory theory [(internal)] that is in deep trouble'.\(^{23}\) This is largely right, but not entirely so. For whilst evidence for cognitive penetrability strongly tells against off-line simulation theory, failure to locate

\(^{22}\) For simulationist responses to the argument from cognitive penetrability see Goldman 1992 and Gordon 1992. For convincing replies see Stich and Nichols forthcoming.
\(^{23}\) Stich & Nichols forthcoming.
evidence for cognitive penetrability does not necessarily tell against the theory theory (internal). For say that folk psychology (internal) turns out to be a really good theory of human psychology, so good that deploying folk psychology (internal) practically always leads to correct predictions. Then it would appear that our predictions of action are not cognitively penetrable, for we recognize that prediction is cognitively penetrated only when prediction diverges from action. Of course, this possibility is rather academic as we already know that folk psychology (internal) does not facilitate perfect prediction – the case given above establishes that. Nevertheless, it is rather ironic that a major defense of the theory theory (internal) turns on the fact that the theory posited by the theory theory (internal) is less than perfect.

The third argument against off-line simulation theory I wish to canvas attacks the off-line simulation theorist's account of the explanation of action. Recall that, according to off-line simulation theory, explanations of action are obtained by generating and testing belief/desire pairs. The belief/desire pair which, were it fed into the reasoner on-line, would generate the action in question, constitutes the explanation of the action. As mentioned earlier, a problem arises in that indefinitely many belief/desire pairs exist which would generate the action in question. The off-line simulation theorist therefore appeals to a principle of humanity: assume that the person whose actions are to be explained is much as you are. Under this assumption, only those belief/desire pairs which you yourself would hold in their circumstances need to be tested.

Paul Churchland, though, has raised an objection to this approach to the explanation of action.24 What if the person whose actions are to be explained is not similar to the person doing the explaining? We can, after all, fairly successfully attribute beliefs and desires to those who differ from ourselves in various ways (for example, those who hold religious views we take to be bizarre). The theory theory (internal) can account for our capacity to explain the actions of those who diverge from us: the tacit theory of human psychology posited by the theory theory (internal) is rich enough to cover folk who diverge from us to some extent.25 How, though, can off-line simulation theory cope with the problem of explaining the actions of those who diverge from ourselves? Adopting a principle of humanity is no use here – the assumption that they are like us is precisely the assumption that is being

25 This is not to say that folk psychology can explain the actions of people who diverge wildly from the norm: folk psychology's impotence in the face of madness has been widely remarked upon. See for example P.M. Churchland 1981.
denied. The problem faced by the off-line simulation theorist arises because off-line simulation theory is information poor. Usually the off-line simulation theorist can avoid the consequences of a paucity of information by relying on the very principles of human psychology that drive the person seeking the explanation. But when the person whose actions are to be explained differs markedly from the person who is seeking the explanation, the off-line simulation strategy collapses.

Alvin Goldman has attempted to defended off-line simulation theory against Churchland's attack. Goldman claims that when explaining the actions of those who diverge markedly from ourselves we rely on induction: we formulate principles of human action based on past experience. Thus he concedes that 'inductive or nomological information is not wholly absent [from the explanation of action], but it is sparser than the [theory theory (internal)] approach alleges'. The idea seems to be that the explanation of action is recursive, with the recursion finally being grounded in explanations of action generated solely by off-line simulation. The fundamental mode of action explanation deploys off-line simulation, but once some explanations have been generated by that method, further explanations can be located by a process of induction.

But now the game is up. For Goldman is positing both an off-line simulation mechanism and an internally represented theory of human psychology (the principles of action gleaned by induction). Parsimony strongly rules out such an approach to the explanation of action. As we have seen, prior to conceding that off-line simulation theory demands a rich body of beliefs about human psychology, parsimony considerations seemed fairly inconclusive in selecting between off-line simulation theory and the theory theory (internal). Now that Goldman has posited an internally represented human psychology, though, parsimony heavily favours the theory theory (internal).

A final argument against off-line simulation theory involves pointing out that it is unacceptably limited in the scope of predictions and explanations with which it can deal. In Section 1.4 I introduced a distinction between action and behaviour. Actions are those behaviours which folk psychology (external) putatively explains by appealing to beliefs and desires. That is, actions result from the interplay of beliefs and desires. There exist, though, behaviours which are not actions – behaviours which are not (putatively) explained by folk psychology (external) in terms of beliefs and desires. Crying out upon touching a hot stove is an example of a

26 See Goldman 1989: 176-177.
behaviour which is not an action. Beliefs and desires are not involved in the (external) folk psychological explanation of crying out upon touching hot stove. As we have seen, folk psychology (external) is largely concerned with actions rather than non-action behaviours. It is, though, concerned with non-action behaviours to some extent.

The off-line simulation theory put forward by Gordon and others presupposes the cognitive architecture posited by (external) folk psychological. It posits two additions to that architecture: a "pretend" belief/desire generator and a mechanism for taking the reasoner off-line. "Pretend" beliefs and desires are fed into the reasoner which is taken off-line. The result is then either issued as a prediction or compared with the target action in order to locate an explanation. But this mechanism cannot account for our capacity to predict and explain non-action behaviours. By definition such behaviours do not arise from the interaction of beliefs and desires. Since they do not arise from the interaction of beliefs and desires, they cannot be explained by a mechanism which works by manipulating "pretend" beliefs and desires. Off-line simulation theory is thus powerless to account for our capacity to predict and explain non-action behaviours.

On the other hand, the theory theory (internal) offers a plausible account of our capacity to predict and explain non-action behaviours. The story is simple: the internally represented theory of human psychology posited by the theory theory (internal) covers both actions and non-action behaviours. Of course off-line simulation theory could be modified to cope with the prediction and explanation of non-action behaviours. For example, the off-line simulation theorist could posit not merely a "pretend" belief/desire generator but also a "pretend" sensation generator. But such modifications will render the off-line simulation story considerably less parsimonious than its rival.

2.4 Concluding Remarks. There is an important aspect of the debate between the (internal) theory theorists and the off-line simulation theorists which I have so far ignored. Both sides have claimed that their theory of the capacity to predict and explain behaviour also yields an account of autism. Autism is typified by an impaired capacity to form and understand social relations and an inability—or unwillingness—to participate in activities that involve pretense. Intriguingly, autistic children perform poorly on false belief tasks. In fact, autistic children perform less

well on false belief tasks than Down's syndrome children with substantially lower IQs. This has suggested to a variety of (internal) theory theorists that autism is to be explained in terms of a failure to develop and/or deploy folk psychology (internal). In particular, Alan Leslie has argued that both pretense and the capacity to predict and explain behaviour involves 'decoupling' a symbol from its normal causal connections. Off-line simulation theorists, on the other hand, have argued that off-line simulation theory can more smoothly account for autism than can the theory theory.

I do not propose to examine the autism debate in detail. Rather, I want to point out that the basic argument structure shared by both the (internal) theory theorists and the off-line simulation theorists is less than convincing. Both sides begin by observing that the deficits characteristic of autism are consistent with the developmental failure of our capacity to predict and explain behaviour. They then go on to identify autism as a condition caused by the developmental failure of the mechanism responsible for our capacity to predict and explain behaviour. The two accounts of autism differ only in that they offer different accounts of the mechanism responsible for the capacity to predict and explain behaviour.

Notice that the premiss held in common by the (internal) theory theorists and the off-line simulation theorists is not beyond question. That is, it is not entirely clear that the deficits characteristic of autism are consistent with the developmental failure of the capacity to predict and explain behaviour. Thus, autistic individuals often have marked difficulty with verbal communication, exhibit executive function deficits, and have trouble with problem solving. It is not obvious how such deficits can be accounted for in terms of a failure to develop the capacity to predict and explain behaviour. It might be responded that the failure of autistic individuals to develop the capacity to predict and explain behaviour accounts for their difficulty with social relations, and that their difficulty with social relations in turn accounts for their poor performance on tests designed to measure their capacity to solve problems and so forth. However, if the poor performance of autistic individuals on such tests is an artifact of their poor social skills, we would expect their difficulty with psychological testing to be quite general. There are,

31 See for example Currie forthcoming.
though, some psychological tests on which autistic individuals perform very well. For example, autistic children perform normally on a variety of perception tasks.\textsuperscript{33} However, even if we accept that the deficits characteristic of autism are consistent with the developmental failure of the capacity to predict and explain behaviour, it does not follow that autism is caused by the developmental failure of the mechanism responsible for the capacity to predict and explain behaviour. When dealing with a system as complex as the human brain, it is exceedingly difficult to pinpoint the cause of a deficit. The brain consists of a large number of processing subsystems which are extensively interconnected. Failure of one subsystem may have unexpected consequences in distant parts of the system: subsystems which are themselves in perfect working order may receive either degraded or entirely spurious input from the damaged subsystem. Consequently, the output of undamaged subsystems may be severely interrupted.\textsuperscript{34} In the case of autism, the mechanism responsible for the prediction and explanation of behaviour may itself be in perfect working order but may receive degraded or spurious inputs from an "upstream" subsystem which has been damaged. Note that this hypothesis not only accounts for the fact that autistic individuals have deficits consistent with the developmental failure of the capacity to predict and explain behaviour; it also suggests an account of those deficits which seem to have nothing to do with the capacity to predict and explain behaviour. For the damaged subsystem may be causally connected to other subsystems besides the mechanism responsible for our capacity to predict and explain behaviour. Those other subsystems are also likely to receive degraded or spurious input and thus produce degraded or spurious output.

One final brief remark before closing this Chapter. Both Gordon and Goldman have claimed that off-line simulation theory has unexpected consequences for the philosophy of mind. In particular they claim that if off-line simulation theory is true, the entire debate concerning eliminativism is utterly misguided.\textsuperscript{35} Since I think that off-line simulation theory is false, I have good grounds for ignoring Gordon and Goldman’s claim. But I am not prepared to leave it at that, for it seems to me that Gordon and Goldman are quite mistaken: the eliminativist debate remains alive and well even if off-line simulation theory turns out to be true. I leave that issue, though, for Chapter 6, which is devoted to assessing various arguments in and around the eliminativism debate. In the next Chapter I return to folk psychology.

\textsuperscript{33} See for example Leslie & Firth 1988.  
\textsuperscript{34} See Kosslyn & Koenig 1992: 108.  
(external), and offer an analysis of the (external) folk psychological concept of belief.
CHAPTER THREE

ANALYZING BELIEF

3.1 Introduction. We saw in Section 1.4 that folk psychology (external) is essentially Humean: action is caused by the interplay of beliefs and desires. We can ask, then, about the nature of beliefs and desires. How are they to be understood? In this Chapter I develop an analysis of belief. Before turning to that analysis, though, a few preliminary remarks are in order.

First, I focus largely on belief. This restriction is demanded by considerations of space. However, the analysis of belief developed below gives some indication of how analyses of other (external) folk psychological terms might be achieved.

Second, it is vital to bear in mind that we are seeking an analysis of a theoretical term of a folk theory. We are seeking an analysis of the (external) folk psychological concept of belief, not of, say, the concept of belief found in (or read into) computational psychology. An example of Robert Cummins' might help to make this point clear.1 A philosopher of physics may seek to analyze the concept of space found in Einsteinian physics. Alternatively, she may seek to analyze the concept of space found in Newtonian physics. These are separate projects that are likely to yield very different results, and the philosopher of physics must keep them firmly apart. Similarly, we must bear in mind that we seek an analysis of 'belief' as that term is used in folk psychology (external) rather than some other psychological theory.

The analysis of belief I endorse was pioneered by David Armstrong and David Lewis and is often called 'commonsense functionalism'.2 I begin by considering and rejecting two alternative approaches to analyzing the (external) folk psychological concept of belief: a behaviourist approach (Section 3.2) and a normative one (Section 3.3). I consider these alternative approaches to analyzing belief not simply for historical completeness but because, whilst flawed, they offer important insights into how a good analysis should go. Then, in Section 3.4, I outline Armstrong and

Lewis' commonsense functionalist approach to belief, emphasizing Lewis' utilization of Ramsey sentences. I close by considering a number of difficulties encountered by commonsense functionalism.

3.2 Philosophical Behaviourism. A variety of distinct doctrines can be identified under the general label 'behaviourism'. For our purposes, it will suffice to distinguish what I call 'methodological behaviourism' from what I call 'philosophical behaviourism'. Methodological behaviourism is, as the name suggests, a claim about methodology in psychology. Specifically, it claims that psychology should concern itself solely with generalizations relating sensory inputs to behavioural outputs. It should have no truck with inner states. Philosophical behaviourism, on the other hand, is a claim about the meanings of psychological terms. Specifically, it claims that the meaning of a psychological expression is to be given solely in terms of sensory inputs and behavioural outputs. Philosophical behaviourism thus wears its logical positivist origins on its sleeve. In what follows I will be concerned only with philosophical behaviourism – with the doctrine about meaning. Moreover, I will only be concerned to examine whether the meanings of (external) folk psychological expressions can be given solely in terms of sensory inputs and behavioural outputs. I will not be concerned with expressions drawn from other psychological theories.

How is a philosophical behaviourist account of the meanings of (external) folk psychological expressions supposed to go? Begin with the case of pain. The meaning of 'S is in pain' is to be analyzed in terms of sensory input clauses (S barked her shin; S touched a hot stove) and behavioural output clauses (S cried out; S withdrew her hand). To be in pain is to bark your shin or touch a hot stove and to cry out or withdraw your hand.6

Already we can see difficulties with this view. Think first about the behavioural output aspects of the analysis just given. How S reacts upon touching a hot stove will vary with S's circumstances. Say that S is hiding from Nazi agents who are close by. In that case she will not cry out upon touching a hot stove. The point is that, even in the case of pain, it is difficult to locate a set of typical associated behaviours. What we do when we are in pain varies with the circumstances in which we find ourselves. Notice that a similar remark can be made about the input side of

---

3 See Lycan 1990 for a more thorough classification of behaviourisms.
4 See Skinner 1933; Watson 1924/1925.
6 Strictly, the definition should contain no intentional expressions. However, the present formulation suffices for our purposes.
the analysis. If I touch a hot stove I will be in pain unless massively distracted by a more pressing problem (perhaps I have just noticed the tiger in the room or the gunman at the window).

Behaviourist analyses of belief face similar problems. This is particularly the case for beliefs not closely tied to the perceptual circumstances of the believer. Consider the belief that the economy is recovering. S's belief that the economy is recovering was caused by a chance remark overheard on a tram to the effect that unemployment is falling. R's belief that the economy is recovering was caused by a newspaper report to the effect that interest rates are declining. S's belief caused her to do nothing in particular; R's belief caused her to invest heavily in real estate.

It is tempting to respond to these problems as follows. In the case of pain, I remarked that S

":'ill not cry out upon touching a hot stove if she is hiding from Nazi agents who are close by. More accurately: S will not cry out upon touching a hot stove if she believes that there are Nazi agents near by and desires to avoid capture.

So it seems that we can push the analysis of pain through if we are willing to include in the analysandum not merely inputs and outputs but also internal states such as beliefs and desires. Similar remarks apply to belief. Given R's desire to make money and her belief that a good way to make money is to invest in real estate as the economy recovers, then her belief that the economy is recovering will typically cause her (ceteris paribus) to invest in real estate.

But the behaviourist who adopts this solution faces a dilemma. One horn of the dilemma is a charge of circularity: mental state terms now appear on both sides of the analysis. This would not be problematic if the definitions were recursive, with the recursion finally being grounded in definitions which were strictly behaviourist; that is, in definitions in which no mental state terms appeared. But no such definitions seem to be available: for no (external) folk psychological expression does a strictly behaviourist analysis seem possible. It will not be possible, then, to build up definitions of all mental state terms recursively, starting with those definitions which are strictly behaviourist.

There is a way, though, of avoiding the charge of circularity. We can take it that the various (external) folk psychological expressions are implicitly interdefined by the roles they play in folk psychology (external). This approach to the definition of theoretical terms was pioneered by Ramsey and developed by Lewis. But now the behaviourist is caught on the other horn of the dilemma. For the position arrived at by

---

7 Lewis 1970 and F. Ramsey 1931.
adoption of the Ramsey approach to the definition of theoretical terms is not behaviourist: (external) folk psychological expressions are no longer being defined solely in terms of sensory inputs and behavioural outputs. Rather, they are being defined in terms of sensory inputs, behavioural outputs, and other mental states. Ironically, the price of saving philosophical behaviourism from circularity is abandoning that very feature of philosophical behaviourism that made it behaviourist.8

Before closing this Section, let me briefly describe a pair of counterexamples to philosophical behaviourism offered by Hilary Putnam.9 Putnam's counterexamples are worth sketching for two reasons. First, they emphasize the problems with philosophical behaviourism which we have been examining in this Section. Second, we will have reason to consider Putnam's counterexamples again in Chapter 5, so a brief introduction to them now will not go amiss.

Taken together, Putnam's counterexamples challenge both the necessity and the sufficiency of the behaviourist analysis of pain. That is, they challenge both the claim that certain patterns of sensory input and behavioural output are necessary for pain, and the claim that certain patterns of sensory input and behavioural output are sufficient for pain. The challenge to the necessity of the behaviouralist analysis takes the form of a superstoic. A superstoic is sometimes in pain but never behaves as though she were. When, for example, the superstoic touches a hot stove, she experiences painful sensations but does not flinch or cry out. The superstoic thus challenges the necessity of pain behaviour for pain: the pain behaviour is absent but the pain present.

The challenge to the sufficiency of the behaviourist analysis takes the form of a perfect pain pretender. A perfect pain pretender never experiences pain but always produces pain behaviour when subject to the sort of sensory stimulation which typically causes pain. For example, when the perfect pain pretender touches a hot stove she always flinches and cries out, even though she experiences no painful sensations. The perfect pain pretender thus challenges the sufficiency of the behaviourist conditions on pain: the conditions are fulfilled but the pretend is not in pain.

So far we only have obtained counterexamples to the philosophical behaviourist analysis of pain. But with a little ingenuity Putnam's strategy can be extended to behaviourist analyses of any (external) folk psychological predicate. (In Chapter 5 we will witness a perfect intelligence pretender.) Notice also that Putnam's

8 Commonsense functionalists, lacking behaviourist scruples, accept the interdefinition of (external) folk psychological expressions. See Section 3.4.  
9 See Putnam 1980.
counterexamples nicely illustrate the dilemma given above. Consider the superstoic. She fails to generate typical pain behaviour even though she is in pain. Why? Because of her desire to repress all pain behaviour. If we were to avail ourselves of mental state terms, the superstoic's pain could be brought within the analysis. But now the dilemma re-emerges: the analyses of mental states are either circular or no longer behaviourist.

3.3 A Normative Approach. In this Section I want to consider – and reject – a normative approach to belief. According to this approach, belief attribution involves a very strong rationality assumption. I have used the label 'normative' because, on this view, we attribute to a cognitive system the beliefs and desires which it ought to have. Normative theorists need to make clear, of course, just what sense of 'ought' is operating here. The normative approach to the analysis of belief has its origin in the work of Quine and Davidson.\textsuperscript{10} It is most commonly associated, though, with Daniel Dennett.\textsuperscript{11} In what follows I will focus exclusively on Dennett's work.

Dennett identifies three stances from which a system's output can be predicted.\textsuperscript{12} (1) The physical stance, which views the system as a collection of the sort of particles quantified over by physics and subject to physical law. Given sufficient knowledge of the initial conditions of the system and of the relevant laws, the output of the system can, \textit{in principle}, be predicted. The qualification 'in principle' is important. For all but the most simple systems the physical stance is untenable: the calculations involved would overwhelm even the most prodigious calculator. Dennett's standard example is the chess-playing computer. Predicting the computer's next move on the basis of its physical structure described at, say, the atomic level, would in practice be impossible. (2) The design stance, which views the system as a product of design, where 'design' is understood to apply both to artefacts and to the products of natural selection. Think again of the chess playing computer. As just remarked, predicting the computer's output from the physical stance would in practice be impossible. However, by familiarizing ourselves with the machine's design we can accurately predict its output. Just as with the physical stance, though, the design stance is limited: if the chess playing computer's program is too involved, the design stance may become unworkable. (3) The intentional stance, which views the system as a locus of beliefs and desires. Those beliefs and desires which the

\textsuperscript{10} See Davidson 1980, 1984; Quine 1960.
\textsuperscript{12} See for example Dennett 1981: 3-6, 1987: 16-20.
system would have, were it fully rational, are attributed to the system, and then the system's output is predicted by practical syllogistics. Thus a chess playing computer should have beliefs which accurately reflect both the state of the board and the rules of the game, plus beliefs about its opponent's beliefs, etc. It should also desire to win, or at least to draw if outright victory is impossible. The beliefs and desires thus ascribed rationally dictate certain moves. Those are the moves predicted from the intentional stance. The intentional stance is relinquished only when the cognitive system breaks down—when it no longer does what it ought to do. At that juncture the physical stance must be adopted.

According to Dennett, then, beliefs and desires are states attributed to a cognitive system under the assumption that the system is fully rational. Dennett is at pains to emphasize, though, that beliefs and desires are to be understood purely instrumentally.13 Our deployment of the intentional stance carries no more ontological commitment than the modern navigator's deployment of Ptolemaic astronomy. In particular, by adopting the intentional stance and talking in terms of beliefs and desires we make no claims at all about the inner causes of behaviour. Again consider a chess playing computer. From the intentional stance we may attribute to it the belief that it should get its queen out early. Nevertheless, the computer's program may explicitly contain no such instruction, nor need there exist any identifiable part of the machine's hardware dedicated to getting the queen out early.14

It is important at this point to recall that we are seeking an analysis of the (external) folk psychological concept of belief. We are not seeking either an idealization of the (external) folk psychological concept of belief nor (what may amount to the same thing) an analysis of 'belief' as that expression is used in some other psychological theory. Dennett makes it quite clear, though, that his project is not ours. Rather than offering a thorough analysis of (external) folk psychological concepts—Dennett describes that project as 'anthropological'—he seeks an idealized account of belief that captures 'the parts of folk psychology worth caring about'.15 For Dennett, those are the parts of folk psychology (external) that present us as agents, free and responsible. But Dennett is also concerned to show that his idealized notion of belief is not too far from the (external) folk psychological one. He therefore

---

14 I am relying here on Dennett's own insistence on his instrumentalism. Moreover, that is the position he is widely regarding as having adopted (see for example Fodor 1990a). Dennett has, though, recently repudiated the terminology of contemporary philosophy of mind. (See Dennett forthcoming.) It is an open question, then, whether he would presently accept being labelled an 'instrumentalist'.
offers at least two arguments which attempt to close the gap between his idealized notion of belief and the one found in folk psychology (external).

**First argument.** Dennett's first argument draws attention to our reluctance to attribute irrationality to those around us. It is at this point that Dennett's indebtedness to Quine is clearest. In *Word And Object*, Quine reflects on the possibility of radical translation. In particular, he considers how we would respond if, according to our present translation manual, our interlocutor asserted both p and not-p. Quine suggests that, faced with such apparent irrationality, we would modify our translation manual so as to render the speaker's utterances consistent. In other words, *the assumption of rationality is a precondition of translation*.16

As we have seen, Dennett makes a similar claim. For Dennett, the assumption of rationality *is a precondition on the attribution of beliefs and desires*. Attributions of irrational beliefs and desires are avoided by making suitable revisions at other points. Thus, rather than attributing to S the belief that p and not-p, we assume that she has not heard; that she does not speak our language; that she has failed to notice a crucial detail; and so forth.17 The attribution of rationality is itself immune from such revisions: it is the bedrock upon which propositional attitude ascription rests.

There is some truth in Dennett's claim. We are reluctant to attribute obviously inconsistent beliefs to our interlocutors. Rather than attributing to S the belief that p and not-p, we tend to assume that she has not heard or has failed to notice something about her environment. Nevertheless, we do at times attribute irrationality to those around us. When one of my logic students commits the fallacy of affirming the consequent I do not assume that she has misunderstood the question. Rather, I think that she has failed to reason properly. Moreover, even if she persistently has trouble with her logic assignments I do not shy away from the intentional stance in my dealings with her. I still attribute to her various beliefs and desires. In other words, the rationality assumption can be revised without retreating from (external) folk psychological belief/desire ascription.18 It follows that the propositional attitudes Dennett ascribes from the intentional stance are not the propositional attitudes found in folk psychology (external). For Dennett's idealized notions of belief and desire are attributable only under a very strong rationality assumption. Whilst folk psychology

---

18 The rationality assumption, though, is not infinitely revisable. Faced with extreme irrationality, folk psychology (external) tends to withdraw belief/desire ascriptions.
(external) has some sympathy with a weak rationality constraint, it has no truck with Dennett's strong assumption.

*Second argument.* Dennett argues that it is safe to make a strong rationality assumption because natural selection will in fact have brought it about that we are all pretty much fully rational. After all, Dennett reasons, more rational organisms will have a selective advantage over less rational ones and therefore natural selection will, over time, tend to result in rational organisms.\(^{19}\)

Straight away we can see that this argument requires, at the very least, refining. The influenza virus is a product of natural selection. Indeed, it is pretty good at survival and reproduction — far too good from our perspective. Nevertheless, the influenza virus is not a paradigm of rationality. So the argument from natural selection to rationality at least needs tidying up. Here is one attempt at improvement. It might be suggested that there exists a number — perhaps an indefinite number — of survival/reproduction "strategies". One strategy involves the formation and manipulation of representations. Those organisms which survive in virtue of their capacity to form and manipulate representations will be under selection pressure to maximize rationality. It follows that since our survival is due to our capacity to form and manipulate representations, we will tend to be highly rational. The influenza virus, which has not adopted the survival/reproduction strategy of forming and manipulating representations, is under no such pressure. Hence natural selection can be cited as a reason for our being pretty much maximally rational whilst the influenza virus remains dumb.

There is a lot wrong with this attempt to patch up Dennett's argument. For one thing, in the absence of a definitive account of representation, it is hard to see why at least some of a virus' activities do not count as forming and manipulating representations. After all, according to one theory of representation — the indicator theory — a state R represents another state S if and only if R covaries with S.\(^{20}\) And surely some internal states of influenza viruses covary with external states. So at least on this sense of 'representation' influenza viruses form and manipulate representations. Of course it could be objected that the indicator theory of representation is inappropriate in this context, or that, whilst the influenza virus does form and manipulate representations, it has not survived *in virtue* of doing so — that it has not adopted the formation and manipulation of representations as a survival/reproduction strategy. It is not obvious, though, that either of these ways of

---

\(^{19}\) See Dennett 1981: 8, 1987: 33.

\(^{20}\) I discuss the indicator theory at length in Ch. 4.
defending Dennett will be successful. Moreover, two other powerful objections to
Dennett's argument from natural selection exist. The first turns on the idea of local
optima in an adaptive landscape. The second draws attention to the costs involved
in maximizing rationality.

Imagine a landscape that includes a pair of hills, one higher than the other.
You are halfway up the lower hill, but would prefer to ascend the higher one. There
is no way to walk to the top of the higher hill from where you are without descending
into the valley. The lower hill that you are on is a local optimum - at least with respect
to altitude. Of course, normally you are free to walk down into the valley and then
ascend the other hill - the global optimum. But if for some reason you were only
permitted to walk in an upwards direction, you would be committed to climbing the
local optimum even if you preferred to climb the global one.

A similar phenomenon threatens in natural selection. Think of altitude in the
landscape just described as representing optimality of design - better designed
organisms are located higher in the landscape than less well designed organisms. And
think of the restriction on walking downwards as representing selection pressure
towards better design. Then once again there exists the threat of capture by local
optima: an organism's descendants may come to occupy a local rather than a global
optimum. It follows that even if natural selection favours more rational organisms -
even if being more rational means being better designed - an organism's descendants
may not end up globally optimally rational. Rather, they may end up merely
occupying a local rationality optimum. So Dennett cannot assume that natural
selection guarantees that we are pretty much ideally rational. The ideally rational
organisms - if they exist at all - may be across the valley on the higher peak.

Dennett in effect assumes that the adaptive landscape has only one peak on it.
But that assumption cannot be made without justification. It seems fairly certain that
the adaptive landscapes of some actual organisms have at least two local optima, with
organisms located on both peaks. Thus consider the human eye. Light passing
through the lens strikes the light-sensitive cells on the retina. These cells are
extensively interconnected. From a design point of view one would expect the
interconnecting "circuitry" to be behind the photo-sensitive cells. But it is not. Light
coming from the lens must pass through the interconnecting circuitry before it strikes
the photosensitive cells. There is nothing necessary about this arrangement. Organisms exist in which the circuitry is located behind the photo-sensitive cells.

---

21 The idea of local optima in an adaptive landscape was introduced by Sewell Wright and is well
Indeed, some of the cephalopods have eyes remarkably similar to the human eye except that the circuitry is behind the photo-sensitive cells of the retina. The human eye thus contains a rather striking design flaw. But getting from an eye like ours to an eye like the cephalopods' is, from a natural selection perspective, close to impossible. The intermediate steps involved in the transition would be far too disadvantageous to us. In short, we are stuck on a local optima with the global optima across the valley occupied by the cephalopods and quite out of reach. Dennett cannot just assume that the situation with regard to rationality is not the same. He must demonstrate that the adaptive environment for rationality does not contain more than one local optima.

The argument we have just considered assumes that natural selection does in fact favour rationality – that there exists a selection pressure towards optimal rationality. But it is not clear that we should concede even that much to Dennett. Consider two organisms, A and B. A is perfectly rational. She carefully weighs up all available evidence and meticulously considers her best course of action. She is invariably right. B is much less rational. She makes a cursory survey of the available evidence and jumps to a conclusion. She is sometimes wrong. She is, though, very fast. She obtains her conclusion in milliseconds. A, on the other hand, is very slow. It often takes her minutes to reach a conclusion. Which organism will natural selection favour? That depends. In an environment in which speed is irrelevant but cognitive errors costly, A will (ceteris paribus) do better. On the other hand, in an environment in which cognitive mistakes are not too serious but speed is essential, B will (ceteris paribus) do better. In other words, there is no simple argument from natural selection to optimal rationality.

Maximizing rationality is expensive. From a natural selection perspective it may be worth economizing on reasoning power and investing elsewhere the resources thus saved. In the example just given we saw a trade off, in organism B, between rationality and speed. B gets things wrong but at least she gets an answer fast, and in some environments that may confer on her a selective advantage. This point becomes particularly acute when we reflect on just how expensive ideal rationality would be. Christopher Cherniak has made this point in a particularly dramatic way. Consider testing a set of 138 propositions for tautological consistency by the truth table method.

---

22 See Dawkins 1982: 40.
23 Sewell Wright has argued that genetic drift can help free populations from local optima. However, the general point against Dennett remains. (See Wright 1980.)
24 See Stich 1990, especially Section 3.2.
Even assuming that each line of the table could be checked in the time it takes a light wave to traverse a distance equal to the diameter of a proton, it would take longer than the entire history of the universe from the big bang to the present to check the entire table. So, if ideal rationality involves tautological consistency, ideal rationality is going to be very expensive indeed. Of course, Dennett could insist that rationality, as he understands it, involves something less than the sort of ideal rationality canvassed here. But if so he owes us an account of just what he has in mind, and of how his new sense of 'rationality' relates to the intentional stance.

Note also that rationality can be costly in terms other than time. Effort spent building a brain is effort not spent building something else — and that "something" may have survival/reproductive advantages. Indeed, there exist organisms which cannibalize their own nervous systems when their environmental circumstances demand it. Thus the newborn sea squirt utilizes its nervous system to swim and locate food. Eventually it finds a suitable location and fixes itself to a rock. At that point it consumes its now redundant nervous system. For the mature sea squirt, a nervous system has become an unacceptable expense. I am not, of course, claiming that the immature sea squirt should be regarded as rational. The point is, rather, that the biological basis of rationality is costly, and natural selection may favour other more economical solutions to the problems posed by the organism's environment.

We are seeking an analysis of the (external) folk psychological concept of belief, and we are trying to determine if Dennett has provided such an analysis. Dennett himself claims to have provided only an idealized conception of the propositional attitudes, a conception which allegedly captures 'those parts of folk psychology worth caring about'. He believes, however, that his idealized conception of belief is not too far from the folk conception. In particular, he claims that the strong rationality assumption underpinning the intentional stance is acceptable for two reasons. We have surveyed those reasons and found them wanting. Dennett's conception of belief is not the (external) folk psychological conception, not even to a first approximation.

Moreover, Dennett's instrumentalist claims are at odds with folk psychology (external). As we have seen, according to Dennett our belief and desire talk carries no ontological commitment. In particular, Dennett denies that beliefs and desires are internal causes of behaviour. Folk psychology (external), on the other hand, is realist.

---

27 In Section 4.1 I briefly sketch a further argument against making using of rationality assumptions when ascribing content. There my target is Lewis, but the point applies equally well to Dennett.
about the propositional attitudes. As we saw in Section 1.4, folk psychology (external) does take the propositional attitudes to be internal causes of behaviour. We have two reasons, then, for denying that Dennett has provided us with an adequate analysis of the (external) folk psychological conception of belief: Dennett's rationality assumption and his instrumentalism.

3.4 Commonsense Functionalism. In Section 3.2 we examined at some length the failure of behaviourist analyses of (external) folk psychological expressions. We saw, for example, that the (external) folk psychological predicate 'is in pain' cannot be analyzed simply in terms of sensory inputs (S touched a hot stove) and behavioural outputs (S cried out). For say that S is hiding from the Nazis. If she believes that Nazi agents are nearby and desires to avoid capture she will not cry out upon touching a hot stove. Nevertheless, she will be in pain.

It is tempting to avoid this difficulty by introducing S's beliefs and desires into the analysis, for it is true (ceteris paribus) that S will cry out upon touching a hot stove if she does not believe that the Nazis are nearby and desire to avoid capture. But, as we have seen, behaviourism thus amended faces a dilemma: it is either circular or no longer behaviourist. The charge of circularity arises because (external) folk psychological expressions now appear on both sides of the analysis. This charge can be evaded by embracing Ramsey's idea that the theoretical terms of a theory are interdefined. But now the other horn of the dilemma looms. The resulting interdefinitions are not behaviourist. (External) folk psychological expressions are no longer being defined solely in terms of relations between sensory inputs and behavioural outputs. Rather, they are being defined in terms of relations between sensory inputs, behavioural outputs, and other mental states.

The behaviourist cannot embrace interdefinition without foregoing her behaviourism. But those of us without verificationist scruples might cheerfully adopt Ramsey's idea of interdefinition and apply it to folk psychology (external). On this view, then, a mental state is a state which bears certain causal relations to sensory inputs and behavioural outputs, and to other mental states. That is, an (external) folk psychological state is the occupant of a causal role involving sensory inputs, behavioural outputs, and other mental states. I call this approach to the analysis of (external) folk psychological states commonsense functionalism.28

---

28 This position has also been called 'analytic functionalism'.

59
Commonsense functionalism is immediately attractive for at least three reasons. First, it seems to capture what is right about behaviourism without falling into the traps that beset behaviourism. For there is something right about behaviourism. Stimuli and behaviour do have a role to play in the analysis of (external) folk psychological mental states, it is just that they are not the whole story. Second, the analyses offered by commonsense functionalism make no claims at all about the occupants of the specified roles. To be in pain is to occupy the (external) folk psychological pain role, not to be made of this stuff or that. It follows that a variety of cognitive systems with diverse physical make-ups could be in pain. And indeed folk psychology (external) is comfortable with the idea that a variety of cognitive systems could experience pain: we are happy to ascribe pains to dogs and Martians. Finally, given that certain neuroscientific claims turn out to be true, commonsense functionalism facilitates the identification of mental states with neural states. The identification proceeds via the transitivity of identity. Thus, commonsense functionalism asserts that

\[
\text{Mental state } M = \text{the occupant of role } R, \\
\text{whilst neuroscience may, in the fullness of time, be able to establish that} \\
\text{The occupant of role } R = \text{neural state } N.
\]

By the transitivity of identity we can conclude that

\[
\text{Mental state } M = \text{neural state } N.^{31}
\]

Clearly these identities may need to be relativized to a species. In humans role \( R \) might be occupied by neural state \( N \), in dogs by neural state \( N^* \), and in Martians by state \( S \) of silicon chip 924 alpha. Indeed, if there is considerable variation within species, these identities might have to be relativized to individuals.\(^{32}\)

---

\(^{29}\) Indeed, the pain role might even be occupied by a non-physical cognitive system. Adopting an expression of Ryle's, J.J.C. Smart has called such analyses topic neutral.

\(^{30}\) Typically, such identifications are taken to be token-token: each token mental state is identified with a token neural state. For an important dissenting position see Jackson, Pargetter and Prior 1982.

\(^{31}\) See Lewis 1966, 1972. See also Jackson forthcoming.

\(^{32}\) Lewis thinks otherwise. See his 'Mad Pain And Martian Pain' (Lewis 1980).
The analysis of (external) folk psychological states we are considering here is most commonly associated with David Lewis. It is worth spending some time looking more closely at Lewis' approach. The core of Lewis' approach is an application of Ramsey's idea of interdefining theoretical terms to folk psychology (external).33 Ramsey claimed that the meaning of a theoretical term was given by the role that term played in the relevant theory. By applying this idea to (external) folk psychological expressions, Lewis is committing himself to the theory theory (external); that is, he is taking our everyday talk about mental states to implicitly constitute a theory of mind. Lewis does not explicitly defend the theory theory (external). Rather, he shows how smoothly Ramsey's approach to theoretical terms can be applied to our everyday talk about mental states, thus making it overwhelmingly plausible that the theory theory (external) is true.

According to Ramsey, the meaning of a theoretical term is given by its use in the relevant theory. It follows that, if the theory theory (external) is true, the meanings of the various (external) folk psychological expressions are given by the collected commonsense platitudes about mental states. Lewis therefore begins by assembling our commonsense platitudes about the mind:

Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. Perhaps we can think of them as having the form: 'When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses'. Add also all the platitudes to the effect that one mental state falls under another — 'toothache is a kind of pain', and the like. Perhaps there are platitudes of other forms as well. Include only platitudes which are common knowledge among us — everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that names of mental states derive their meaning from these platitudes.34

As Lewis remarked in the passage quoted, our everyday platitudes about mental states take various forms. Some platitudes relate sensory inputs and mental

33 See F. Ramsey 1931.
34 Lewis 1972: 256.
states to behavioral outputs and other mental states. Thus the collection of platitudes will include platitudes of the form:

\[(1) \ (x)(S_1x \& M_1x \rightarrow M_2x \& B_1x).\]

(1) says that anyone in mental state \(M_1\) who experiences sensory input \(S_1\) will move into mental state \(M_2\) and produce behavioral output \(B_1\). Lewis' first step is to form the conjunction of all the assembled platitudes, \(T\). The conjunction will contain the names of all the (external) folk psychological mental states \(M_1, M_2, \ldots, M_n\). It will also include various expressions which do not denote mental states. Lewis calls these the '0-terms'.\(^{35}\) We can therefore express \(T\) by:

\[(2) \ T[t],\]

where \(t\) denotes the \(n\)-tuple of mental state terms \(<M_1, M_2, \ldots, M_n>\). Lewis then uniformly replaces the (external) folk psychological mental state expressions \(M_1, M_2, \ldots, M_n\) with free variables \(m_1, m_2, \ldots, m_n\), and prefixes the appropriate existential quantifier:

\[(3) \ \exists x \ T[x],\]

where \(x\) denotes the \(n\)-tuple \(<m_1, m_2, \ldots, m_n>\). (3) is the Ramsey sentence for folk psychology (external). It says that there exists an \(n\)-tuple of entities which realizes \(T\); that is, there exists an \(n\)-tuple of entities which occupy the roles specified by our everyday talk about mental states. Note that (3) contains only 0-terms (plus logical connectives).

Lewis takes it that the theoretical terms of a theory are denotationless if the \(n\)-tuple is multiply realized. He therefore constructs the modified Ramsey sentence for \(T\) by replacing the existential quantifier in (3) with a unique existence quantifier:

\[(4) \ \exists! x \ T [x].\]

\(^{35}\) Lewis is at pains to emphasize that the O-terms are not necessarily observation terms in, say, Carnap's sense. They are simply terms other than those denoting mental states. See Lewis 1972: 250.
(4) says that there exists a unique n-tuple of entities which satisfies T. Lewis now constructs what he calls the 'modified Carnap sentence' of T, which says that if T is uniquely realized then the (external) folk psychological expressions M₁, M₂, ..., Mₙ name the unique realizors of T:

\[(5) \exists! (x) \ T [x] \rightarrow T [t].\]

We have seen that Lewis takes it that the theoretical terms of a theory are empty unless T is uniquely satisfied. Lewis also takes it that the theoretical terms of a theory are empty if no n-tuple satisfies T:

\[(6) \neg \exists! x \ T [x] \rightarrow t = *, \]

where \( t = * \) means that each \( tᵢ \) is denotationless. Taken together, (5) and (6) imply:

\[(7) t = ! x \ T [x].\]

That is, the mental state terms M₁, M₂, ..., Mₙ name the entities which uniquely realize the roles specified by the conjunction of platitudes T.

As we have just seen, Lewis takes it that the theoretical terms of a theory are denotationless if the n-tuple of free variable \( <m₁, m₂, ..., mₙ> \) is multiply realized. I argued above, though, that one of the strengths of commonsense functionalism is that it makes no claims about the occupants of the folk-specified causal roles characteristic of mental states. That is, commonsense functionalism respects the (external) folk psychological intuition that mental states can be multiply realized. I therefore propose to derive

\[(5') \exists (x) \ T [x] \rightarrow T [t] \]

from the Ramsey sentence (3), and hence obtain

\[(7') t = x \ T [x].\]

(7') says that the mental states M₁, M₂, ..., Mₙ name the entities which realize the roles specified by the conjunction of commonsense platitudes, T, and allows that they may be multiply realized.
As Lewis himself remarks, this approach to the analysis of (external) folk psychological terms amounts to treating them as definite descriptions. We can therefore identify three elements in Lewis' approach to (external) folk psychological expressions. (1) Lewis is committed to the theory theory (external). He must accept that our everyday talk about mental states constitutes a theory of mind. (2) He is committed to a Russellian definite description theory of meaning, at least for the expressions of folk psychology (external). That is, he is committed to the claim that the meanings of 'belief', 'desire', and so forth are given by an associated definite description. (3) He is committed to the claim that (external) folk psychological mental states are to be identified with the occupants of certain causal roles mediating between sensory inputs, mental states, and behavioural outputs. These three elements are not independent of one another. In particular, the third follows naturally from the first two. As we have seen, if we adopt the theory theory (external) and a definite description theory of meaning for theoretical terms, we can straightforwardly identify (external) folk psychological mental states with the occupants of certain causal roles.

Lewis' approach to commonsense functionalism is very attractive. However, as just noted, it is committed to a definite description theory of meaning for theoretical terms – or at least it is committed to a definite description theory for the theoretical terms of folk psychology (external). The definite description theory of meaning is, however, controversial. Standardly, it is contrasted with the causal-historical theory of reference according to which the referent of a name is fixed by a causal-historical chain linking each meaningful utterance of that name with the object to which it refers. The chain begins with a naming event – Aristotle's parents called him 'Aristotle' – and passes from speaker to speaker. Thus I succeed in referring to Aristotle because I learnt the name 'Aristotle' from someone who learnt the name 'Aristotle' from someone who . . . . and so on back to the original naming event. A slight development of the theory yields an account of the meanings of kind terms. Again there is an original naming event: someone points to a sample of, say, water and says 'water'. And again contemporary usage of the term succeeds in referring because of a causal-historical chain linking present usage with the naming event. However, the name 'water' applies not just to the original sample but to the entire kind to which the original sample belongs.

36 See Lewis 1972: 252.
37 See Russell 1957. For a discussion and critique see Devitt & Sterelny 1987: 39-43.
38 See Donellan 1972; Kripke 1980; Putnam 1975. For an exposition and defense of the causal-historical theory see Devitt & Sterelny 1987, especially Ch. 4.
We can apply the causal-historical theory of reference to (external) folk psychological expressions. On this view, my use of the term 'belief' picks out the kind beliefs in virtue of a causal-historical chain linking my utterance to a member of that kind via an original naming event. In the case of 'water', the kind picked out is a natural kind, namely, H2O. In the case of 'belief', the kind picked out is a functional kind, namely, the occupant of a certain causal role. Note that allowing the causal-historical chain to terminate at a functional rather than a natural kind is not a concession made in an ad hoc manner specifically to deal with (external) folk psychological expressions. An analogous story must be told for other functional terms such as 'carburetor' and 'pain killer'.

Which theory of reference is correct? In Chapter 6 I endorse a variant of the definite description theory of reference – the cluster theory. According to the cluster theory, the referent of a term is fixed not by a single definite description but by a "cluster" of descriptions; that is, by a disjunction of definite descriptions. The cluster theory enjoys a considerable advantage over the classical definite description theory in that it allows some elements of the description to be mistaken without reference failure occurring. Thus, say that the definite description, 'the pupil of Plato, teacher of Alexander, and author of the Nichomachean Ethics', is associated with the referring term 'Aristotle', but that in fact Aristotle did not write the Nichomachean Ethics. Then, by the classical definite description theory, 'Aristotle' does not refer to Aristotle. However, if the description, 'the pupil of Plato and teacher of Alexander', is sufficient to pick out Aristotle, by the cluster theory 'Aristotle' will refer to Aristotle. That is, one of the disjoined descriptions – the one that makes no mention of the author of the Nichomachean Ethics – will secure the reference of 'Aristotle'.

As we might expect, the Lewis/Ramsey approach to theoretical terms is compatible with the cluster theory of reference. Thus, in the case of (external) folk psychological terms, rather than conjoining all the commonsense platitudes about the mind we can disjoin all the conjunctions of most of the commonsense platitudes about the mind. In Chapter 6 I will argue that only the cluster theory of reference allows us to make sense of our ontological practices. In particular, only the cluster theory allows us to make sense of the fact that sometimes we eliminate the posits of a false theory whilst at other times we retain them. Moreover, I believe that the argument most commonly cited against the cluster theory – Putnam's Twin Earth argument – is

---

Twin Earth is an exact replica of Earth except that the stuff that falls from the sky, runs in rivers, fills the lakes, and pours from the tap on Twin Earth is XYZ not H2O. Whilst XYZ can be distinguished from H2O by chemical analysis, its run-of-the-mill causal properties are identical to those of H2O. There exists a widespread intuition that when an inhabitant of Earth uses the expression 'water' she refers to H2O, whereas when her Twin on Twin Earth uses the expression 'water' she refers to XYZ. Note, though, that the cluster of descriptions associated with the expression 'water' by the inhabitants of Earth seems to be identical to the cluster of descriptions associated with the expression 'water' by the inhabitants of Twin Earth. On both Earth and Twin Earth the expression 'water' is associated with the description: 'falls from the sky, runs in rivers, fills the lakes, and pours from the tap'. Since both H2O and XYZ satisfy that description, it follows that the cluster theory is committed to the claim that the expression 'water', when used by either an inhabitant of Earth or an inhabitant of Twin Earth, refers to H2O-or-XYZ, thereby contradicting the widespread intuition that 'water' on Earth refers to H2O whilst 'water' on Twin Earth refers to XYZ.

Note, though, that if the cluster of definite descriptions associated with the expression 'water' by the inhabitants of Earth included the description 'is H2O', whilst the cluster of definite descriptions associated with the expression 'water' by the inhabitants of Twin Earth included the description 'is XYZ', the problem just sketched for the cluster theory would disappear. For then only H2O would satisfy the description associated by the inhabitants of Earth with the expression 'water', whilst only XYZ would satisfy the description associated by the inhabitants of Twin Earth with the expression 'water'. Two responses to this move come to mind. First, it might be argued that the inhabitants of Earth who do not realize that (Earth) water is H2O still refer to H2O when they use the term 'water'. I am inclined to think that the intuition that such people refer to H2O rather than H2O-or-XYZ is not particularly strong. However, for present purposes I propose to take that intuition seriously. Second, it might be argued that, since Earth and Twin Earth are in effect identical, the cluster of descriptions associated with the expression 'water' on Earth must be the same as the cluster of descriptions associated with the expression 'water' on Twin Earth. Thus, the cluster theorist cannot avoid the problem presented by the Twin Earth example by positing one cluster of definite descriptions for the inhabitants of Earth and another for the inhabitants of Twin Earth.

40 See Putnam 1975.
Indexical expressions like *here* which non-rigidly pick out the world in which they are uttered offer a way to avoid the difficulties which the Twin Earth example presents to the cluster theory. Thus, if the cluster of definite descriptions associated with the expression 'water' by both the inhabitants of Earth and the inhabitants of Twin Earth included a description to the effect that water is a natural kind found around here in abundance, then XYZ would fail to satisfy the cluster of descriptions on Earth and H2O would fail to satisfy it on Twin Earth. It follows that, 'water' refers to H2O when used by the inhabitants of Earth, but refers to XYZ when used by the inhabitants of Twin Earth.

What about uses of the expression 'water' by inhabitants of either Earth or Twin Earth who do not include in their associated cluster a description to the effect that water is a natural kind found around here in abundance? At this point I am prepared to bite the bullet and declare that for such people 'water' refers to H2O-or-XYZ. The expression has become ambiguous.41

As we have seen, the Lewis/Ramsey approach to theoretical terms commits Lewis to the definite description theory of reference or to its refinement, the cluster theory. I am prepared to grant Lewis the cluster theory of reference, and thus I am prepared to grant his package (1)–(3), above. I realize, however, that the cluster theory is not fashionable at present, and so it is important to note that the claim that (external) folk psychological states are the occupants of certain causal roles can be divorced from Lewis' controversial theory of meaning. Advocates of the causal-historical theory are encouraged to read on!

I want to endorse commonsense functionalism; that is, I think that (external) folk psychological mental states are loci in a causal network. The causal network is narrow in that it is characterized in terms of states – sensory inputs, behavioural outputs, and other mental states – which do not extend beyond the skin. And the causal network is potential in that a mental state is characterized by its potential rather than its actual causal relations. Thus, my desire for a piece of chocolate cake is partly characterized by the causal relations it would have if I were to come to believe that there is a piece of chocolate cake in the cupboard. However, the claim that mental states are loci in a narrow, potential, causal network presents three difficulties.

First difficulty. Consider Stich's example about Mrs T. As a young woman, Mrs T took a lively interest in politics. She was able to discourse, for example, on President McKinley's assassination and its consequences for American politics.

41 My discussion of the Twin Earth case owes a great deal to Lewis forthcoming.
Regrettably, a few years ago, Mrs T fell victim to a degenerative brain disorder which caused her to forget more and more of her beliefs. This process continued until today she can remember practically nothing. Mrs T is still capable, though, of correctly answering the question 'What happened to McKinley?'. However, she cannot state whether an assassinated person is dead, whether McKinley is dead, what dying is, or whether she herself is dead.

There is a final twist to Stich's tale. Although Mrs T holds practically no beliefs, she has lost none of her capacity to reason. Thus, if she did come to believe that p and that if p then q, she would (ceteris paribus) come to believe that q. Of course, Mrs T's capacity to reason is entirely unexercised, for she has no beliefs to reason about.\(^\text{42}\)

Now Mrs T's case is very odd. Such people are uncommon and we may be at a loss to explain how they came to be as they are. But note that folk psychology (external) is quite clear on at least this much: Mrs T does not believe that McKinley was assassinated. Mrs T is merely mouthing the words. She no more believes that McKinley was assassinated than my answering machine believes that I cannot come to the phone right now.

Call the state within Mrs T which causes her to respond correctly to the question 'What happened to McKinley?', state $S$. Notice that, since Mrs T's reasoning apparatus is, by hypothesis, in perfect working order, the narrow, potential, causal relations of state $S$ are identical to the narrow, potential, causal relations of the belief she had as a young woman that McKinley was assassinated. It follows that, if beliefs are loci in a narrow, potential, causal network, then Mrs T believes now that McKinley was assassinated. In other words, state $S$ is the belief that McKinley was assassinated. Since, as we have seen, folk psychology (external) denies that Mrs T believes that McKinley was assassinated, (external) folk psychological beliefs cannot be identified with loci in a narrow, potential, causal network. Rather, beliefs are individuated in part by their actual causal relations to other mental states. We therefore need to reformulate our analysis of belief. Thus, a belief is a locus in a narrow, potential, causal network which has, moreover, at least some actual causal relations to other mental states.

Two comments are in order. First, notice that we arrived at the reformulated analysis by reflecting on the commonsense platiitudes elicited by Stich's story. It is thus a commonsense functionalist analysis. The earlier analysis was simply mistaken:

\(^{42}\) See Stich 1983: 55.
it arose from an incomplete survey of folk platitudes about the mind. Second, a consequence of including actual causal relations in the analysis of belief is that commonsense functionalism is committed to what I will call ontological holism— the doctrine that an intentional system cannot consist of just one propositional attitude.

Second difficulty. Recall Putnam’s Twin Earth cases, presented earlier in this Section.43 Twin Earth is a molecule-for-molecule replica of Earth with one exception: the stuff that falls from the sky, runs in the rivers, fills the lakes, and pours from the tap on Twin Earth, is XYZ not H2O. Oscar, an inhabitant of Earth, truthfully says that the stuff in the glass in front of him is water. Twin Oscar, Oscar’s molecule-for-molecule replica on Twin Earth, also says that the stuff in the glass in front of him is water. Do Oscar and Twin Oscar have the same belief about the stuff in the glass? If we focus on behavioural dispositions then there is a strong inclination to say that Oscar and Twin Oscar have the same belief. After all, we explain why they both say that the stuff in the glass is water by saying that they both believe that the stuff in the glass is water. On the other hand, if the focus is on the object of belief then there is a strong temptation to say that they have different beliefs: Oscar has a belief about H2O; Twin Oscar has a belief about XYZ.

It is sometimes said that Oscar’s belief is narrowly identical to Twin Oscar’s but not widely identical. If we say that two beliefs are narrowly identical if and only if they have identical narrow causal relations, then commonsense functionalism can straightforwardly account for the narrow identity of Oscar and Twin Oscar’s beliefs. We still require, though, an account of the wide divergence of their beliefs.

Third difficulty. We need an account of the semantic properties of (external) folk psychological propositional attitudes. That is, we need a theory of content. Moreover, that theory of content must be compatible with commonsense functionalism: it must show how semantic properties can be attributed to the occupants of the causal roles characteristic of beliefs and desires. I address this difficulty in Chapter 4.

---

CHAPTER FOUR

THE PROBLEM OF CONTENT

4.1 Introduction. We saw in Section 1.4 that (external) folk psychological beliefs and desires have semantic properties: beliefs are true or false; desires satisfied or not. Standardly, the semantic properties of propositional attitudes are accounted for by taking believing and desiring to be relations to representations.\(^1\) A belief is true if and only if the represented state of affairs pertains; false otherwise. Similarly, a desire is satisfied if and only if the represented state of affairs pertains; unsatisfied otherwise. We therefore require an account of mental representation. In virtue of what is my belief that \(p \) about \(p\)? This is the problem of content.

There exist a number of putative solutions to the problem of content. Before turning to possible solutions, though, it is worth trying to get the problem into sharper focus by considering exactly what the fundamental unit of semantic evaluation is supposed to be. We can identify at least three candidates.\(^2\) (1) Entire intentional systems, with propositional attitudes having content only derivatively. (2) Propositional attitudes, which themselves have no semantically-relevant internal structure. (3) What we might call 'subattitudinal states'; that is, components out of which propositional attitudes are assembled.

According to option (1), the fundamental unit of semantic evaluation is the entire intentional system. Thus Lewis remarks that '[t]he contentful unit is the entire system of beliefs and desires... Its content is defined, insofar as it is defined at all, by constitutive rationality on the basis of its typical causal role'.\(^3\) Following Fodor, I want to suggest that it is wrongheaded to appeal to rationality when ascribing content.\(^4\) The difficulty is that a necessary condition on being rational is having contentful states: rationality involves having contentful states which are related to one another and to behaviour in certain ways. Thus, one is rational if, say, one's beliefs are related to one another in accordance with the first order calculus, and one's beliefs

\(^1\) Note that I do not claim that, standardly, the semantic properties of the propositional attitudes are accounted for by taking believing and desiring to be relations to *sentence tokens in the language of thought*. The language of thought hypothesis is but one way of spelling out the nature of mental representation.


\(^3\) Lewis forthcoming.

\(^4\) See Fodor 1990a: 8. Fodor's target is Dennett rather than Lewis, but the point generalizes to any attempt to appeal to rationality when ascribing content. See Section 3.3, above, for a discussion of Dennett's appeal to rationality in the analysis of belief.
and desires are related to one's actions in accordance with the practical syllogism. It follows that

the rational systems are a species of the intentional ones rather than the other way around. If that is so, then it is misguided to appeal to rationality in the analysis of intentionality since, in the order of explanation, the latter is the more fundamental notion.\footnote{Fodor 1990a: 8.}

Without appealing to rationality, however, it is difficult to see how option (1) can be exercised.

What of the remaining two options? According to option (2), the unit of semantic evaluation is the propositional attitude which itself has no semantically-relevant internal structure. So-called functional role semantics stands alone in endorsing option (2). Functional role semantics exploits the fact that the causal relations of propositional attitudes approximately mirror their semantic relations. Thus, not only is it the case that there is a \textit{semantic} relation between the propositions \((p \& q)\) and \(p\); it is also the case that there is – typically – a \textit{causal} relation between S's believing that \((p \& q)\) and her believing that \(p\). According to functional role semantics, content is ascribed to the propositional attitudes of an intentional system so as to preserve the isomorphism between the causal network to which the propositional attitudes belong and the semantic network to which their propositional objects belong.

The trouble with functional role semantics is that it leads to an unacceptable degree of semantic underdetermination. With a little ingenuity, we can construct indefinitely many semantic networks isometric with any real causal network. Thus, for example, the propositions \(p\) and \((p \& (q \lor \neg q))\) have identical semantic relations, and so any causal network isomorphic with a semantic network containing \(p\) will also be isometric with the semantic network obtained by systematically substituting \((p \& (q \lor \neg q))\) for \(p\). It follows that there will be indefinitely many propositional objects for any given propositional attitude. To put this point another way, functional role semantics cannot discriminate between the belief that \(p\) and the belief that \((p \& (q \lor \neg q))\), whilst folk psychology (external) insists on distinguishing between them.

It seems pretty clear that functional role semantics will not yield an adequate theory of content for (external) folk psychological beliefs and desires. Let us therefore abandon option (2) and turn to option (3). According to option (3), the fundamental unit of semantic evaluation is the subattitudinal state. Typically, such
units are called \textit{concepts}.\footnote{I rest very little weight on this terminology. The expression 'concept' has a variety of uses, many at odds with the use adopted here.} Propositional attitudes get their content in virtue of the contentfulness of their constituent concepts, and whole systems of intentional states get their content in virtue of the contentfulness of their constituent propositional attitudes. The so-called \textit{language of thought} hypothesis claims that the construction of propositional attitudes from concepts is directly analogous to the construction of sentences from words: concepts are combined into propositional attitudes in accordance with purely syntactical rules.\footnote{See Fodor 1975, 1987: 135-154. See also Field 1980; Harman 1973.} Note that the language of thought hypothesis is compatible with — although not entailed by — commonsense functionalism. Grafting the language of thought hypothesis onto commonsense functionalism yields the claim that "mentalese" sentences occupy the causal roles characteristic of (external) folk psychological beliefs and desires. I want to argue that the language of thought hypothesis allows us to account for certain properties of thought which cannot otherwise be readily accounted for. Since the language of thought hypothesis entails option (3), by defending the language of thought hypothesis I am defending the claim that the fundamental unit of semantic evaluation is the subattitudinal concept.

I propose to consider four arguments put forward by Jerry Fodor in favour of the language of thought hypothesis. In each case, a property of thought is located and it is then argued that the language of thought hypothesis smoothly accounts for that property. Of course, arguments of this kind are defeasible. Nevertheless, taken together, the four arguments add up to an impressive case for the language of thought hypothesis.

\textit{First Argument.} Fodor suggests that we adopt the following principle:

\begin{quote}
Suppose there is a kind of event $C_1$ of which the normal effect is a kind of event $E_1$; and a kind of event $C_2$ of which the normal effect is a kind of event $E_2$; and a kind of event $C_3$ of which the normal effect is a complex event $E_1 \& E_2$. Then, \textit{ceteris paribus}, it is reasonable to infer that $C_3$ is a complex event whose constituents include $C_1$ and $C_2$.\footnotemark[8]
\end{quote}

Applying this principle to cases of mental causation, let $C_1$, $C_2$, and $C_3$ be the inner causes of, respectively, the behaviours $E_1$, $E_2$, and $E_3$. It follows (\textit{ceteris paribus}), that the mental state $C_3$ is a complex event whose constituents include mental states $C_1$ and $C_2$. But in that case at least some mental states are built up combinatorially
from semantically-evaluable constituent parts. But to concede that much is to grant
the language of thought hypothesis. 9

Second argument. We need to account for not only the semantic properties
attributed by folk psychology (external) to beliefs and desires, but also for the causal
properties attributed by folk psychology (external) to beliefs and desires. That is, we
need not only a theory of mental content but also a theory of mental processes. A
powerful account of mental processes is the computer metaphor. In a computer,
representational states causally interact solely in virtue of their syntax. The machine
is, however, programmed in such a way that the causal interactions of the
representational states mirror their semantic relations. Such programming is only
possible if the symbols are drawn from a finite vocabulary, with more complex
expressions being built up combinatorially. Without a combinatorial syntax based on
a finite vocabulary, the machine would have to be able to recognize and respond to an
indefinite number of syntactic entities. By analogy, mental processes involve the
causal interactions of mental symbols. Such interactions occur solely in virtue of the
symbols’ syntax, but nevertheless preserve semantic relations. But again the story is
only plausible if the mental symbols are drawn from a finite lexicon, with more
complex expressions built up combinatorially. The computer metaphor thus commits
us to the language of thought hypothesis. 10

Third Argument. There are, in principle, indefinitely many English sentences,
and the explanation of this fact is quite straightforward: English sentences are
generated combinatorially from a finite lexicon. Similarly, it seems that we can think
indefinitely many thoughts. By analogy with the case of English, the obvious way to
explain this striking fact is by postulating that thoughts are generated combinatorially
from a finite lexicon of concepts. But to thus explain the productivity of thought just
is to put forward the language of thought hypothesis. 11

Fourth Argument. Mastering a language is not like memorizing a phrase
book. We do not learn languages sentence by sentence. Rather, when we master a
language we gain the ability to generate sentences that we have never heard before
and which may not have been uttered before. There is an obvious explanation of this
phenomenon: when we learn a language we learn not a list of sentences but a lexicon
and the rules for combining lexical items into sentences. We also have the capacity
to generate new thoughts we have not had before and which perhaps no one has had
before. By analogy with language, we do not think by deploying a pre-established
mental phrase book; rather, thinking involves a lexicon of concepts and a set of rules

10 See Fodor 1987: 143-147.
for combining concepts into thoughts. But to thus explain the systematicity of thought is to accept the language of thought hypothesis.\textsuperscript{12}

I want to endorse the language of thought hypothesis: the occupants of the causal roles implicitly defined by our commonsense talk about beliefs and desires are most plausibly regarded as sentences in the language of thought. Only by taking the occupants of the causal roles to be sentences in the language of thought can we account for the existence of mental processes and the productivity and systematicity of thought. But taking the language of thought hypothesis seriously requires us to account for the content of subattitudinal concepts. The rest of this Chapter is devoted to that task.

4.2 The Resemblance Theory. In a nutshell, the problem of content is this. What makes my belief that Saba is a tiger about Saba? One suggestion is that my thought about Saba is about Saba because my belief resembles Saba.\textsuperscript{13}

Difficulties for the resemblance approach are legion. Two standard objections suggest that resemblance relations are too scarce to do the job demanded of them. In what way does my thought of Saba resembles Saba? My thought is, presumably, located somewhere in my head; Saba is prowling around outside. My thought came into existence sometime last week; Saba came into existence about 4 years ago. There does not seem to be a suitable resemblance relation between my thought of Saba and Saba herself. Resemblance relations are too scarce. A second standard objection emerges when we consider that we can have a variety of thoughts about a particular under various descriptions. I can think about Saba as a tiger, as striped, as a mammal, as four-legged. But how on the resemblance approach can my thought of Saba as a tiger not also be a thought of Saba as a mammal? Any thought I have of Saba would seem to resemble a mammal just as much as it resembles a tiger.\textsuperscript{14} Again a scarcity of suitable resemblance relations seems to lie behind the problem. The problem would be avoided if we could locate a resemblance relation between my thought of Saba as a tiger and her being a tiger, and a distinct resemblance relation between my thought of Saba as a mammal and her being a mammal.

It might be replied that both of these problems - and indeed the general difficulty of a lack of suitable resemblance relations - arise because we are too restrictive in the way we think about resemblance. Traditionally, the resemblance relations which purportedly underpin content have been thought of as pictorial resemblance relations. But pictorial resemblance is only one sort of resemblance

\textsuperscript{12} See Fodor 1987: 148-153.
\textsuperscript{13} Hume claimed that conceiving, knowing, believing, and so forth consisted in the having of ideas. Ideas were, in turn, taken to be copies of sensory impressions. See Hume 1739, Book I, Part IV.
\textsuperscript{14} This problem was first raised by Berkeley. See Berkeley 1710.
relation; there exist a multitude of others. Thus, even though my thought of Saba might bear no pictorial relation to her, some resemblance relation exists between my thought of Saba and Saba. Similarly, the reply continues, some resemblance relation exists between my thought of Saba as a tiger and her being a tiger, which does not exist between my thought of Saba as a tiger and her being a mammal.

I think it should be granted that such resemblance relations exist and that therefore the resemblance theory of content is not guilty as charged. The problem with resemblance theories is not that there are too few resemblance relations but that there are too many. Resemblance relations are too cheap: there are resemblance relations between everything and everything else. We need some principled way of distinguishing the resemblance relations that determine content from those that do not, and none is forthcoming.

It might be suggested that resemblance relations cannot be as cheap as I make out because we find them so useful. I have just bought a new bicycle and I want to tell you what it is like. I say, quite unproblematically, that it is more like John's bike than Betty's. If everything really is like everything else, how could my remark possibly be informative? An obvious suggestion is that context determines which resemblance relations are relevant and which are not. Similarly, the resemblance theorist might hope to locate the content-determining resemblance relations by appealing to context. But now circularity threatens. For appeals to context will almost invariably involve appeals to the propositional attitudes of the persons involved, but the content of propositional attitudes is the very thing we set out to explain.

4.3 The Indicator Theory. According to the indicator theory of content, content is determined by counterfactual causal relations between a thought and its object. That is, my thought is a thought about p because, in relevant worlds, p, and only p, causes p-thoughts. 'Relevant worlds' is then spelt out in terms of nomic possibility: a world is relevant if and only if it shares our laws of nature. Putting this still other ways, my thought that p is about p because my thought indicates, or covaries with, or carries information about, p.\textsuperscript{15}

The fundamental problem faced by the indicator theory is the problem of misrepresentation.\textsuperscript{16} According to the indicator theory, my thought about Saba is about Saba because in all relevant worlds Saba, and only Saba, causes me to think of Saba. Say that it is true that my tokens of 'Saba' are caused by, and only by, Saba. Then misrepresentation is impossible. If I have a Saba thought it must be true: Saba

\textsuperscript{15} See Dretske 1981, 1988; Fodor 1990b; Stampe 1977.

\textsuperscript{16} Also called the disjunction problem or the error problem.
must indeed have caused it. On the other hand, say that I sometimes think of Saba when I have seen not Saba but Tiga. That is, on dark nights I sometimes token 'Saba' when I have in fact been confronted by Tiga. This looks like a case of misrepresentation: I tokened 'Saba' falsely. But note that if I token 'Saba' when presented with Tiga-on-a-dark-night, my tokenings of 'Saba' counterfactually covary not with Saba, but with Saba-or-Tiga-on-a-dark-night. So once again misrepresentation is impossible. I have not misrepresented Saba if I token 'Saba' on a dark night when confronted by Tiga. Rather, I have (truthfully) represented Saba-or-Tiga-on-a-dark-night.17

One way of dealing with this problem is to distinguish a narrowly circumscribed set of situations under which meaning is determined from a broader set of situations under which misrepresentation is possible. Following Fodor, let us call the former situations 'type 1' and the latter 'type 2'.18 In type 1 situations, whatever would cause a symbol to be tokened lies within its extension; in type 2 situations a symbol can be caused by things which lie outside its extension. Thus, symbols tokened in type 1 situations must be true, but symbols tokened in type 2 situations may be false.

So far we do not so much have a solution to the misrepresentation problem as a suggestion about how such a solution might be located. What is required is an account of type 1 and type 2 situations. Dretske has suggested identifying type 1 situations with the circumstances under which we learn the meaning of a symbol, and identifying type 2 situations with the circumstances under which we use it.19 During the learning period my teachers ensure that all my tokenings of 'Saba' are caused only by Saba herself. Tokens of 'Saba' thus represent Saba rather than, say, Saba-or-Tiga-on-a-dark-night. Later on – outside the learning period – I might token 'Saba' in the presence of Tiga-on-a-dark-night, but such tokens are misrepresentations: I have misrepresented Tiga as Saba.

An obvious problem with Dretske's proposal is that it demands a strict learning period/using period distinction. But of course no such distinction exists. We muddle along, learning and using as we go. A somewhat more subtle objection is as follows.20 Dretske's idea is that during the training period all my tokenings of 'Saba' are caused by Saba: the teacher sees to it that I am not exposed to Tiga-on-a-dark-night and thus do not get the chance to token 'Saba' inappropriately. However, according to the indicator theory, content is determined not by the actual causal relations between thoughts and their objects, but by the counterfactual causal

---

17 See Fodor 1990b.
18 See Fodor 1990b: 64.
20 See Fodor 1990b.
relations between thoughts and their objects. And it remains true that, had I been exposed to Tiga-on-a-dark-night during the training period, I would have tokened 'Saba'. The training period/use period distinction is thus of no help: if it is true that I would have tokened 'Saba' when confronted with Tiga-on-a-dark-night during the training period, then 'Saba' represents Saba-or-Tiga-on-a-dark-night. The problem of misrepresentation remains.

In the next two Sections of this Chapter I will examine in depth the two most prominent solutions to the problem of misrepresentation. Before moving on, though, I would like to briefly sketch two further difficulties for the indicator theory: what I call the inverse misrepresentation problem and the Swampman objection.

The misrepresentation problem arises because a token of a subattitudinal concept can be caused by things which lie outside its extension: 'Saba' tokens are sometimes caused by Tiga-on-a-dark-night. Figure 4.1, below, represents the misrepresentation problem diagrammatically (the arrows stand for causal relations).

![Figure 4.1](image)

In the standard misrepresentation problem, mental tokens of a single type are caused by events of two distinct types. In the inverse misrepresentation problem, mental tokens of two distinct types are caused by events of a single type. Thus, confronted with the planet Venus I might token 'Morning Star' or I might token 'Evening Star'. Again, confronted with a sample of H2O, I might token 'H2O' or I might token 'water'. The inverse misrepresentation problem is represented diagrammatically in figure 4.2, below.

Since both my 'Morning Star' tokens and my 'Evening Star' tokens counterfactually covary with Venus, the indicator theory of content is committed to
the claim that my 'Morning Star' tokens have the same content as my 'Evening Star' tokens. However, commonsense functionalism takes my belief that the Morning Star has just risen to be distinct from my belief that the Evening Star has just risen. After all, the former causes me, ceteris paribus, to say that the Morning Star has just risen; the latter causes me, again ceteris paribus, to say that the Evening Star has just risen. Since the two beliefs have different effects, they are, according to commonsense functionalism, different beliefs. But the two beliefs vary only in that one contains a 'Morning Star' token where the other contains an 'Evening Star' token. It follows that 'Morning Star' and 'Evening Star' are different concepts. Yet, by the indicator theory of content, they have the same content. If we take it that concepts are individuated solely by content, the indicator theory is in trouble: it identifies only one belief where folk psychology (external) recognizes two. We will have cause to revisit this problem when we discuss Fodor's theory of content in Section 4.5.

The final difficulty I would like to discuss is the well-known Swampman case. The Swampman example is widely taken as an objection to teleological theories of content, and I will discuss Swampman again in that context (Section 4.4). However, Swampman also raises problems for the indicator theory of content. Swampman is an atom-for-atom replica of myself who emerged from a chance reaction of organic molecules in the local swamp.21 There is a strong (external) folk psychological intuition that Swampman has contentful states. Indeed, there is a strong (external) folk psychological intuition that Swampman has the same contentful states as I do. However, whilst it is at least plausible that my token of 'Saba' covaries with, and only with Saba, it is not at all plausible that Swampman's token of 'Saba' covaries with, and only with, Saba. After all, he has never had causal contact with Saba. It follows

21 See Davidson 1987.
that the indicator theory of content is at odds with the (external) folk psychological attribution of belief.

4.4 The Teleological Theory. As we have seen, the indicator theory of content has difficulty coping with misrepresentation. We have also seen that one way to approach the problem of misrepresentation is by distinguishing between type 1 situations in which meaning is determined and type 2 situations in which misrepresentation is possible. Dretske appealed to just this strategy. Teleological theories can also be understood in terms of the type 1/type 2 distinction. In teleological theories, type 1 situations are identified with Normal situations. If the situation is Normal then the content of my thought is determined by whatever counterfactually covaries with my thought. That is, when the situation is Normal, misrepresentation is impossible. When the situation is abNormal, however, misrepresentation can occur.

What is it for a situation to be Normal? Normalcy is spelt out in terms of natural selection. A situation is Normal for a mechanism when it is such that the mechanism can function as it was selected to function. The beating of my heart has a variety of effects. For example, it both pumps blood and makes a small noise. According to standard accounts of biological function, the function of my heart is to pump blood (rather than make a small noise), because my ancestors survived at least in part because their hearts pumped blood (rather than made a small noise). Circumstances are Normal when my heart can fulfil its function; that is, when my heart can do what it was selected to do, namely, pump blood.

There are two ways in which the idea of a biological function can be applied to the problem of misrepresentation. On the first way, functions are attributed to thoughts themselves. A situation is type 1 for my thought that p if it is such that my thought can fulfil its function in that situation; that is, if the situation is Normal for my thought that p. Type 2 situations, on the other hand, are identified with situations in which my thought that p cannot fulfil its function; that is, with situations which are abNormal for my thought that p.

The problem with this way of applying the idea of biological function to the problem of misrepresentation is that it is difficult to see in what way my thought that p has a biological function in the required sense. We said that the function of my heart was to pump blood because my ancestors survived at least in part because their hearts pumped blood. It follows that for some characteristic C of organism O to have a function, C must have been present in O's ancestors. After all, if C was not present

22 See Fodor 1990b.
it could not have been selected for. But I have thoughts that my ancestors most
certainly did not have; that is, I have thoughts which could not possibly have a
selective history because they did not exist to be selected. Thus, my thoughts about
Macintosh computers are, from the perspective of evolutionary history, novel. There
is no way that my evolutionary ancestors entertained those ideas.

The suggestion that thoughts have biological functions is tempting because it
seems reasonable – at least prima facie – to attribute biological functions to cognitive
mechanisms. It seems plausible that the mechanisms that underpin my ability to
produce and manipulate thoughts were selected for: presumably my ancestors
survived at least in part because they could produce and manipulate thoughts. But the
claim that cognitive mechanisms have functions does not entail that the thoughts
those mechanisms produce and manipulate have functions. As Fodor remarks, 'You
could perfectly well have a machine whose function is to produce things that are
themselves functionless'. This brings us to the second way in which we might try to
apply the idea of biological function to the problem of misrepresentation: attribute
functions not to thoughts themselves but to the cognitive mechanisms that produce
and manipulate thoughts.

Intuitively, misrepresentation occurs when something goes wrong: when a
cognitive mechanism is working as it should, misrepresentation will not occur. So
perhaps we can solve the misrepresentation problem by distinguishing between the
situations in which everything is going well and those in which it is not. In other
words, we can identify type 1 situations with situations in which the cognitive
mechanism is working as it should, and type 2 situations with all other occasions.
What is it for a cognitive mechanism to be working as it should? At this point appeal
is made to natural selection. A cognitive mechanism is working as it should when it
is working Normally; that is, when it is doing what it was selected to do. In sum,
when my cognitive mechanisms are doing what they were selected to do, the
extension of each of my mental token is whatever it covaries with. Sometimes,
though, my cognitive mechanisms work abnormally – they do not do what they were
selected to do. In those cases tokens may be caused by things not in their extension.
In other words, misrepresentation can occur.

The standard example is a frog snapping at a fly. Apparently frogs are fairly
bad at discriminating flies from other small black objects moving quickly across their
visual fields. Thus, they snap at beebees thrown in their direction. It is natural to say
that when the frog snaps at a beebee it has misrepresented the beebee as a fly. Note,
though, that since the state within the frog which causes snapping behaviour covaries
with both flies and beebees, the indicator theory is committed to saying that the state

within the frog represents flies-or-beebees. In other words, the frog does not misrepresent beebees as flies; it (truthfully) represents beebees as flies-or-beebees. Misrepresentation has become impossible. This is, of course, just another example of the problem of misrepresentation. The teleological theorist argues, however, that the frog represents flies rather than flies-or-beebees because the cognitive mechanism which is triggered by flies or beebees and which directs snapping behaviour, has the function of catching flies. There were no beebees in the environments of the modern frog's ancestors, and anyway catching beebees would not have raised their reproductive success. Thus the mechanism is a fly catcher, not a fly-or-beebee catcher, and so the representation which causes snapping behaviour is a representation of a fly. Therefore, when the frog snaps at a beebee, it misrepresents the beebee as a fly. There are at least three objections which have been offered against this story.

First Objection. Fodor has pointed out that there is in fact no good reason to describe the frog's mechanism as a fly catcher rather than as a fly-or-beebee catcher. Since in the frog's ancestors' environments flies and flies-or-beebees were coextensive, natural selection offers no way of distinguishing between the hypothesis that the frog's mechanism is a fly catcher and the hypothesis that it is a fly-or-beebee catcher. The mechanism would have, ceteris paribus, conferred exactly the same survival advantage on the frog whether it was a fly or a fly-or-beebee catcher. To put the point another way, in the environment we are concerned with the set of flies is coextensive with the set of flies-or-beebees. Natural selection is indifferent between coextensive properties, but intentionality is not: the belief that X is a fly is distinct from the belief that X is a fly-or-beebee. I propose to consider two responses to Fodor's objection.

Kim Sterelny has suggested that what Fodor is objecting to is in fact a specific case of a general problem with gerrymandered or gruesome predicates. The predicates 'is a fly' and 'is a fly-or-beebee' are assumed to be locally coextensive. Similarly, the predicates 'is green' and 'is grue' are locally coextensive. (An object is grue if and only if it is green and first observed before midnight January 1, 2000, or blue and first observed after midnight January 1, 2000.) Notoriously, gruesome predicates create problems in the philosophy of science. In particular, they are not projectible. Thus, whilst it is true that 'Every emerald I have seen so far has been green' provides evidence for 'Every emerald is green', it is not the case that 'Every emerald I have seen so far has been grue' provides evidence for 'Every emerald is grue'. What is required to avoid this problem is an account of gruesomeness – an

---

26 See Fodor 1990b: 71-73.
28 See Goodman 1965, especially Ch. 3.
account of which properties are projectible and which are not. Notoriously, such an account is hard to come by. But if we did have an account of gruesomeness, reasons Sterelny, we could solve the teleologist's problem by ruling out predicates such as 'is a fly-or-beebee' as unacceptably gruesome. We can, though, hardly blame the teleologist for not having a solution to the grue problem. No one has a solution to the grue problem.

Sterelny's objection turns on the gerrymanderedness of the predicate 'fly-or-beebee'. He argues that everyone has a problem with gerrymandered predicates, so it is not surprising to see that problem pop up in this context. But it is not clear that the predicates which create problems in the philosophy of science are gerrymandered in the same sense as the predicates which create problems for the teleological theory of content. That is, it is not clear that 'is a fly-or-beebee' is gerrymandered in the same sense as 'is grue'. Note that whilst 'is grue' is not projectible, 'is a fly-or-beebee' is projectible. 'Every small black thing I've seen so far has been a fly-or-beebee' does provide evidence for 'Every small black thing is a fly-or-beebee'. So it cannot be the case that the problem confronting the teleological theory of content is just the old problem about gruesome predicates: 'is a fly-or-beebee' is not gruesome.

Perhaps 'is a fly-or-beebee' is gerrymandered in virtue of being disjunctive. After all, 'is grue' is disjunctive. But disjunctiveness simpliciter cannot be the problem with 'is a fly-or-beebee' (or with 'is grue' for that matter), because we can have perfectly good, useful, and projectible disjunctive predicates. Thus I can coin the predicate 'is rellow'. Something is rellow if and only if it is red or yellow. 'Is rellow' is a perfectly well-behaved predicate. It is projectible and, moreover, might even come in handy. ('Stop at rellow traffic lights' is quite a good piece of advice.) Similarly, 'is strongly acid or strongly alkaline' is a disjunctive predicate. Nevertheless, there is nothing wrong with it. A gardening manual might instruct us not to plant a certain species of tree in soil which is strongly acid or strongly alkaline.

To sustain his attack on Fodor, Sterelny needs an account of gerrymanderedness such that both 'is grue' and 'is a fly-or-beebee' are gerrymandered. We have examined two analyses of the property of being gerrymandered: a predicate is gerrymandered if and only if it is gruesome (ie nonprojectible) and a predicate is gerrymandered if and only if it is disjunctive. Neither analysis supports Sterelny's conclusion. I think, though, that there is a more general argument to be had against Sterelny's suggestion. Note that I can have thoughts about gerrymandered properties. I can have thoughts about flies-or-beebees or about grueseness. This suggests that, if the teleologist wants a general account of content, she had better not appeal to gerrymandered predicates to solve the problem of misrepresentation. We can develop this intuition by seeing what happens when we try to applying the teleological theory to thoughts about gerrymandered properties.
Change the frog example so that in its natural environment the frog snaps at both flies and slugs. Moreover, assume that the frog’s ancestors flourished because they ate flies and slugs. Intuitively, when the frog snaps at a beebee it misrepresents the beebee as a fly-or-slug. The indicator theorist, though, is committed to the claim that the frog (truthfully) represents the beebee as a (fly-or-slug)-or-beebee. Once again it is open to the teleologist to respond that since there were no beebees in the frog’s ancestor’s environments, the frog’s snap mechanism is a fly-or-slug catcher not a (fly-or-slug)-or-beebee catcher. And once again it is open to Fodor to observe that flies-or-slugs and (flies-or-slugs)-or-beebees were locally coextensive in the frog’s ancestors’ environments, and that natural selection cannot distinguish between locally coextensive properties. Now if Sterelny’s appeal to gerrymandered properties is going to provide a general solution to the teleologist’s difficulties it should settle whether the frog’s ‘fly-or-slug’ token represents flies-or-slugs or (flies-or-slugs)-or-beebees. But it does not. The predicates ‘is a fly-or-slug’ and ‘is a (fly-or-slug)-or-beebee’ are both gerrymandered, and so we cannot settle the referent of the frog’s token on the grounds that one is gerrymandered but the other not. 29

The second response to Fodor’s objection to the teleological theory of content points out that, whilst flies and flies-or-beebees are coextensive at this world, they are not coextensive at other worlds. Thus, whilst it is true that a fly-or-beebee catching mechanism would have conferred exactly the same reproductive success at this world as a fly catching mechanism, in worlds in which flies and flies-or-beebees are not coextensive the fly catching mechanism would have conferred greater reproductive success. In other words, if the frog’s ancestors had evolved in environments in which there were beebees, natural selection would have favoured those frogs which could distinguish flies from flies-or-beebees, so the actual frog’s mechanism is a fly catcher not a fly-or-beebee catcher. 30

This proposal in effect modifies the account of biological function appealed to by the teleosemanticist. Following Millikan and Neander, we have thus far assumed that characteristic C of organism O has function F if and only if O’s ancestors successfully reproduced in virtue of possessing C which did F. Thus, as Fodor points out, the frog’s snap mechanism can truthfully be described as having the function of catching flies-or-beebees because the frog’s ancestors successfully reproduced in virtue of their possessing a snap mechanism which caught flies-or-beebees. Now it is being suggested that characteristic C of organism O has function F if and only if O’s ancestors, and their counterparts in other worlds, successfully reproduced because

29 Sterelny might try to argue that ‘is a (fly-or-slug)-or-beebee’ is more gerrymandered than ‘is a fly-or-slug’, and thus that the frog’s token represents the latter rather than the former. Without a detailed account of the notion of degrees of gerrymanderedness it is hard to evaluate such a proposal.
30 This objection was first brought to my attention by Greg Currie and Greg O’Hear.
they possessed C which did F. In other words, the notion of biological function is being intensionalized. On this revised account of biological function, the frog's snap mechanism can no longer be truthfully described as having the function of catching flies-or-beebees. Had some of the frog's ancestors' counterparts — those that lived in worlds rich in beebees — possessed a mechanism which could not distinguish between flies and beebees, they would have failed to successfully reproduce.

Fodor's own response to this move is to point out that in some worlds in which flies and flies-or-beebees are not coextensive, natural selection will favour a mechanism which fails to distinguish between flies and flies-or-beebees. For example, in worlds in which beebees are nutritious, frogs that eat both flies and beebees (and do not waste resources on mechanisms capable of distinguishing them) will, ceteris paribus, be better off. Thus, once again, appeals to natural selection have failed to resolve the disjunction problem. The teleosemanticist could respond by pointing out that worlds in which there were edible beebees in the frog's ancestors' environments are more distant from the actual world than worlds in which there were inedible beebees in the frog's ancestors' environments, and then claim that only the nearest worlds in which flies and flies-or-beebees are not coextensive need be considered when assigning content. In other words, the teleosemanticist can once again modify the account of biological function appealed to by her theory of content. The new proposal is that characteristic C of organism O has function F if and only if O's ancestors, and their counterparts in nearby worlds, successfully reproduced because they possessed C which did F.

However, examples can readily be generated which overturn this new proposal. Let me tell a little tale. In the local zoo there is a large collection of frogs. Unfortunately, the frog keeper is having trouble securing a supply of flies for her charges to eat. Ingeniously, she develops small pellets of food which, when thrown near a frog, causes the frog to snap. The pellets of food are, in fact, more nutritious than flies. The frog keeper has observed that frogs fed either exclusively on the artificial food pellets or on a fly/artificial food pellet mix do better — have more offspring — than frogs fed exclusively on flies.

I think that it is reasonable to say that the frogs in the zoo misrepresent the food pellets as flies, just as it was reasonable to say of the frog in the original example that it misrepresented beebees as flies. On the indicator theory of content, the frogs in the zoo do not misrepresent the food pellets as flies; rather, they (truthfully) represent the food pellets as flies-or-food-pellets. This is, of course, just another example of the problem of misrepresentation. The teleosemanticist proposes to solve the misrepresentation problem by appealing to natural selection. The frogs'
mechanism was selected as a fly catcher rather than a fly-or-food-pellet catcher — there were, after all, no artificial food pellets in the frogs' ancestors' environments — and so the frogs misrepresent the food pellets as flies. It is now open to Fodor to repeat the move against the teleosemanticist that he made earlier. Flies and flies-or-food-pellets were coextensive in the frogs' ancestors' environments so natural selection cannot distinguish, as the teleosemanticist requires, between a fly catching mechanism and a fly-or-food-pellet catching mechanism. Finally, the teleosemanticist invites us to consider the nearest worlds in which flies and flies-or-food-pellets were not coextensive in the frog's ancestors' environments.

What happens in those worlds? The nearest worlds in which flies and flies-or-food-pellets were not coextensive in the frog's ancestors' environments are worlds in which there were food pellets in the frog's ancestors' environments. But, as we have seen, eating food pellets confers a reproductive advantage on frogs. Thus, in the nearest worlds in which flies and flies-or-food-pellets were not coextensive in the frog's ancestors' environments, natural selection will favour those frogs which do not bother to discriminate between flies and food pellets. Thus, on the (revised) teleological theory of content, the frogs in the zoo do not misrepresent food pellets as flies; rather, they (truthfully) represent food pellets as flies-or-food-pellets. Once again appealing to natural selection has failed to resolve the disjunction problem.

Second objection. As we have seen, teleologists appeal to natural selection to solve the problem of misrepresentation. More specifically, they appeal to what a cognitive mechanism was selected to do. This assumes, of course, that the cognitive mechanisms involved in the fixation of beliefs and desires were in fact selected. But we can question that assumption. The point is not that the theory of natural selection might be false. That is a singularly unlikely hypothesis. The point is, rather, that interesting and important properties of organisms may not be the products of natural selection. They may be, as it were, side-effects of natural selection.32

The idea that all the interesting and important characteristics of an organism were directly selected is called adaptationism. Perhaps the most famous attack on adaptationism is Gould and Lewotin's 'The Spandrels of San Marco and the Panglossian Paradigm'.33 They point out that one of the most striking architectural features of the San Marco cathedral, Venice, results not from the deliberate exercise of architectural genius but is a mere by-product of the arrangement of other architectural features. A more biological and oft-cited example is the shape of the human chin. Our chins are the shape they are because of selection pressure on,

amongst other things, the human jaw. Chin shape is a by-product of natural selection, not a direct product of it.

It is possible, then, that our cognitive mechanisms were not directly selected. They may be a by-product of natural selection rather than a direct result of natural selection. If this is the case then the teleological story about content collapses, for the teleological story about content takes as premiss the claim that our cognitive mechanisms are the product of natural selection. The teleological theory thus rests on a major empirical premiss, a premiss which, moreover, is almost never scrutinized by the teleological camp.

Fodor and Sterelny are correct in pointing out that teleological theories of content rest on the (often unacknowledged) assumption that our cognitive mechanisms are in fact the products of natural selection. But this observation only constitutes a genuine objection to teleological theories if adaptationism about our cognitive mechanisms is false. There exists, though, a substantial body of evidence in favour of the claim that our cognitive mechanisms are the products of natural selection. So called evolutionary psychologists take as their starting point paleoanthropological evidence about the selection pressures to which our Pleistocene hunter-gatherer ancestors were subject, and then argue that various cognitive mechanisms existing in modern humans can best be explained as adaptations to those pressures. Mechanisms thus (putatively) explained include the mechanisms underpinning cooperative behaviour, mate selection, sexual proprietariness, food aversion during pregnancy, maternal vocalization, natural language acquisition and processing, spatial abilities, and responses to landscapes.34 So it seems that many, if not all, significant cognitive processes have selective histories and therefore Normal functions. The teleosemanticist does not have to be worried on that score.

Third objection. Recall that we are seeking an account of the content of (external) folk psychological propositional attitudes. As (external) folk psychologists, though, we need neither know nor care about evolutionary history. This suggests that (external) folk psychological judgements about the contentfulness of mental states might diverge from judgements derived from teleological theories of content. Thus consider Swampman, introduced in the previous Section. Swampman is an atom-for-atom replica of myself. However, whilst I am the result of a long process of natural selection, Swampman arose from a chance reaction of organic molecules in the local swamp.35 There is a strong (external) folk psychological intuition that Swampman has contentful states. Not so, says the teleologist. For a mental token has the content 'p' in virtue of its being caused by, and only by, p under

34 See the papers collected in Barkow et al 1992. But see also Sterelny forthcoming for a critique of Barkow et al and the evolutionary psychology program generally.
35 See Davidson 1987.
Normal conditions. But 'Normal' is, as we have seen, spelt out in terms of natural selection. Since Swampman does not have a selective history, he does not have any contentful states. The teleological theory thus fails to capture the notion of content found in (external) folk psychology.

4.5 Fodor's Theory. In 'A Theory of Content II', Fodor develops a theory of content first put forward in his *Psychosemantics*. Fodor's theory is, very roughly, a type of indicator theory. However, Fodor does not attempt to solve the problem of misrepresentation by distinguishing between type 1 and type 2 situations. Rather, Fodor appeals to asymmetric nomic dependencies between the property of being p and the property of being a cause of 'p' tokens. I will give two glosses of this idea, relying each time on Fodor's own words. First gloss:

Cows cause "cow" tokens, and (let's suppose) cats cause "cow" tokens. But "cow" means cow and not cat or cow or cat because there being cat-caused "cow" tokens depends on there being cow-caused "cow" tokens, but not the other way around. "Cow" means cow because, as I shall henceforth put it, noncow-caused "cow" tokens are asymmetrically dependent upon cow-caused "cow" tokens. "Cow" means cow because but that "cow" tokens carry information about cows, they wouldn't carry information about anything. 36

If "cow" tokens are to carry information about cows, the generalization cows cause 'cows' must be counterfactual supporting. Fodor argues that a generalization of the form X's cause Y's is counterfactual supporting if and only if there is a covering law relating the property of being X to the property of being a cause of Y's; that is, if and only if the property of being X and the property of being a cause of Y's are nomically related. 37 This leads us to the second gloss:

"cow" means cow if (i) there is a nomic relation between the property of being a cow and the property of being a cause of "cow" tokens; and (ii) if there are nomic relations between other properties and the property of being a cause of "cow" tokens, then the latter nomic relations depend asymmetrically upon the former. 38

---

36 Fodor 1990c: 91. His emphasis.
37 Fodor 1990c: 93.
38 Fodor 1990c: 93. His emphasis.
I propose to demonstrate how Fodor's theory can deal with the difficulties which thwarted the indicator and teleological approaches to content (Sections 4.3 and 4.4): Swampman; the problem of misrepresentation; and what I have called 'the inverse misrepresentation problem'. I will also discuss three objections which arise more specifically in the context of Fodor's theory. Surprisingly, it turns out that Fodor's theory is committed to a degree of content holism. I say 'surprisingly' because Fodor is an avowed opponent of content holism. I close this Section by arguing that the holism inherent in Fodor's account of content is benign.

First objection. How does Fodor handle the problem of misrepresentation? Given that my 'Saba' tokens are caused not only by Saba but also by Tiga-on-a-dark-night, why does 'Saba' represent only Saba rather than Saba-or-Tiga-on-a-dark-night? By the first gloss (above), my 'Saba' tokens represent Saba rather than Saba-or-Tiga-on-a-dark-night if, if Saba did not cause 'Saba' tokens then neither would Tiga-on-a-dark-night. Think of the nearest world in which Saba does not cause me to token 'Saba' - a world in which, say, I have never seen Saba nor had a thought about her. In such a world I would not mistakenly token 'Saba' in the presence of Tiga-on-a-dark-night; that is, Tiga-on-a-dark-night would not cause me to token 'Saba'. Thus, Fodor's asymmetric dependency approach to content avoids the problem of misrepresentation which dogged the indicator and teleological theories.

Second objection. Can Fodor's theory handle the inverse misrepresentation problem? Presumably my 'Morning Star' tokens are nomically dependent on Venus. Moreover, if my 'Morning Star' tokens were also nomically dependent on, say, Alpha Centuri, then the nomic dependence of my 'Morning Star' tokens on Alpha Centuri would itself be asymmetrically dependent on the nomic dependence of my 'Morning Star' tokens on Venus. Similar remarks apply to my 'Evening Star' tokens. Presumably my 'Evening Star' tokens are nomically dependent on Venus. Moreover, if my 'Evening Star' tokens were also nomically dependent on, say, Halley's Comet, then the nomic dependence of my 'Evening Star' tokens on Halley's Comet would itself be asymmetrically dependent on the nomic dependence of my 'Evening Star' tokens on Venus. In other words, my 'Morning Star' tokens and my 'Evening Star' tokens have the same content: they both represent the planet Venus.

However, my belief that the Morning Star just rose is distinct from my belief that the Evening Star just rose. After all, they have different effects. For example, the former belief causes me to say, ceteris paribus, that the Morning Star just rose whilst the latter belief causes me to say, again ceteris paribus, that the Evening Star just rose. Since these beliefs differ only in that the former contains a 'Morning Star' token where the latter contains an 'Evening Star' token, 'Morning Star' tokens cannot be identical to 'Evening Star' tokens. It follows that, by Fodor's theory of content,
two concept tokens can have identical contents and yet be of different types. That is, concepts are not individuated solely by content.

This raises a serious issue. What else are concepts individuated by besides content? The obvious way to solve this conundrum is by introducing a sense/reference distinction: my 'Morning Star' tokens have the same reference as my 'Evening Star' tokens, but differ in sense. However, Fodor's theory of content offers only an account of reference, not of sense, and thus seems to be fundamentally inadequate.

Fodor has an ingenious response to this difficulty—a response which is, in my view, exactly right. Recall that concepts have syntactic properties. It follows that two concept tokens can have identical content but differ in type because they differ in syntax. Thus, whilst my 'Morning Star' tokens and my 'Evening Star' tokens have identical content because they are both suitably nomically dependent on Venus, they differ in syntax and are thus of distinct type. Moreover, a concept's syntax partly determines the causal properties of the propositional attitudes of which it is a constituent part. Thus, the proposed difference in syntax between my 'Morning Star' tokens and my 'Evening Star' tokens accounts for the divergent causal relations of my belief that the Morning Star just rose and my belief that the Evening Star just rose.

Third Objection. Since there are no unicorns, there can be no causal relations between unicorns and tokens of 'unicorn'. How, then, can 'unicorn' tokens represent unicorns?42 This objection fails to notice that Fodor's theory of content is given in terms of nomic relations between properties, not causal relations between individuals. That is, this objection fails to notice the subtle shift that has occurred between the first and second glosses given above. The first gloss is given in terms of causal relations between individuals, and the objection under consideration does indeed hold against it. However, the second gloss is given not in terms of causal relations between individuals, but in terms of nomic relations between properties. Whilst it is true that there can be no causal relations between nonexistent individuals, there can be nomic relations between uninstantiated properties; in particular, there can be nomic relations between the property of being a unicorn and the property of being a cause of 'unicorn' tokens. Thus, as long as we accept the second gloss (rather than the first) as

39 This is, of course, directly analogous to the move made by Frege for public languages. See Frege 1952.
40 See Fodor 1990d.
41 Note that I am concerned here only with the semantics of mental concepts. I am not endorsing Fodor's solution as an alternative to the sense/reference distinction for public languages.
42 See Fodor 1990c: 100.
representing Fodor's proposal, Fodor can account for thoughts about nonexistent unicorns.43

What about cases involving objects which do not merely not exist, but which necessarily do not exist? I can have a thought about a square circle, but there can be no nomic dependencies between the property of being a square circle and the property of being a cause of 'square circle' tokens. Fodor takes the moral of this example to be that mental tokens which represent necessarily nonexistent objects cannot be primitive but must be built up by a process akin to definition.44

Before moving onto the next objection, note that Fodor's theory can handle Davidson's Swampman example (Sections 4.3 and 4.4). Swampman, recall, is an atom-for-atom replica of myself which arose from a chance reaction of organic molecules at the local swamp. In a way, every predicate is for Swampman just like 'is a unicorn' is for us. None of us has had causal contact with a unicorn; similarly, Swampman enjoys no semantically relevant causal relations with trees, fish, rocks, the Sydney Opera House . . . the list is a long one. Nevertheless, there exists a nomic relationship between the property of being the Sydney Opera House and the property of being a cause of 'Sydney Opera House' tokens. After all, such a nomic dependence exists in my case, and Swampman is an atom-for-atom replica of me. In other words, were Swampman confronted with the Sydney Opera House 'he would token 'Sydney Opera House'. Fodor's theory thus succeeds in attributing content to Swampman.

Fourth objection. Can Fodor account for the meanings of logical connectives? After all, it is highly implausible that tokens of, say, 'or' mean or because of nomic relations between the property of being or and the property of being a cause of 'or' tokens.

Fodor replies by endorsing, in effect, a functional role semantics for the logical connectives: 'I'm inclined to think that there is no objection to the idea that "+", "and", "all" and the like have the meanings they do because they play a certain causal role in the mental lives of their users'.45 Fodor insists, though, that he only wants to endorse functional role semantics for the logical connectives; indeed, he mentions approvingly Gilbert Harman's suggestion that the logical vocabulary can be distinguished from the non-logical vocabulary by the applicability of functional role semantics to the former but not the latter.

Fifth objection. Recall the Twin Earth example introduced in Section 3.4. Twin Earth is a molecule-for-molecule replica of Earth with one exception: the stuff that falls from the sky, flows in the rivers, fills the lakes, and pours from the tap on

---

45 Fodor 1990c: 110.
Twin Earth is XYZ not H2O. Here is the concern. If Oscar, an inhabitant of Earth, were exposed to XYZ he would token 'water'. There is therefore a nomic dependence of 'water' tokens not simply on H2O but on H2O-or-XYZ. So why do not Oscar's 'water' tokens represent H2O-or-XYZ? 46

This is, of course, just another version of the problem of misrepresentation: 'water' tokens represent H2O not H2O-or-XYZ even though they would be caused by H2O-or-XYZ. Fodor can avoid concluding that Oscar's 'water' tokens represents H2O-or-XYZ if he can show that the nomic dependence of 'water' on XYZ itself asymmetrically depends on the nomic dependence of 'water' on H2O. Fodor points out that we intend to use 'water' to refer to the natural kind to which the stuff around here belongs. This intention establishes the required asymmetric dependence. Thus Fodor remarks that 'Given that people are disposed to treat "water" as a kind term (and of course that the local samples are all in fact H2O) it follows that – all else being equal – they would apply it to XYZ only when they would apply it to H2O; specifically, they would apply it to XYZ only when they mistake XYZ for H2O; only when (and only because) they can't tell XYZ and H2O apart. Whereas, given a world in which they can tell XYZ and H2O apart (and in which their intentions with respect to "water" are the same as they are in this world), they will continue to apply "water" to H2O and refrain from applying it to XYZ'. 47

Notice the structure of Fodor's solution. The nomic dependence of 'water' tokens on XYZ is itself asymmetrically dependent on the nomic dependence of 'water' tokens on H2O, and that nomic dependence is "mediated" by the believer's intentions. It is the believer's thoughts about water – that it is a natural kind found around here, for example – that set up the required asymmetric dependence. But content holism now threatens, for whether my 'water' tokens represent H2O or H2O-or-XYZ is now determined in part by the contents of my other beliefs. This is a particularly ironic result as Fodor is perhaps the harshest contemporary critic of content holism. 48 I propose to close this Chapter by quickly sketching Fodor's anti-holism argument, and then showing how it has almost no applicability to his theory of content. 49

According to the doctrine of content holism, the content of any one propositional attitude in an intentional system is determined in part by the contents of the other propositional attitudes in that system. Since my total set of propositional

---

46 See Fodor 1990c: 114.
47 Fodor 1990c: 115.
48 See especially Fodor 1987, Ch. 3.
49 Fodor insists that all that matters is that the asymmetric nomic dependencies exist, not that they are mediated by other propositional attitudes. (See Fodor 1990c: 109-110.) But the existence of the asymmetric dependencies and their mediation by other propositional attitudes cannot be so readily pulled apart. If the asymmetric dependencies determine content, and if they in turn are in part determined by other propositional attitudes, then content is at least in part determined by other propositional attitudes. And that is exactly what the doctrine of content holism claims.
attitudes is almost certainly distinct from yours, it follows that none of my propositional attitudes are likely to be content-identical with yours: my belief that Saba is a tiger will differ in content from your belief that Saba is a tiger. Consequently, generalizations which range over all agents in circumstance C with propositional attitudes $P_1, P_2, \ldots, P_n$, are at best likely to be true of only one agent. Folk psychology (external), however, requires that there exist generalizations over agents in C with propositional attitudes $P_1, P_2, \ldots, P_n$, which are true of most (if not all) agents. Content holism thus represents a severe difficulty for folk psychology (external), and indeed for scientific intentional psychology.\textsuperscript{50}

The argument just sketched assumes that (external) folk psychological propositional attitudes are individuated by content. However, my belief that the Morning Star just rose is content identical to my belief that the Evening Star just rose – both are about the planet Venus – and yet they are distinct beliefs. It follows that (external) folk psychological propositional attitudes are not individuated solely by content; rather, they are individuated by both syntax and semantics. This result is hardly surprising given that we have already acknowledged (second objection, above) that subattitudinal concepts are individuated by both syntax and semantics.

One way to respond to the argument from content holism is, therefore, to turn revisionist. If we typed propositional attitudes purely syntactically – that is, purely narrowly – the argument from content holism collapses. Interestingly, Fodor argues in another context that we should revise folk psychology (external) and individuate propositional attitudes purely narrowly (see Section 8.2, below). I do not, however, wish to take the drastic step of revising the (external) folk psychological individuation of belief. Rather, I propose to argue that the content holism to which Fodor’s asymmetric dependency theory of content is committed is quite benign.

Recall the point at which Fodor appeals to the believer’s intentions and thus embraces content holism. According to Fodor, intentions sometimes mediate the asymmetric dependencies which establish what a mental concept represents. For example, it is my intention to use ‘water’ to represent the natural kind found in abundance around here which establishes that the nomic relation between the property of being XYZ and the property of being a cause of ‘water’ tokens is asymmetrically dependent on the nomic relation between the property of being H$\text{H}_2\text{O}$ and the property of being a cause of ‘water’ tokens, and thus ensures that my ‘water’ tokens represents H$\text{H}_2\text{O}$ not H$\text{H}_2\text{O}$-or-XYZ. In other words, my intention to use ‘water’ to represent the natural kind found around here in abundance determines that my ‘water’ tokens refer to H$\text{H}_2\text{O}$ rather than H$\text{H}_2\text{O}$-or-XYZ. It follows – other things being equal – that for your ‘water’ tokens to differ in content from mine you must lack the

\textsuperscript{50} See Fodor 1987, Ch. 3.
intention to use 'water' to represent the natural kind found around here in abundance. But now the original difficulty about content holism looks much less problematic. The content my 'water' tokens is determined not by my entire set of propositional attitudes but by whether or not I take 'water' to represent a natural kind found in abundance around here. Since it is typically very easy to determine if an intentional agent intends to use the token 'water' for the natural kind found around here in abundance, it is typically very easy to form classes of agents whose 'water' tokens have the same content. In other words, the destructive effects of content holism are strongly contained in Fodor's approach.

4.6 Concluding Remarks. I think it would be helpful to bring together the results of the last two Chapters. Following Armstrong and Lewis, I have argued that an (external) folk psychological belief is a locus in a causal network embracing sensory inputs, behavioural outputs, and other mental states. Stich's Mrs T example encourages us to see that the causal network involves actual as well as merely potential causal relations. It follows that folk psychology (external) is committed to what I have called ontological holism – the doctrine that an intentional system cannot consist of just one propositional attitude. Note that the doctrine of ontological holism is distinct from the doctrine of content holism discussed in the previous Section. The latter entails the former but not vice versa.

Propositional attitudes have content. Following Fodor I argued that the fundamental unit of content is the subattitudinal concept and that the content of a concept token is determined by asymmetric nomic dependencies. Also following Fodor, I endorsed the language of thought hypothesis. Whilst commonsense functionalism is not committed to the language of thought hypothesis, it is compatible with it. Grafting the language of thought hypothesis onto commonsense functionalism yields the claim that the characteristic causal roles identified by commonsense functionalism are in fact occupied by "mentalese" sentences.

Sentences in the language of thought have both syntactic and semantic properties. Two beliefs are widely identical if and only if their characteristic causal roles are occupied by sentence tokens in the language of thought which are both syntactically and semantically identical. On the other hand, two beliefs are narrowly identical if and only if their characteristic causal roles are occupied by sentences tokens which are syntactically identical. It follows that wide identity entails narrow identity but not vice versa.

The Morning Star/Evening Star case discussed in the previous Section led us to postulate the existence of what we might call 'synonyms' in the language of thought: concept tokens with identical content but divergent syntax. The 'Morning Star' and the 'Evening Star' are example of synonyms. Since they diverge in syntax,
the 'Morning Star' and the 'Evening Star' diverge in their causal properties. It follows that the belief that the Morning Star just rose will have distinct causal relations from the belief that the Evening Star just rose. And indeed it does: the former causes me to say, ceteris paribus, that the Morning Star just rose whilst the latter causes me to say, again ceteris paribus, that the Evening Star just rose.

Notice that the language of thought also contains what we might call 'homonyms': concept tokens with identical syntax but divergent content. Thus, the tokens 'H2O' and 'XYZ' have divergent content – 'H2O' refers to H2O whilst 'XYZ' refers to XYZ. They have, however, identical syntax. After all, the causal properties of Oscar's 'H2O' tokens are identical to the causal properties of Twin Oscar's 'XYZ' tokens.
CHAPTER FIVE

DEFENDING FUNCTIONALISM

5.1 Introduction. In Chapter 3 I argued that the mental states posited by folk psychology (external) are functional states; that is, they are the occupants of certain functional roles which mediate between sensory inputs, behavioural outputs, and other mental states. I am, then, committed to functionalism, although on this view functionalism is a meta-theory - a theory about the states posited by folk psychology (external). Functionalism has, though, been extensively attacked, principally by Ned Block and John Searle. It is not clear that Block and Searle explicitly saw themselves as attacking a theory of the states posited folk psychology (external). Nevertheless, their attack can be taken that way and loses none of its force by being thus construed.

In a nutshell, Block and Searle charge functionalism with being overly liberal; that is, both Block and Searle claim that the functionalist is forced to attribute mental states to systems which lack them. There are two obvious strategies for dealing with such a charge. (1) We can insist that the systems Block and Searle describe do in fact have mental states. (2) We can argue that functionalism is not committed to ascribing mental states to those systems. Against Block I adopt the second strategy. More specifically, I argue that commonsense functionalism, as described in Chapter 3, does not ascribe mental states to the system Block describes. Other sorts of functionalism do indeed succumb to Block's attack; I take that as evidence in favour of commonsense functionalism. Against Searle I try the other tack. Intentional states can quite properly be attributed to the system Searle describes - or at least to Searle's system suitably extended and developed. Searle's presentation, though, contains a rhetorical device that makes us reluctant to attribute mental states to the system he discusses. Or so I shall argue in Section 5.3.

There is another aspect to Block's charge of illiberalism: he claims that the functionalist is forced to attribute mental states to systems which lack the qualia characteristic of those states. This objection goes back to Wilfred Sellars, who both proposed a functionalist account of mental states and rejected it on the grounds that functionalism could not account for qualia. In Section 5.4 I discuss the qualia

---

1 See Block 1980, 1981; Searle 1980a.
2 As we shall see, Block argues that attempts to repair functionalism's excessive liberalism lead to chauvinism. I address these details in Section 5.2.
3 See Sellars 1963.
objection, focussing of Frank Jackson's perspicuous presentation of the qualia problem. 4

Before getting to the hard cases, though, it is worth spending just a few minutes dealing with yet another objection to functionalism. This objection is a close cousin of the objections I will consider in detail in Sections 5.2 and 5.3. Various philosophically identical versions of the objection exist, trading under a variety of folksy names. I will focus on the so-called 'bucket of water' example which is widely attributed to Ian Hinkfuss. 5

Imagine a bucket of pond water standing in the sunlight. A vast and complex array of dynamic interactions ceaselessly occur. Solar radiation is heating the water's surface, causing a sharp temperature differential between the top and bottom strata. Complex convection currents result, stirring the water gently but continuously. Heat is being lost through the bucket's walls, and H2O molecules are being driven off the water's surface. Oxygen is dissolving into the water from the atmosphere whilst carbon dioxide, produced by the metabolic activity of aerobic bacteria, is being given off. Microbes flourish, divide, and die. Photosynthesizing algae blooms and even a tadpole struggles in the depths. Enough said. We are dealing with a very complex, highly interrelated, system.

The mind, too, is a very complex, highly interrelated, system. Hinkfuss' worry is that functionalism does not contain the resources to separate the mind from the bucket. Functionalism abstracts away from the physical details of the system. That is one of its great virtues as an account of the mental states posited by folk psychology (external). Folk psychology (external) accepts that mental states can be multiply realized, and functionalism explains how that is possible. But, argues Hinkfuss, the bucket of pond water standing in the sunlight might just happen (by an extraordinary coincidence) to occupy the functional roles identified by the functionalist with mental states. The functionalist is therefore committed to attributing mental states to such a bucket. But that is absurd. Buckets of pond water do not have beliefs and desires, so there is something wrong with functionalism.

We can convey the force of Hinkfuss' objection by applying to our scientific claims about the bucket the Ramsey sentence approach that Lewis applied to our commonsense platitudes about the mind (see Section 3.4, above). Recall Lewis' idea. Form a vast conjunction of all the (external) folk psychological platitudes which are common knowledge amongst us. It will contain platitudes of the form: 'If a person is in mental state M1 and experiences sensory input S, then she will move into mental state M2 and produce behaviour B'. Replace the mental state terms with free variables.

4 See Jackson 1982. See also Jackson 1986.
5 See Sterelny 1990, Ch. 1.
and prefix the appropriate existential quantifier binding those variables. The resulting sentence says that there exists an n-tuple of entities which occupy the roles specified by our everyday talk about mental states. Note that the sentence will contain no mental state terms, for they were all systematically replaced by free variables. Now form a vast conjunction of all the scientific "platitudes" which describe the complex activities going on in the bucket. It will contain statements like 'If the bucket is in physical/chemical state $S_1$ and solar energy $E$ is transmitted across its surface, then the bucket will move into physical/chemical state $S_2$ and oxygen will be released'. Systematically replace all the expressions referring to the physical and chemical states of the bucket by free variables, and prefix the appropriate existential quantifier binding those variables. The resulting sentence says that there exists an n-tuple of entities which occupy the roles specified by our scientific talk about the bucket. Note that the sentence will contain no terms referring to the physico-chemical states of the bucket, for they were all systematically replaced by free variables. And of course it could turn out that the Ramsey sentence just formed about the bucket is identical to the one formed from our (external) folk psychological platitudes. That is, it could turn out that various states of the bucket occupy exactly the functional roles specified by folk psychology (external).

It might be objected that it is incredibly unlikely that a bucket of pond water could occupy the functional roles implicit in folk psychology (external), and thus the objection does not have to be taken seriously. But functionalism (as I am construing it) claims that (external) folk psychological states are to be identified with the occupants of certain functional roles: if the role is occupied the state exists. So the mere improbability of the state being occupied is irrelevant. To put this point another way, like Lewis I claim it to be a conceptual truth that (external) folk psychological mental states are identical with the occupants of certain functional roles. Such truths hold in all worlds, including the distant one in which a bucket of water comes to occupy the functional roles implicitly specified by folk psychology (external).

I do think, though, that pointing out the sheer implausibility of a bucket of water occupying the relevant functional roles does have a part to play in defusing this sort of objection. I want to bite the bullet and insist that if a bucket of pond water really did come to occupy the functional roles specified by folk psychology (external) then it really would have beliefs and desires. We are tempted to deny that the bucket of water has beliefs and desires because we fail to reflect upon what a very special bucket of water this is. We simply imagine any old bucket and then ask ourselves if that bucket is likely to have beliefs and desires. And of course the answer is 'No'. It is very unlikely that that bucket will have mental states because it is very unlikely that it occupies the functional roles specified by folk psychology (external). Nevertheless, were it to come (incredibly) to occupy the (external) folk psychological roles, it would
have (external) folk psychological mental states. To insist that the bucket would have
mental states if it occupied the relevant functional roles is just to remind ourselves that
mental states can be multiply realized. In Lewis' Martians the functional roles
specified by folk psychology (external) are occupied by the states of various hydraulic
devices. The bucket case is on a par with Lewis' Martian example: in both cases the
functional roles characteristic of mental states are occupied in rather fanciful ways.

There is a lesson in rhetoric to be learnt here - a lesson we will have cause to
recall before the end of the Chapter. In the next Section I turn to Block's look-up tree
example, known rather affectionately as the 'Blockhead'.

5.2 The Blockhead. Block asks us to imagine a chess playing robot whose
internal structure is as follows. Consider a tree - strictly, a directed graph - such that
every root-to-tip branch represents a complete game of chess. Each node is a
representation of a possible state of the board, and an arc connecting two nodes
represents a single legal move which, were it performed on the board as represented
by the first node, would yield the board as represented by the second. There are two
sorts of arcs - call them 'white arcs' and 'black arcs'. The white arcs represent moves
available to the white player; the black arcs represent moves available to the black
player. The arcs - and hence the graph itself - are directed in that not all chess moves
are reversible: it may be possible to get from board configuration A to board
configuration B by a single legal move, but not vice-versa. (Reversible moves would
be represented by a pair of oppositely directed arcs.) A panel of human chess experts
reviews the tree. Any arc which represents a move they would not themselves make
under the circumstances represented by the first node, is erased. The tree is then
etched onto a silicon chip which in turn is implanted in a suitable robot. The robot
plays chess by locating the node which represents the current board configuration and
executing any move represented by an appropriately coloured arc directed away from
that node.

The robot just described is input-output identical to a human chess expert
within the chess-playing domain. Of course, outside that domain it flounders
completely. Block therefore extends his example by describing a robot which is input-
output identical to an intelligent human being in any domain. In the new robot, each
node represents the complete sensory input which could be experienced by the robot at
any moment, and each arc represents an intelligent output given that input. In what
follows I will focus, for simplicity, on Block's chess playing machine. The argument
equally applies, however, to Block's more complicated example.

6 See Lewis 1980.
7 See Block 1980. See also Block 1981.
How does Block's thought experiment constitute a challenge to functionalism? Block states that 'functionalism can be said to "tack down" mental states only at the periphery — ie through physical, or at least nonmental specifications of inputs and outputs'. But Block's robot is input-output identical to a human chess expert, so the functionalist must ascribe to it those mental states we are prepared to ascribe to a human chess expert; or at least the functionalist must ascribe to Block's robot those mental states we are prepared to ascribe to a human chess expert qua chess expert. Once we are familiar with the robot's internal structure, however, we have no inclination to ascribe mental states to it. Block therefore charges functionalism with excessive liberalism: functionalism ascribes mental states to a robot which clearly lacks them.

It might be thought that we can argue against Block by pointing out that his robot is physically impossible in quite a strong sense. The task we envision the robot executing is, strictly speaking, computable. However, the directed graph which lies at the heart of Block's thought experiment is extraordinarily vast. To make things easier, let us restrict attention to chess games no more than 40 moves deep. It has been estimated that there are about $10^{100}$ such games. The universe is believed to be about 15 billion years old so, even working solidly from the Big Bang to the present, the panel of experts pruning the directed graph would have had to have check approximately $2 \times 10^{83}$ games per second to have completed their task by now! Block's robot is unbuildable in anything like real time and therefore his thought experiment tells us nothing about the nature of mental states. Or so the objection might run.

This objection is, of course, just an analogue of the objection against the bucket of water example we saw in the last Section. We were encouraged to dismiss the bucket of water example on the grounds that it is highly unlikely that such a bucket could exist, and we are now being encouraged to ignore Block's example on the grounds that the Blockhead is unbuildable in real time. But the objection misses its mark. For like Lewis I take the identification of (external) folk psychological mental states with the occupants of certain functional roles to be a conceptual truth — a truth which follows from the meanings of (external) folk psychological predicates. (Think of how we arrived at the commonsense functionalist position: we attended to various (external) folk psychological platitudes.) Conceptual truths are true at all worlds, including worlds in which constructing the directed graph for Block's robot is a straightforward engineering problem. Therefore we cannot shield commonsense functionalism from Block's attack by pointing out how distant from the actual world...
the world of Block's robot is. It is enough for Block's robot to be logically possible for it to be a challenge to commonsense functionalism.

Note that the argument to excessive liberalism is underpinned by a particular conception of functionalism which I will call 'input-output functionalism'. According to input-output functionalism, mental states supervene entirely on inputs and outputs; that is, if two systems are input-output identical they enjoy identical mental states. It is by assuming input-output functionalism that Block makes the crucial move from the input-output identity of the robot and a human chess expert to their having identical mental states. Block's charge of excessive liberalism could be blocked, therefore, by locating a form of functionalism that does not license the crucial inference from input-output identity to identity of mental states. We require a form of functionalism that does not "tack down" mental states only at the [sensory and behavioural] periphery. Block himself considers a form of functionalism that meets this requirement — a form of functionalism he calls psychofunctionalism. According to the psychofunctionalist, mental states are the occupants of functional roles characterized in terms of human information processing; that is, in terms of the psychophysical details of the human cognitive apparatus. The psychofunctionalist is therefore not committed to attributing mental states to Block's chess-playing robot. After all, the robot processes information in a fashion radically distinct from a human chess expert. But, Block argues, psychofunctionalism, with its emphasis on human information processing, escapes the charge of excessive liberalism only by becoming unacceptably chauvinistic. For surely there can exist believers and desires who do not process information in just the way that we do.

Block's anti-functionalist argument thus takes the form of a dilemma. The functionalist must choose between input-output functionalism and psychofunctionalism. Input-output functionalism is unacceptably liberal; psychofunctionalism unacceptably chauvinistic. So much the worse for functionalism. But it should be clear that this is a false dilemma. For commonsense functionalism, defended in Chapter 3 as providing the correct analysis of (external) folk psychological mental states, is distinct from both input-output functionalism and psychofunctionalism. Moreover, commonsense functionalism is neither overly liberal nor chauvinistic. We can begin to see this by returning to an example of Putnam's introduced in Chapter 3, above.

Recall Putnam's example of a perfect pain pretender.9 A perfect pain pretender never experiences pain but always produces pain behaviour when subject to the sort of sensory stimulation which typically causes pain. For example, when the perfect pain pretender touches a hot stove she always flinches and cries out even though she

---

9 See Putnam 1980.
experiences no painful sensations. Putnam offers the perfect pain pretender example as part of an attack on behaviourism. The perfect pain pretender challenges the sufficiency of the behaviourist analysis of pain: the conditions spelt out by the behaviourist analysis are met but the pretender is not in pain. Note, though, that Putnam's perfect pain pretender example also provides the basis for an attack on input-output functionalism. For the perfect pain pretender is input-output identical to a person who genuinely is in pain. Thus, since input-output functionalism holds that mental states supervene on inputs and outputs, the input-output functionalist is committed to attributing pain to the perfect pain pretender. But, by hypothesis, the perfect pain pretender is not in pain.

As we saw in Chapter 3, commonsense functionalism deals smoothly with Putnam's perfect pain pretender example. Commonsense functionalism identifies mental states with the occupants of certain functional roles which mediate between sensory inputs, behavioural outputs, and other mental states. It is therefore not committed to attributing pain to the perfect pain pretender because the pretender's behaviour is mediated by a set of beliefs and desires characteristic not of pain but of pretending to be in pain. For example, the pain pretender desires to appear to be in pain and believes that she can appear to be in pain by behaving thus and so. Such beliefs and desires play no part in the production of veridical pain behaviour. By appealing to mental states, the commonsense functionalist can distinguish the person truly in pain from the pretender, and thus avoid the problem which beset both behaviourism and input-output functionalism.

Notice that Block's robot is a perfect pretender: it is input-output identical to a human chess expert but does not actually possess the mental states which we attribute to the human chess expert. The input-output identity of the robot with a human chess expert commits the input-output functionalist to ascribing mental states to the robot, just as the input-output identity of the perfect pain pretender with an ordinary human being who is in pain commits the input-output functionalist to ascribing pain to the perfect pain pretender. But we have already seen that commonsense functionalism has no difficulty handling Putnam's perfect pretender case; similarly, commonsense functionalism has no difficulty handling Block's robot case. The human chess player's behavioural output is mediated by a rich network of beliefs and desires about the game - beliefs and desires which are wholly absent in the robot's case. Thus, since according to commonsense functionalism mental states are loci in causal networks embracing inputs, outputs, and other mental states, the commonsense

---

10 Block himself has noted the parallel between his robot and Putnam's perfect pain pretender. See Block 1981: 22.
functionalist is not committed to attributing the same mental states to the human chess expert and the Blockhead.

Notice that it is not merely that the robot's input-output profile is mediated by the wrong set of beliefs and desires. Rather, commonsense functionalism holds that the robot has no beliefs and desires at all. Recall Stich's Mrs T case, outlined in Section 3.4, above. Mrs T suffered from a degenerative brain disorder which steadily eroded her corpus of beliefs. Thus, as a young woman Mrs T had a wide knowledge of American politics, and could discourse extensively on McKinley's assassination and its political consequences. As the disease progressed, however, her grip on these events became more and more tenuous. She could still respond 'He was assassinated' to the question 'What happened to McKinley?', but she could not say who McKinley was, what an assassination was, or whether McKinley was alive or dead.¹¹ We do not ascribe to Mrs T the belief that McKinley was assassinated. She is merely mouthing the words; she no more believes that McKinley was assassinated than does a cassette recorder on which that sentence has been taped.

Commonsense functionalism has no difficulty accounting for our reluctance to ascribe the belief that McKinley was assassinated to Mrs T. For commonsense functionalism is committed to ontological holism; that is, beliefs necessarily have actual causal relations to other mental states. Thus, the belief that McKinley was assassinated is in part individuated by a network of actual causal relations which involves the belief that McKinley is dead; that death terminates life; that the dead play no active role in politics; and so forth. No such rich structure of actual causal relations exists in Mrs T's case, and thus commonsense functionalism is not committed to attributing to her the belief that McKinley was assassinated. Similarly, commonsense functionalism identifies the belief 'My king is being threatened' with a node in a network of actual causal relations which involves the belief that the king is a chess piece; that the object of the game is to defend the king; that the game is finished when the king can no longer be moved into a position in which it is secure; and so forth. But whilst Block's robot has internal representational states (the nodes and arcs of the directed graph), those states are not richly structured in the way demanded by belief ascription. We cannot attribute to Block's robot the belief that his king is being threatened, for the robot has no internal representational state which has actual causal relations to other states representing the purpose of the game, the nature of moving into and out of check, and so forth. Block's robot is directly analogous to Stich's Mrs T, and commonsense functionalism refuses to attribute mental states to either of them.

By individuating beliefs in part holistically - as loci in extensive networks of actual causal relations involving other mental states - commonsense functionalism

respects the (external) folk psychological intuition that neither Mrs T nor Block's robot have beliefs. Commonsense functionalism is thus not illiberal. It does not ascribe mental states to Block's robot. But what about Block's other charge? Is commonsense functionalism excessively chauvinistic?

Recall how Block's charge of chauvinism arose. We recognize that Block's robot lacks mental states because we are familiar with its internal structure—a structure incompatible with the possession of mental states. What is the basis for that judgement? Block objects to the idea of a universal psychology, apparently because of the difficulty of locating generalizations capable of covering all possible cognizers. Our judgement that the internal structure of Block's robot is incompatible with the possession of mental states cannot be grounded, therefore, in universal considerations. Rather, our refusal to ascribe mental states to Block's robot must be grounded, according to Block, in a comparison of its psychophysical structure with our own. Thus functionalism escapes the charge of liberalism only by emphasizing human information processing characteristics, an emphasis which seems unacceptably chauvinistic.

We have already seen how commonsense functionalism defuses Block's charge of excessive liberalism. But it should be clear that commonsense functionalism can defuse Block's other charge as well. For Block's argument to chauvinism proceeds via a denial of the possibility of a universal psychology. But commonsense functionalism is such a psychology. The analyses of mental states offered by commonsense functionalism are universal in the sense that they yield conceptual truths. As such they apply to all possible believers and desirers. Commonsense functionalism is thus not chauvinistic as charged by Block. In particular, it abstracts away from the psychophysical details of human information processing.

In this Section I have defended commonsense functionalism against Block's disjoined charges of chauvinism and excessive liberalism. Commonsense functionalism is neither. In the next Section I turn to John Searle's Chinese room argument.

5.3 The Chinese Room. Searle's main target in his 'Minds, Brains, And Programs' is what he calls 'strong artificial intelligence (AI). Strong AI is the thesis that 'the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states'. I intend to construe Searle's attack on strong AI as an attack on functionalism. In principle, a computer could be programmed to occupy those states

---

12 See Block 1980: 292-293.
which, according to functionalism, are identical with mental states. Thus, if functionalism is true, so is strong AI. By contraposition, if strong AI is false then functionalism is false. The difference between the two theses is largely a matter of emphasis. Hence, if we are to defend functionalism we must say something plausible about Searle's objection to strong AI.

By way of background to Searle's objection, let me just briefly remark that Searle focusses on Roger Schank's attempts to program a computer to answer questions about a story. Schank's approach to this problem involves giving the computer a 'script', that is, a knowledge base about the domain in which the action of the story is played out. The computer is then programmed to utilize the script when answering questions about the story. Thus for example Schank has written a restaurant script which the program can access when answering questions about stories set in restaurants. If Schank has got the script and program right, then, according to strong AI, the computer literally understands the story. Searle attempts to undermine strong AI by locating an example in which the program is executed but understanding is absent. In other words, he seeks to show that following a program is not sufficient for understanding. Here is how Searle sets up his objection:

Suppose that I am locked in a room and given a large batch of Chinese writing. Suppose furthermore . . . that I know no Chinese, either written or spoken . . . . Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all these symbols call the first batch "a script", they call the second batch "a story", and they call the third batch "questions". Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions", and the set of rules in English that they gave me, they call "the program" . . . . Suppose also that after a while I get so good at following the instructions

for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view . . . . my answers to the questions are indistinguishable from those of native Chinese speakers . . . . [Nevertheless,] as far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program. [But] it seems to me to be quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of a Chinese speaker, and I can have any program you like, but I still understand nothing.\footnote{Searle 1980a: 417-418.}

Searle's objection, construed as an attack on functionalism, is this. The Chinese room occupies the functional roles characteristic of understanding but clearly understands nothing. The occupation of functional roles therefore cannot be sufficient for the having of mental states. But functionalism insists that mental states are nothing over and above the occupants of certain functional roles. Therefore functionalism is false.

How can Searle's attack be resisted? Unfortunately, Searle muddies the waters considerably by introducing talk about syntax and semantics. Thus in a separate publication he summarizes the Chinese room argument as follows:

\begin{enumerate}
\item Programs are formal (syntactical).
\item Minds have contents (semantic contents).
\item Syntax is not identical nor sufficient by itself for semantics.
\end{enumerate}

Therefore

\begin{enumerate}
\item Programs are not sufficient for nor identical with minds.\footnote{See Searle 1991: 526.}
\end{enumerate}

Searle's idea seems to be that the Chinese room deals solely with syntax. Minds, on the other hand, have thoughts which in turn have semantic properties. Syntax alone does not yield semantics. Therefore the Chinese room is not a mind. Nobody is going to quibble with premiss (3), but, as Jerry Fodor has pointed out,
symbols acquire semantics in virtue of being in certain relations to the world.\(^{18}\) (See Chapter 4, above, for an extensive discussion of what the content-conferring relations might be like.) Thus, if Searle's example is suitably augmented so that the Chinese symbols enjoy the appropriate, content-conferring relations, Searle's worry about semantics evaporates.

Searle's concern about \textit{understanding}, though, remains in place. Searle has pointed out to Fodor that, even if the content-conferring relations were in place, the agent within the Chinese room would still not understand Chinese.\(^{19}\) The possibility arises, though, that the entire room, rather than merely the agent within it, understands Chinese. Functionalism is committed to the view that, if the functional role identified with understanding Chinese is occupied by states of the room, the room really does understand Chinese. This could be true, however, without it being the case that the agent within the room understands Chinese. The agent is but a subsystem of the larger system, and so the system can have properties that she lacks. Searle's example thus contains a clever rhetorical device. We would not understand Chinese if we were in the room, so by encouraging us to imaginatively identify with the agent within the room, Searle deflects us from recognizing that the room itself understands Chinese.

Searle has considered the possibility that the room as a whole (rather than the agent within it) understands the Chinese story.\(^{20}\) He responds—ingeniously in my view—by pointing out that if the agent were to memorize the script and program, states \textit{within her} would occupy the causal roles identified by functionalism with understanding, but she still would not understand Chinese. This response, though, involves an equivocation between \textit{understanding} and \textit{believing that one understands}. A cognitive system can be said to understand a domain if it exhibits a certain skill or facility in that domain. Thus it can be said, for instance, that I understand how to use the computer on which I am presently typing; that I understand English; that I understand the boss' instructions. In each case what is being said is that I have a certain skill or facility in the stated domain: I can operate the computer successfully; I can read, write, and converse in English; I know what the boss expects of me. Note that one's skills and facilities can be the objects of one's own beliefs. Thus when someone acquires the skill or facility characteristic of understanding, they typically believe that they have done so.\(^{21}\) It is clear that understanding and believing that one understands are distinct. Someone can have the relevant skill or facility and yet not believe that they understand; and someone can believe that they understand and yet in

\(^{18}\) See Fodor 1980.

\(^{19}\) See Searle 1980b: 454.


\(^{21}\) When I believe that I understand \(p\) I often experience a characteristic feeling: the feeling of understanding. In this Section I want to focus on the belief that one understands \(p\), rather than the accompanying qualia. I discuss qualia in Section 5.4.
fact lack the relevant skill or facility. It follows that the agent in Searle's example may understand Chinese without believing that she does so; that is, she may have the relevant skill or facility but not realize it. Thus, that we do not imagine ourselves believing that we understand Chinese when we imaginatively identify with the agent who has memorized the script and program, does not entail that understanding is absent.

Nevertheless, it might still seem that something is missing. Merely memorizing a script and a list of instructions would not, it seems, give rise to understanding in the relevant sense. I suggest that our reluctance to ascribe understanding in such a case comes about because we fail to adequately reflect upon what would in fact be involved in memorizing the script and program. Understanding a story — Macbeth, say — involves having a very large corpus of beliefs about the physical and social worlds, and having those beliefs in turn involves having yet further beliefs. Moreover, we would not say that someone understood Macbeth unless they also had the capacity to follow, say, A Tale Of Two Cities. (The system which can reply only to questions about Macbeth is akin to Stich's Mrs T who could only answer the question 'What happened to McKinley?'.) The script and program required to generate understanding of Macbeth would thus be enormous. If an agent mastered this wealth of details — if she came to hold all the relevant beliefs about the social and physical worlds — then she really would understand the story. And notice that we would no longer be tempted to withhold the ascription of understanding from her. We ascribe understanding to people who possess (and can appropriately manipulate) the details relevant to the domain in question.

Such a person, though, need not believe that they understand: they may still not have within them a state which occupies the functional role identified with believing that one understands. However, whilst they might not have the relevant functional role occupied within them, I think that any normal person in their position would in fact come to have that role occupied. After all, it is usually the case that when we come to understand something we also come to believe that we understand. This suggests that the functional role characteristic of believing that one understands typically comes to be occupied as a direct causal consequence of the functional role characteristic of understanding being occupied. So when the person comes to understand the Chinese story they will in all probability also come to believe that they understand. Note, though, that the Chinese room may not be constituted such that the occupation of the role characteristic of understanding leads to the occupation of the

---

22 This suggests that building a machine to which we could truthfully ascribe (external) folk psychological mental states would in practice be very difficult. In other words, there are real world limits that strong AI is unlikely to overcome.
role characteristic of believing that one understands. It may have no beliefs at all about its own mental states.

5.4 Troubles With Qualia. Commonsense functionalism identifies mental states with the occupants of certain causal roles. Thus pain, for example, is identified with the occupant of what we might call the pain role. Presumably, in humans the pain role is occupied by a neural state; that is, by a physical state. This immediately suggests the following difficulty.\(^\text{23}\) There is a characteristic feel to pain: it hurts. To use Thomas Nagel's expression, there is something *that it is like* to be in pain.\(^\text{24}\) How can we account for the hurtfulness of pain in terms of the brain states with which pain has been identified? How does the merely physical give rise to the mysteriously phenomenological?

Sensory states like pains are not the only states that have phenomenological properties. Thus there is something that it is like to see a red surface in good light. Following standard usage I will use the expression 'qualia' to refer to the something that it is like to have a pain or see a red surface. The difficulty we face is accounting for qualia within the physicalist framework to which commonsense functionalism is committed.\(^\text{25}\)

I want to discuss three separate aspects of the qualia problem. First, I want to consider the claim that the incorrigibility of our qualia reports is strong evidence for the immateriality of qualia, and is thus strong evidence against the identification of mental states with brain states standardly made by the commonsense functionalist. Second, I will argue that our inability to account for qualia in physical terms is a reflection upon our limited cognitive capacities rather than evidence for immaterialism. My argument here is drawn from the work of Colin McGinn.\(^\text{26}\) Finally, I will consider Frank Jackson's particularly elegant formulation of the qualia problem.\(^\text{27}\)

At least since Descartes philosophers have been tempted by the view that we stand in a special, privileged, epistemic relation to our qualia. My belief that I have just touched a hot stove may be mistaken, but I cannot be mistaken that I am experiencing pain. If I think I am in pain then I am in pain, and if I am in pain then I think that I am. Similarly, whilst I might be mistaken that I am standing before a red

\(^\text{23}\) Suggests rather than entails. As Colin McGinn has pointed out, one of the peculiarities of the qualia problem is the difficulty of formulating it in a rigorous way. (See McGinn 1991, Ch. 1, fn 1.)

\(^\text{24}\) See Nagel 1973.

\(^\text{25}\) In principle, commonsense functionalism could identify the occupant of the characteristic causal role R with an immaterial object. Commonsense functionalism is, to use J.J.C. Smart's Rylean phrase, *topic neutral*: it is neutral about the ontological status of the occupiers of the folk-specified roles. However, commonsense functionalists typically expect the roles characteristic of (external) folk psychological mental states to be occupied by brain states. See for example Armstrong 1968; Lewis 1966.

\(^\text{26}\) See McGinn 1991, especially Ch. 1.

\(^\text{27}\) See Jackson 1982. See also Jackson 1986.
table, I cannot be mistaken that I am having an experience of redness. If I think I am having a red experience then I am having a red experience, and if I am having a red experience then I think that I am. What sort of things must qualia be if they are to be such that my epistemic access to them is thus guaranteed? Obviously they must be very special things – things quite different from the material objects of the external world about which there is no such guarantee of epistemic success. Thus we are led quite naturally to the idea that qualia are not material objects but immaterial objects. Since qualia are immaterial they cannot be identified with the physical properties of the brain, and they thus raise a problem for the functionalist who seeks to identify mental properties with the occupants of certain causal roles and thereby with brain processes.

I want to resist the claim that we have privileged access to qualia, and thus pull away one of the props supporting the idea that qualia are immaterial and so cannot be accounted for by commonsense functionalism as it is ordinarily understood. We can begin by considering two examples put forward by Paul Churchland.28 These examples can be resisted in the sense that a determined immaterialist can refuse to concede them without losing too much credibility. Nevertheless, the examples point the way to developing a powerful argument against the idea that we have privileged access to our qualia.

First example. You are being tortured by repeated applications of a red hot metal bar to your bare back. You are blindfolded and thus cannot see the bar. Nevertheless, you can certainly feel it, and it feels painful. On the twentieth occasion your captor presses against your back not the hot bar but a block of ice. Your initial response is to cry out, for the block of ice feels just as the hot bar did. After a moment's reflection, though, you realize your mistake. You were not in pain at all; it only momentarily felt as though you were. Thus the having of a pain and the qualia of pain can come apart. You can think you are in pain without being in pain.

Second example. Due to an unusual medical condition, your C-fibres have failed to work for the last fifty years. Consequently, you have had no experience of pain during that period. Medical science discovers the cure for your condition. The cure is instantaneous: a single injection is all that is required to immediately restore your C-fibres to normal function. Soon after the administration of the injection you stand on a tack. You do not immediately realize that you are in pain – after all, it is fifty years since anything like this last happened to you. You are in pain but you do not think that you are.

Churchland's examples suggest that pain and the belief that you are in pain can come apart and thus that sensation reports are not incorrigible. As I said above, these examples are not entirely compelling: the advocate of incorrigibility can insist with

28 See P.M. Churchland 1988: 77-78.
some plausibility that the cases have been misdescribed. They do, however, function
nicely as an introduction to Wilfred Sellars' powerful argument against the
incorrigibility of sensation reports. 29

Sellars argues against the incorrigibility of qualia reports by constructing a
"myth" which describes a way in which our mental state talk could have arisen. The
myth is not to be read as history – Sellars does not believe that the events he described
actually took place. Rather, the myth shows that our mental state talk could have
 arisen in a way that does not entail privileged epistemic access to our qualia and thus
casts doubt on the claim of incorrigibility.

Sellars' myth unfolds in three stages. The first stage is 'a stage in pre-history
in which humans are limited to what I shall call a Rylean language, a language in
which the fundamental descriptive vocabulary speaks of public properties of public
objects located in Space and enduring through Time'. 30 At this stage, our 'Rylean
ancestors' have no terms in their language for beliefs, desires, and other mental states.
In the second stage of the myth, Sellars describes how it came about that inner causes
of behaviour are first postulated by 'a genius – let us call him Jones'. 31

[IlIn the attempt to account for the fact that his fellow men behave
intelligently not only when their behavior is threaded on a string of overt
signals – that is to say, as we would put it, when they 'think aloud' – but
also when no detectable verbal output is present, Jones develops a theory
according to which overt utterances are but the culmination of a process
which begins with certain inner episodes. And let us suppose that his
model for these episodes which initiate the events which culminate in overt
verbal behavior is that of overt verbal behavior itself. In other words,
using the language of the model, the theory is to the effect that overt verbal
behavior is the culmination of a process which begins with 'inner
speech'. 32

At the second stage of the myth, just described, the theory that overt behaviour
is the culmination of inner episodes (specifically, 'inner speech') is applied only to
other people. In the third stage, Jones and his friends learn to apply it to themselves.
Initially they infer the existence of inner causes of behaviour in themselves by
applying the theory to their observations of their own behaviour, but finally they learn
to recognize the inner causes of behaviour through introspection:

29 See Sellars 1956.
30 Sellars 1956: 309.
31 Sellars 1956: 314.
32 Sellars 1956: 317-318. Emphasis Sellars'.
Once our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compatriots to make use of the theory in interpreting each other's behavior, it is but a short step to the use of this language in self-description. Thus, when Tom, watching Dick, has behavioral evidence that warrants the use of the sentence (in the language of the theory) "Dick is thinking 'p'" . . . . Dick, using the same behavioral evidence, can say, in the language of the theory, "I am thinking 'p'" . . . . And it now turns out — need it have? — that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly, by applauding utterances by Dick of "I am thinking that 'p'" when the behavioral evidence strongly supports the theoretical statement "Dick is thinking that 'p'"; and frowning on utterances of "I am thinking that 'p'", when the evidence does not support this theoretical statement. Our ancestors begin to speak of the privileged access each of us has to his own thoughts. What began as a language with a purely theoretical use has gained a reporting role.33

According to Sellars' myth, our introspective reports are theory-driven and therefore not incorrigible — the theory may be false or it may have been misapplied. Since we cannot rule out the possibility that our introspective reports are theory-driven, we cannot rule out the possibility that they are not incorrigible. Sellars has thus provided a powerful argument against the incorrigibility of introspection.

The argument against incorrigibility just sketched takes as its premiss the possibility that our introspective reports are theory-driven. According to the theory (internal), our introspective reports actually are theory-driven: they are driven by folk psychology (internal). On this view, Paul Churchland's examples, given above, involve cases in which highly unusual circumstances lead to folk psychology (internal) being misapplied. There exists, though, a much more powerful set of considerations pressing the immaterialist's viewpoint upon us. It is to those considerations I now turn.

Immaterialism about qualia draws much of its force from our inability to offer plausible physicalist accounts of qualia. A variety of attempts to identify qualia with properties which are broadly speaking physical have been made, but they invariably fail to convince.34 Our inability to account for qualia in physicalist terms poses the

33 Sellars 1956: 320. Emphasis Sellars'.
34 See, for example, Johnson-Laird 1983, Ch. 16.
following dilemma. Either qualia are non-physical or else they do not exist at all. The latter option – eliminativism about qualia – is implausible. The metaphysics of our internal life might be an open issue, but that we have an internal life seems undeniable. We are thus forced to conclude that qualia are immaterial.

This unwelcome conclusion can, however, be resisted: we can insist that our inability to locate a satisfying physicalist account of qualia is a reflection on our limited cognitive capacities rather than proof of the immateriality of qualia. This approach to the problem of qualia is Colin McGinn's, and I draw heavily upon his work.35

How could it be that qualia are in fact physical and yet we are unable to locate a physicalist explanation of their existence? We can make some progress towards answering this question by introducing the idea of cognitive closure. A cognitive system is closed with respect to a property if it is incapable of forming a concept of that property.36 Thus, for example, the mind of a dog is (I take it) closed to the property has prime factors 3 and 5. Presumably, the human mind too is closed to certain properties – we are to some properties as the dog is to has prime factors 3 and 5. Indeed, to think otherwise is to flirt with idealism. A robust realism demands that we concede that there could be properties that lie forever beyond our ken.37 But the task at hand requires more than just conceding cognitive closure. We need to show not that there are, in all probability, properties we cannot grasp; rather, we need to show that the property which, if grasped, would make qualia intelligible to us, is beyond our reach. In other words, we need to show that our minds are cognitively closed with respect to property $P$, where property $P$ is that physical property which, if physicalism is true, figures significantly in explaining how qualia depend on brain states. If we could show that our minds would in fact be closed to property $P$, we would be able to resist immaterialism about qualia: better to posit a physical property – even an unknowable one – than a non-physical property. Considerations of parsimony weigh so heavily against the existence of non-physical properties that we would be well advised to accept unknowable physical properties than dabble in the immaterial.

We might try to argue to the claim that our minds are closed with respect to $P$ by pointing out that, despite years of efforts, no physicalist solution to the qualia problem is forthcoming. But this argument is doubly dubious. To begin with, it may be that we have not tried hard enough, or that the solution is "just around the corner".38 More importantly, notice that whilst the truth of the cognitive closure claim would explain our historical failure to locate property $P$, our historical failure to locate

---

35 See McGinn 1991, especially Ch. 1.
36 See McGinn 1991: 3.
$P$ is equally well explained by immaterialism. If immaterialism about qualia is true, then of course we cannot locate property $P$—property $P$ does not exist! We need to look elsewhere for support for the cognitive closure principle. In particular, we need an argument that shows why $P$ is closed to minds like ours.

Such an argument might be constructed as follows. Think first about the possibility of having direct perceptual or introspective access to $P$. That we might see $P$ is implausible. No amount of looking at the brain is going to reveal the property which, if grasped, would allow us to understand how qualia arise from grey matter. And of course this remark applies to the other perceptual modalities: we cannot expect to directly smell, touch, hear, feel, or taste $P$. Whatever $P$ is, it is not the sort of thing that we can directly perceive. Nor, for that matter, is $P$ the sort of property that we can expect to observe with instruments (microscopes and the like) that enhance our perceptual powers. In insisting that this is the case I am, of course, just airing the sort of intuitions that gave rise to the qualia problem in the first place. I have no argument to offer here; I am instead relying on the brute intuition that the mystery of consciousness will not yield to simple perception of the brain. Let us turn then to introspection. Could we directly introspect $P$? Again it seems the answer is no. I argued above against the incorrigibility of qualia reports. Nevertheless, I take it that introspection is on the whole trustworthy. Except in the sort of unusual circumstances to which Paul Churchland drew our attention, introspection seems to be highly reliable. But that reliability has a price. Reliability is purchased at the cost of limited range: introspection yields nothing but the contents of consciousness. Indeed, my introspecting yields nothing but the contents of my own consciousness. But property $P$ is not part of the content of my consciousness. Therefore introspection will not reveal property $P$.39

Science's capacity to postulate new properties is not rigidly bound by the limits of introspection and perception, whether unaided or otherwise. We also infer the existence of new properties by argument to the best explanation. Could it be that we can discover property $P$ by this method? Let us consider in a very general way what an argument to the best explanation that posited $P$ would be like. Exactly what would it be a best explanation of? It seems very unlikely that $P$ will be posited as part of the explanation of the perceivable properties of the brain—its mass, colour, and so forth. We already have, or are well on the way to having, good explanations of those properties, and there seems to be no role in those explanations for property $P$. What about the explanation of behaviour? Is there a role for property $P$ there? Again it

39 In the case of perception, we considered the use of perceptual aids such as microscopes. Interestingly, no introspective analogue of the microscope seems available, although it was once fashionable to claim for certain psychoactive drugs, and for hypnosis and meditation, just that function.
seems unlikely. Qualia seem to be entirely irrelevant – epiphenomenal if you like – to the explanation of behaviour. After all, I have argued that a person could have a state within her which occupied, say, the pain role, and yet not experience the qualia usually associated with that role. Thus it seems that we can explain pain behaviour (by reference to the pain role's being occupied) without making any reference to qualia. But if we need make no reference to qualia in order to explain behaviour, then we need make no reference to the property which explains the existence of qualia in order to explain behaviour. It follows that no argument to the best explanation of behaviour will posit property \( P \).

It seems that property \( P \) will be part of an argument to the best explanation of only one thing: the qualia themselves. We posited \( P \) in order to account (in physical terms) for the existence of qualia, but 'property \( P \)' is merely functioning as a placeholder, the details of which we will never be able to provide. The immaterialist argument is, however, blocked, for that argument assumed that our inability to account for qualia in physical terms entailed the immateriality of qualia. But we can now see that an alternative explanation of our inability to account for qualia in physical terms is available. We thus have two options: we can either posit non-physical properties or accept that our minds are cognitively closed on this crucial issue. As stated above, parsimony considerations strongly favour the latter.

There is, however, a serious problem with the approach to qualia I have been sketching over the last several pages. Recall that our aim is to defend commonsense functionalism. One of the great virtues of commonsense functionalism is that it explains the (external) folk psychological intuition that mental states can be multiply realized. In humans the functional role \( R \) characteristic of mental state \( M \) is occupied by neurological state \( N \); in Martians (whose "brains" are physically quite different from ours) \( R \) is occupied by state \( S \) of silicon chip 924 alpha. From the physical perspective, \( N \) and \( S \) are quite distinct – \( N \) is a state of grey matter, \( S \) a state of a silicon chip – but functionalism abstracts away from the physical details of the implementation level. Property \( P \), though, is a physical property: it is that property of the brain which, if grasped, would figure significantly in the explanation of the existence of qualia. But since the occupant of \( R \) in Martians is physically quite distinct from the occupant of \( R \) in humans, it is highly unlikely that property \( P \) is instantiated by the Martian "brain". We could postulate a further property, \( P_M \), which, if grasped, would play a significant role in the explanation of the existence of Martian qualia, but that seems hopelessly ad hoc. It would be nothing short of a miracle if a \( P \)-style property conveniently turned up in a variety of quite distinct sorts of matter.

There are (at least) two ways of dealing with this problem.

---

**First way.** Rather than postulate a physical property $P$ which is cognitively closed to us, we could postulate a functional property $F$ which is cognitively closed to us and which, if grasped, would play a significant role in the explanation of the existence of qualia. Since functional properties can be multiply realized, there is no reason to suppose that $F$ cannot be realized by the Martian.

Tempting as this line is, I do not want to endorse it. It is crucial to the entire enterprise that $F$ be cognitively closed to us. But what sort of an argument can be made to the effect that $F$ is cognitively closed to us? How could a functional property be so abstract, so recondite, that we could not grasp it? The concept of $F$ is a concept of something which occupies a certain role in a rich causal network. It is not at all obvious how such a property could be closed to us. Indeed, it seems that such properties are amongst the properties which we can most readily grasp. Seeing things in causal terms, locating things in causal networks: these are mental processes we seem to be particularly good at.

**Second way.** We can bite the bullet and say that since Martians do not instantiate property $P$, they do not have qualia. Martians are unfeeling robots. Certainly they are very smart but, unlike us, they do not have feelings. Only humans have feelings; the having of feelings is a characteristically human thing.\(^{41}\) I think that this is the right response, but it has a consequence which I suspect some will find unpalatable. Martians can of course have states within them that occupy the pain role. The functionalist is therefore committed to saying that they can be in pain. But if Martians do not have qualia, it follows that pain and the feeling of pain have come completely apart in the Martian case. Martian pain does not hurt.

I argued above that we can feel pain but not be in pain, or be in pain but not feel it. Paul Churchland’s examples push us towards that conclusion, and Sellars’ myth about our Rylean ancestors explains just how it could be that our qualia reports are fallible. I stand by those arguments, and I take the Martians to be merely an extreme example of a case already established.

I have argued that we should take our inability to give a physicalist account of phenomenological properties as a reflection of our cognitive limitations rather than as evidence for the immateriality of qualia. If our cognitive powers were somehow miraculously extended, we would see that the neurological states identified by commonsense functionalism as the occupants, in us, of the causal roles characteristic of mental states, possess property $P$, and we would be able to explain the existence of qualia in terms of property $P$. Much of the force of the qualia objection to commonsense functionalism is turned aside by these considerations, but not all of it:

\(^{41}\) Property $P$ may be a property of mammalian brains, or of any brain made out of the same stuff as mammalian brains. In that case the having of feelings may not be as exclusively human as I have made out.
Frank Jackson's knowledge argument still presents the functionalist with certain difficulties. Jackson invites us to consider Mary:

Mary is confined to a black-and-white room, is educated through black and white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physics about us and our environment, in a wide sense of 'physical' which includes completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all of this, including of course functional roles. If physicalism is true, she knows all there is to know. For to suppose otherwise is to suppose that there is more to know than every physical fact, and that is just what physicalism denies . . . . It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. This is rightly described as learning – she will not say "ho, hum". Hence, physicalism is false.42

Note that the argument from cognitive closure given above takes some of the sting from Jackson's example. On the cognitive closure view, if Mary really did know all the physical facts then the existence of qualia would be no mystery to her. She would know all about property P and thus would be able to explain the existence of qualia in terms of property P. Of course, by the cognitive closure argument, that Mary knows all the physical facts entails that Mary is radically different from us, and that fact gives Jackson's example considerable rhetorical power. Property P is cognitively closed to us, and thus when we imagine ourselves in Mary's position we do not imagine ourselves becoming familiar with a property that would explain how qualia arise in the brain. Nevertheless, there is no good reason to doubt either the existence of such a property nor that Mary, having grasped that property, would be capable of explaining the existence of qualia in terms of it.

A problem, however, remains. For although imprisoned Mary can explain how phenomenological properties are related to physical properties, it still seems that she acquires something when she escapes from her monochromatic chamber.43 Prior


43 Colin McGinn would, it seems, deny that Mary presents us with a further problem. Thus he argues that if we knew the property P1 which is prominent in the explanation of how the bat's brain
to her leaving the black and white room her knowledge seemed too theoretical. It seemed that something practical was absent, and that that absence could only be remedied by exposure to the coloured world. I think that the hunch that Mary lacked something practical is exactly right. What she lacked when she was confined to the black and white room was not knowledge. Rather, what she lacked was a particularly useful way of representing and classifying various parts of the world—a way available to normally-sighted people. After leaving the room she acquired the ability to quickly classify objects whose surfaces have a disposition to cause certain effects in normally-sighted people. Prior to leaving the room she could have performed such classifications, but only much more laboriously. David Lewis has provided a nice analogy:

Imagine a smart database. It can be told things, it can store the information it is given, it can reason with it, it can answer questions on the basis of its stored information. Now imagine a pattern recognition device that works as follows. When exposed to a pattern it makes a sort of template, which it then applies to patterns presented to it in the future. Now imagine one device with both faculties, rather like a clock radio. There is no reason to think that any such device must have a third faculty: a faculty for making templates for patterns it has never been exposed to, using its stored information about these patterns. It has a full description about a pattern but no template for it, it lacks an ability but it doesn't lack information. (Rather, it lacks information in a usable form.) When it is shown the pattern it makes a template and gains abilities, but it gains no information. We might be rather like that.

It is time to bring my defense of functionalism to a close. If I have not entirely overturned the objections to functionalism put forward by Block, Searle, and others, I hope I have at least made clear that various countermoves are available to the commonsense functionalist. In the next Chapter the focus shifts from the nature of folk psychology (external) to its status. Even if, as I have argued, commonsense functionalism provides an adequate meta-theory for folk supports the bat's consciousness, we would know what it was like to be a bat: 'By grasping $P_1$ it would be perfectly intelligible to us how the bat's brain generates $B$-experiences [that is, the experiences of being a bat] . . . . But then it seems to follow that grasp of the theory that explains $B$-experiences would confer a grasp on the nature of those experiences . . . . How could we grasp the nature of $B$-experiences without grasping the character of those experiences?' McGinn 1991: 9. I claim below that Mary would still lack something even if she did grasp the explanatory property $P_1$: she would lack a skill.

44 This argument is drawn from Nemirow 1980 and Lewis 1983c.
45 Lewis 1983c.
psychology (external), it remains to be seen whether the roles picked out by commonsense functionalism are occupied at the actual world.
CHAPTER SIX

ELIMINATIVISM

6.1 Introduction. Let me briefly survey the progress so far. In chapter 1 I distinguished between two quite separate senses of the 'theory theory'. On the one hand, the theory theory (internal) claims that our ability to predict and explain our own and others' behaviour is underpinned by a largely tacit, internally represented, theory of human behaviour. I called that theory 'folk psychology (internal)'. On the other hand, the theory theory (external) claims that our everyday talk about mental states implicitly constitutes a theory of human behaviour. I called that theory 'folk psychology (external)'. Chapter 2 was devoted to defending the theory theory (internal) against its only rival – off-line simulation theory. In Chapter 3 I argued that commonsense functionalism provides the correct meta-theory for folk psychology (external); that is, the mental states posited by folk psychology (external) are the occupants of certain characteristic causal roles. Chapter 4 provided an account of how mental states thus construed could have intentional properties, and finally Chapter 5 attempted to deal with a variety of objections that have been offered against functionalism.

Folk psychology (external) is an empirical theory of human psychology and therefore might be false. If we accept that the posits of a false theory do not exist, then the falsity of folk psychology (external) would entail that there are no such things as beliefs and desires. This startling doctrine is known as eliminativism. A variety of arguments for and against eliminativism have been offered. I survey those arguments in Sections 6.2 and 6.3, respectively. I construe the pro- and anti-eliminativist arguments discussed in Sections 6.2 and 6.3 as concerned with the status of folk psychology (external). In Section 6.4 the focus shifts to folk psychology (internal). Before getting to those issues, though, it is worth forestalling four misguided discussions of objections to the eliminativist enterprise.

First objection. Eliminativism is, roughly, the doctrine that there are no such things as beliefs. That is, the eliminativist believes that there are no such things as beliefs. But that is a contradiction, so eliminativism is false.

The error here is clear. The argument is being begged against the eliminativist. If she is being careful she will deny both that she believes there are no beliefs and that she believes that there are beliefs. But of course most of the time she will not be careful – she will fall into the idioms of folk psychology (external). And that is fine as
long as we recall that, for the eliminativist, it is just a way of talking and conveys no ontological commitment to the posits of folk psychology (external). It may even turn out that the temptation to use the language of folk psychology (external) is pretty much irresistible. It seems plausible that our everyday talk about mental states is largely the way it is because folk psychology (internal) is the way it is. Moreover, some of our mental faculties are not under our voluntary control. I cannot resist, for example, assigning a meaning to the list of graphemes CHAIR, nor can I choose not to see, say, an apple as a three dimensional object. If the module which houses folk psychology (internal) is similarly involuntary and is, furthermore, hardwired in the sense in which grammar is reputedly hardwired, then we may find ourselves deploying the language of folk psychology (external) against our better judgment.

Second objection. Beliefs are given to us incorrigibly by our faculty of introspection. Therefore eliminativism must be false.

In Section 5.4 I introduced and endorsed Sellars' argument against the incorrigibility of introspection. At that time the focus was on qualia, but the point carries over against the claim that we can incorrigibly introspect beliefs. One way to view the incorrigibility objection is as a denial of the claim that beliefs and desires are the posits of a theory – albeit a folk one. Eliminativism takes that claim as a premiss, and then argues that the theory is false and thus its theoretical terms empty. Perhaps an anti-eliminativist position could be established by denying Sellars' claim, but it would have to involve argument rather than bare assertion.

Third objection. Folk psychology (external) is not in the business of offering causal explanations of behaviour. Rather, the language of folk psychology (external) facilitates the making of promises, the issuing of warnings, the offering of congratulations, and myriad other interpersonal transactions. It carries no ontological commitment to entities with causal relations and thus eliminativism has missed the point.¹

I think it should be conceded immediately that the language of folk psychology (external) plays a very important role in, as it were, lubricating the mechanisms of social life. But the argument just sketched requires more than that; it requires that folk psychology (external) do nothing but lubricate the mechanisms of social life in the way described. The exclusive claim is implausible because it entails that folk psychology (external) has no predictive or explanatory role whatsoever, and it is totally implausible that folk psychology (external) has no such role. We construct such predictions and explanations on a daily basis.

Fourth objection. Jerry Fodor sometimes seems to tacitly accept, and Lynne Rudder Baker considers, an argument to the effect that, if folk psychology (external)

plays no part in the serious scientific investigation of the mind, then it is false and eliminativism vindicated.\(^2\) The argument can be set out as follows.

\begin{enumerate}
\item The only properties which exist are the properties quantified over by respectable scientific theories.
\item (External) folk psychological properties are not quantified over by respectable scientific theories, nor will they be quantified over by respectable scientific theories in the future.
\end{enumerate}

Therefore

\begin{enumerate}
\item (External) folk psychological properties do not exist; that is, eliminativism is true.
\end{enumerate}

In Chapter 8 I will deny premiss (2), arguing that we have as yet no good reasons to suspect that folk psychology (external) will not form the basis of a respectable scientific psychology, and indeed that we have some good reason to think that it might. However, even setting aside any qualms about premiss (2), the argument just sketched is unconvincing because premiss (1) is implausible. Why should we accept that the only properties which exist are the properties quantified over by respectable science? To borrow an example from Robert Cummins and Georg Schwartz, the property \textit{is a warm coat} is not quantified over by any respectable scientific theory.\(^3\) Nevertheless, warm coat eliminativism is an insane doctrine. I insist that warm coats exist! Moreover, objects that instantiate the property \textit{is a warm coat} can be involved in causal explanations in virtue of possessing that property. Thus, that X is a warm coat can be part of the causal explanation of Y's surviving the blizzard.\(^4\) Similarly, even if it turned out to be the case that (external) folk psychological properties are not quantified over by respectable scientific theories, it would not follow that they do not exist nor that they are not involved in causal explanations.

What is the relationship between properties like \textit{is a warm coat} and the properties quantified over by respectable science? Most everyday properties like \textit{is a warm coat} do not reduce to the properties of, say, physics. Rather, they \textit{supervene} upon them; that is, if we fix all the physical properties we fix all the "clothing"

\(^2\) Passages hinting at this view can be found in Fodor 1981 and Fodor 1987. Terence Horgan attributes this position to Fodor, although he admits that it is not explicit in Fodor's work. See Horgan 1992. See also Baker 1987, especially Ch. 1.


properties. Similarly, (external) folk psychological properties such as \textit{wants a cappuccino} supervene on the physical properties: fix the physical properties and you fix the (external) folk psychological ones.\footnote{For an account of explanation in terms of supervenient properties see Jackson & Pettit 1990b.}

I turn now to a series of considerably more powerful arguments.

### 6.2 Arguing For Eliminativism

Eliminativist arguments are typically variations on the following general strategy.

1. Beliefs, desires, and so forth are posits of a commonsense theory of mind, namely, folk psychology (external).

2. Folk psychology (external) is false.

Therefore

3. Beliefs, desires, and so forth do not exist.

It is not immediately clear that the argument from (1) to (3) is valid. Sometimes when we discover that a theory is false we eliminate its posits - the phlogiston theory of combustion is an obvious case. When the phlogiston theory was overthrown by the oxidation theory, it was accepted that there was no such thing as phlogiston. However, we do not always eliminate the entities posited by a false theory: sometimes we correct the theory and take it that we have made a discovery about the true nature of its posits. Thus for example modern atomic theory has overturned 19th Century theories of the atom. Nevertheless, atoms have not been eliminated; rather, our view of atoms has been corrected.

When does the failure of a theory entail the elimination of its posits? This issue is closely tied to the issue of reference. William Lycan has pointed out that if we adopt the causal-historical theory of reference for natural kind terms then the falsity of a theory does not entail the elimination of its posits.\footnote{See Lycan 1988: 31-32. For the causal-historical theory of reference see Donellan 1972; Kripke 1980; Putnam 1975. See also Devitt & Sterelny 1987.} According to the causal-historical theory, a natural kind term 'K' refers to the natural kind K because there exists a causal-historical chain linking present usage of 'K' with an original naming event involving an example of the kind K. Roughly, 'K' refers to everything that belongs to the same kind as the original example. What is crucial for our purposes is that, according to the causal-historical theory, speakers can succeed in referring to K's in utter ignorance of the nature of K: all that is required is that they are appropriately
causally-historically linked to the sample of K present at the original naming. Thus, speakers can refer to K even if their theory of K is completely false. The failure of our theory of K is no ground for denying the existence of K's. In particular, the failure of folk psychology (external) is no ground for denying the existence of beliefs and desires. Elimination has become impossible. 7

Now consider the classical Russellian theory of reference according to which a natural kind term 'K' succeeds in referring to the kind K because it stands for, or abbreviates, a definite description satisfied by K. 8 On this account of reference, elimination is almost inevitable. For it is very likely that we will come to quibble with some of the details of the definite description, but that will entail that there are no such things as K's.

We do not have to look far to locate the remedy to this problem. The cluster theory of reference proposed by Strawson and Searle allows for both the correction of mistaken views about theoretical posits and for their outright elimination. 9 On this view, a natural kind term 'K' succeeds in referring to a natural kind K because it stands for a disjunction of definite descriptions, where each disjunct varies to some extent from the others. Thus we can discover that our theory of K is mistaken—that one or more of the disjoined definite descriptions is unsatisfied at the actual world—and still not be forced into elimination. The remaining disjuncts carry, as it were, the burden of reference. Alternatively, it could turn out that all of the disjuncts are unsatisfied. In that case, the term fails to refer and elimination occurs.

I said that the disjoined definite descriptions vary from one another to some extent. If we are convinced that all the K's necessarily have some property P—if, that is, we hold essentialist views about K—then each conjunct will existentially quantify over P. Then, should we discover that nothing has P essentially, elimination will follow. Cases of this nature do exist, although sometimes properties which we might once have taken to be essential turn out not to be. Thus, we might once have been tempted to say that it is analytic that atoms are indivisible, but that proved not to be the case. The discovery of the electron revealed something about the nature of atoms; it did not entail their elimination.

In Chapter 3 I endorsed Lewis' Ramsey sentence approach to theoretical terms. As it is standardly presented, that approach is indistinguishable from the Russellian definite description approach. However, if we form not a conjunction of the folk platitudes about the mind but a disjunction of conjunctions of most of the

7 Note that the argument just given from the causal-historical theory of reference to the impossibility of elimination assumes, as Lycan does, that the causal-historical chain can terminate in a multiply realized functional kind rather than in a natural kind narrowly construed.
8 For the classical Russellian theory of reference see Devitt & Sterelny 1987; Russell 1957.
9 See Searle 1958, 1969: 162-174; Strawson 1959: 180-183 and 190-194. See also Devitt & Sterelny 1987, especially Ch. 3. I introduced the cluster theory of reference in Section 3.4, below.

123
platitudes, Lewis' procedure yields a cluster theoretic approach to the reference of (external) folk psychological expressions. One of my reasons for endorsing Lewis' approach to theoretical terms should now be apparent: it alone is capable of accounting for our ontological practices. As we have seen, the causal-historical theory of reference cannot account for the fact that we do, at times, eliminate theoretical posits, whilst the classical Russellian theory cannot account for the fact that at times we retain theoretical posits even though much of what we had believed about them turned out to be false.10

Intriguingly, whilst the cluster theory of reference is presently somewhat unfashionable, it is endorsed by significant writers on both sides of the eliminativism debate. We have already noted Lewis' commitment to the cluster theory. Less often remarked upon is Churchland's endorsement of a Lewis-style position on theoretical terms. Thus he writes that

the semantics of the terms of our familiar mentalistic vocabulary is to be understood in the same manner as the semantics of theoretical terms generally: the meaning of any theoretical term is fixed by the network of laws in which it figures.11

Let us return to the eliminativist strategy sketched above and examine the ways in which eliminativists have attempted to flesh out premise (2). In what follows I will focus on Paul Churchland's 'Eliminative Materialism and the Propositional Attitudes', and on William Ramsey, Stephen Stich, and Joseph Garon's 'Connectionism, Eliminativism and the Future of Folk Psychology'.

Churchland offers us three considerations which purport to show that folk psychology (external) is false. First, he claims that folk psychology (external) has extensive and serious explanatory failures.12 Second, he argues that the history of folk psychology (external) reveals it to be a degenerating research program.13 And finally he claims that folk psychology (external) does not fit comfortably within the corpus of modern science.14

I can do no better than quote Churchland's own account of folk psychology (external)’s shortcomings.

---

13 See P.M. Churchland 1981: 74-75.
14 See P.M. Churchland 1981: 75.
As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of FP [folk psychology (external)], consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals. Consider our utter ignorance of the nature and psychological function of sleep, that curious state in which one third of one's life is spent. Reflect on the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball. Consider the internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas. Consider the rich variety of perceptual illusions, visual and otherwise. Or consider the miracle of memory, with its lightning capacity for relevant retrieval. On this and many other mental phenomena, FP sheds negligible light.

One particularly outstanding mystery is the nature of the learning process itself, especially where it involves large-scale conceptual change, and especially as it appears in its pre-linguistic or entirely nonlinguistic form (as in infants and animals), which is by far the most common form in nature.¹⁵

I think that we should concede at once that folk psychology (external) is much as Churchland describes. Nevertheless, that folk psychology (external) has certain limits does not show that it is largely false. Folk psychology (external) has nothing at all to say about sleep or the grounds of intelligence or perceptual illusions. But so what? It is not a theory of sleep or intelligence or perceptual illusion. It is, rather, a theory of a certain class of human behaviour. The class of behaviour for which it seeks to account can only be identified by reference to folk psychology (external) itself: folk psychology (external) seeks to account for precisely those behaviours it seeks to account for. Other theories are similarly involved in the demarcation of their subject matter. What is the domain of biology? The biological phenomena. Which phenomena are the biological phenomena? Those that biology seeks to account for. Sometimes we have a prior conception of the domain which a theory is supposed to cover – presumably this was the case with biology. But inevitably our conception of the domain changes as the science of it progresses. Thus the growth of chemical crystals was once considered a biological phenomenon but is no longer so regarded; and subterranean oil reserves might once have struck us as falling outside the domain of biology and yet we now know that biological properties are central to explaining

¹⁵ P.M. Churchland 1988: 73.
their existence. Folk psychology (external) seeks to explain, then, precisely what it seeks to explain, namely, a certain class of behaviour, and that it fails to account for a wide range of phenomena that might loosely be called 'mental' does nothing to establish its falsity.

Moreover, the possibility exists that folk psychology (external) might somehow be extended to cover some of the phenomena Churchland lists. This observation relates directly to Churchland's second objection: that folk psychology (external) is a degenerate research project. Churchland claims that the history of folk psychology (external) is a 'story of retreat, infertility, and decadence' and that 'both the content and success of FP have not advanced sensibly in two or three thousand years'. It is not clear how failure to change establishes the falsity of a theory. Indeed, it might even seem that the opposite is the case: the truth is immutable and so a history of modifications suggests that at least most of the time the theory was false! Moreover, it is not clear that folk psychology (external) is a degenerating research program. A great deal of scientific psychology takes folk psychology (external) as its starting point. To take just one example, there is a vast and fascinating literature on the psychology of the emotions. Much of this literature individuates emotions very much as folk psychology (external) does, and finds it theoretically rewarding to do so. So it seems to be false both that the history of folk psychology (external) is that of a stagnant and degenerate research program and that folk psychology (external) might not be successfully extended into the domains Churchland accuses it of being silent about.

Finally, Churchland claims that folk psychology (external)'s 'intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus [of scientific discoveries]'. That folk psychology (external) does not reduce to some other theory (say neuroscience) is not at all surprising. In previous Chapters I have argued extensively for the view that folk psychology (external) is best understood as a functionalist theory: (external) folk psychological mental states are identified with the occupants of certain causal roles. It follows that (external) folk psychological mental states can be multiply realizable. Indeed, one of the virtues of commonsense functionalism is that it accounts for the (external) folk psychological intuition that a wide variety of different kinds of systems could have minds. But a consequence of the multiple realizability of (external) folk psychological mental states is that folk psychology (external) will not reduce to neurobiology; that is, it will not be

---

17 I take up this issue in detail in Ch. 8.
19 P.M. Churchland 1988: 75.
reduced by the theory of what is merely one amongst many possible realizations. What we can hope for is that, in our case at least, (external) folk psychological mental states *supervene* on neurological states.

Is there a problem with the idea that (external) folk psychological states supervene on neurological states? Does the purported supervenience relation entail, imply, or in some weaker sense suggest, that the posits of folk psychology (external) are ripe for elimination? I see no reason to think so. Economic properties supervene on a variety of physical and psychological properties, but that does not entail, imply, or in some weaker sense suggest eliminativism about the posits of economic theory. Similarly, the properties posited by Mendelian genetics supervene on properties posited by molecular biology, but again nothing disastrous seems to follow.

Churchland could respond by attempting to debunk functionalism, and in fact he attempts to do so. His anti-functionalist argument, though, is singularly weak. He points out that alchemy could have been defended against elimination by modern chemistry if only the alchemists had been cunning enough to recast their theory in functionalist guise. Churchland is quite right that alchemy could have been recast as a functionalist theory. Thus the alchemic property 'ensouled by mercury' could have been identified with 'the disposition to reflect light, to liquefy under heat, to unite with other matter in the same state, and so forth'. It is no doubt true that *something* occupies what we might call the 'mercury role'. But it is not clear that recasting alchemy as a functionalist thesis would have saved alchemy. For presumably the alchemists believed that ensouled mercury was essentially a natural kind. That is, every definite description in the cluster of definite descriptions which secured the meaning of 'ensouled mercury' included the claim that ensouled mercury is a natural kind. But the occupant of the mercury role will not be a natural kind – far from it. It will, rather, be a strange concatenation of natural kinds. So functionalizing alchemy will not save alchemy; rather, it will transform it into something the alchemist would not have recognized. Notice also that (external) folk psychology has not been transformed into a functionalist theory in some *ad hoc* manner. It always was a functionalist theory and thus the analogy with a "saved" alchemy breaks down.

Moreover, even if it were true that alchemy could have been "saved" by treating it as a functionalist thesis, that does nothing to establish that all functionalist theories are in some way bogus. If Churchland's argument worked it would sweep away all functionalist properties, including perfectly respectable functionalist properties such as *is a carburetor*. What we need is a specific argument against

---

22 X is a carburetor if and only if X occupies the carburetor role; that is, if and only if X takes as inputs fuel, air, and a throttle setting, and releases as outputs air/fuel mixtures of a ratio specified by the throttle setting.
functionalism as a meta-theory of folk psychology (external), and none is forthcoming.

Let us rewrite Churchland's third eliminativist argument without the problematic reference to reduction. Churchland's concern is that folk psychology (external) will not mesh smoothly with the rest of science. No argument is given to support the move from a theory's failure to mesh with the rest of science to its falsity, and it is not clear how such an argument would go: perhaps psychology does indeed stand in glorious isolation from the rest of science. Nevertheless, it must be conceded that it would be a source of disquiet if folk psychology (external) proved permanently recalcitrant to integration with the rest of science. There seems, though, little reason to fear that folk psychology (external) will prove to be thus isolated. I have already mentioned that much scientific psychology can be seen as a natural extension of folk psychology (external), and in fact much of economics takes (external) folk psychological mental states pretty much for granted. This suggests that at least partial integration with certain sciences has already occurred. Moreover, commonsense functionalism offers at least the possibility of integrating folk psychology (external) with the neurosciences. After all, if the functional roles specified by folk psychology (external) are occupied in us at all, they are occupied by neurobiological states. So it seems that there is at least some hope that folk psychology (external) will mesh smoothly with neuroscience. What, though, if neuroscience demonstrated that there is nothing in the brain that could plausibly be identified with beliefs and desires? This, in a nutshell, is Ramsey, Stich and Garon's concern. Let us turn now to their paper.

Ramsey, Stich and Garon offer an argument with the following general structure.

(1) (External) folk psychological beliefs and desires have a certain property, $P$.

(2) If a certain model of cognition is true, then there is nothing in the brain that has property $P$.

Therefore

(3) If that model of cognition is true there are no such things as (external) folk psychological beliefs and desires.\(^{23}\)

\(^{23}\) Ramsey, Stich and Garon's argument is thus an elaborate attempt to establish premiss (2) of the general form of eliminativist argument given at the start of this Section.
Ramsey, Stich and Garon identify property $P$ as *propositional modularity*, and they identify the model of cognition cited in premiss (2) as *widely distributed connectionist* models. Let us take these in turn.

According to Ramsey, Stich and Garon, the propositional attitudes posited by folk psychology (external) exhibit propositional modularity; that is, they are *functionally discrete, semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behaviour.\textsuperscript{24} It is clear that (external) folk psychological propositional attitudes have semantic properties and that they are causally related to both behaviour and other mental states. But what about the claim that (external) folk psychological propositional attitudes are functionally discrete? Ramsey, Stich and Garon motivate their functional discreteness claim as follows.

[I]t typically makes good sense to claim that a person has acquired (or lost) a single memory or belief. Thus for example on a given occasion it might plausibly be claimed that when Henry woke up from his nap he had completely forgotten that the car keys were hidden in the refrigerator, though he had forgotten nothing else.\textsuperscript{25}

As we shall see, the claim that (external) folk psychological propositional attitudes are functionally discrete needs a good deal of tightening, but the general idea is clear: if (external) folk psychological propositional attitudes are functionally discrete then they can be added or subtracted – committed to memory or forgotten – one by one.

Ramsey, Stich and Garon focus on widely distributed connectionist models of cognition. A widely distributed connectionist network is a computational device which maps certain inputs onto certain outputs.\textsuperscript{26} It consists of a large number of simple units typically arranged in three layers and connected largely in parallel. The three layers are standardly referred to as the 'input', 'hidden', and 'output' layers. Unlike a conventional serial computer, there is no central executive in a widely distributed connectionist network. Rather, the output of each unit is entirely determined by information locally available to it. Widely distributed connectionist networks exhibit some fascinating properties. (1) They can learn new input-output profiles without explicit programming. Learning involves repeatedly presenting the network with inputs drawn from the relevant domain and adjusting the units in the

\textsuperscript{24} Ramsey et al 1990: 504. Their emphasis.
\textsuperscript{25} Ramsey et al 1990: 504-505.
\textsuperscript{26} For extensive discussions of connectionist networks see in particular Rumelhart & McClelland 1986.
hidden layer until the correct output is achieved. The details of the adjustment process need not concern us here. (2) They can often produce the correct output for an input to which they have never previously been exposed. That is, they can generalize from a limited number of examples. Moreover, they can often produce the correct output in response to a degraded input. (3) They often exhibit the property of 'graceful degradation'; that is, if some part of the network is damaged they tend to produce either correct or degraded output rather than no output at all.

In some connectionist networks, each unit is assigned a specific meaning, with each unit representing exactly one feature of the domain. However, in widely distributed connectionist networks, the network carries information about the relevant domain in the connection strengths between its units, and there is no simple one-one correlation between parts of the network and features of the domain. It is this characteristic which has lead to their being called widely distributed connectionist networks.

Ramsey, Stich and Garon take as their example a network which has learnt to respond (i.e., answer 'true' or 'false') to questions about a domain of sixteen propositions. Since the network can answer correctly, there is a clear sense in which it has formed an accurate representation of the domain. Nevertheless - and this is crucial for our purposes - the network's representation of the domain is widely distributed in the sense just given. No discrete part of the network can be identified as representing any one proposition; rather, the representation of a particular proposition supervenes on the connection strengths of the entire system (or at least of extended parts of the entire system).

I think I have already said enough to make Ramsey, Stich and Garon's eliminativist worry apparent. For they have argued that (external) folk psychological propositional attitudes exhibit propositional modularity, in particular, that they are functionally discrete. But the representational states within widely distributed connectionist networks are not functionally discrete; quite the opposite - they are highly indiscrete. Thus, if it turns out that widely distributed connectionist networks are a good model of human cognition then there are no such things as (external) folk psychological propositional attitudes.

It is important to emphasize that Ramsey, Stich and Garon have offered a conditional: If a certain class of connectionist networks satisfactorily model human cognition then there are no such things as beliefs and desires. Too little is as yet known about connectionist networks and human cognition, they claim, to detach the consequent. I do not, though, want to argue against Ramsey, Stich and Garon by

27 See for example McClelland, Rumelhart and Hinton's 'Jets and Sharks' example. McClelland et al 1986.
denying the antecedent of their conditional. Nor do I want to object to the general strategy of locating an empirical claim about human cognition or the brain and then arguing that, if that claim turns out to be true, folk psychology (external) is false. Folk psychology (external) is an empirical theory and thus may be falsified by empirical data. Rather, I want to argue that the particular conditional Ramsey, Stich and Garon have offered us is false. Indeed, I would go as far as claiming that, if widely distributed connectionist networks turn out to be good models of human cognition, so much the better for folk psychology (external).

I remarked above that Ramsey, Stich and Garon’s claim that (external) folk psychological propositional attitudes are functionally discrete could do with some tightening. We need to distinguish the claim that we can add or subtract (external) folk psychological beliefs and desires one at a time from the claim that we can add or subtract (external) folk psychological beliefs and desires without having any other effects on the system. The former is clearly true: I can acquire a single new belief or forget something I previously knew. But according to commonsense functionalism the latter claim is false. Recall that commonsense functionalism is committed to ontological holism: (external) folk psychological mental states necessarily have actual causal relations to other mental states. Thus when I acquire a new belief, I acquire a new node in an existent network of actual causal relations. It follows that when I acquire a new belief at least some of my pre-existing beliefs gain new causal relations and are to that extent changed.

Does Ramsey, Stich and Garon’s eliminativist argument go through on either interpretation of what it is for the propositional attitudes to be functionally discrete? I argued that (external) folk psychological propositional attitudes are functionally discrete in the sense that they can be added or subtracted one by one. But the representational states of widely distributed connectionist networks are also functionally discrete in that sense of ‘functionally discrete’. Given suitable retraining, a single representation can be added to, or subtracted from, a widely distributed connectionist network, just as I can remember or forget a single belief. What about the second sense of ‘functionally discrete’? As we have seen, I cannot gain or lose a belief without altering the actual causal relations of my other beliefs, so in that sense (external) folk psychological propositional attitudes are not functionally discrete. However, in the process of retraining a widely distributed network so that it gains or loses a representation, the actual causal relations of the other representational states will be altered to some extent. Thus widely distributed networks, too, are indiscrrete in the second sense of ‘functionally indiscrrete’. It follows that Ramsey, Stich and Garon’s argument does not go through. On the first sense of ‘functionally discrete’ both the network’s representational states and the propositional attitudes of folk psychology (external) are functionally discrete; on the second sense, neither the
network's representational states nor the propositional attitudes of folk psychology (external) are functionally discrete. The mismatch between widely distributed connectionist models of cognition and folk psychology (external) required by Ramsey, Stich and Garon's argument simply does not exist.

There is a further element to Ramsey, Stich and Garon's idea of functional discreteness. They claim that, according to folk psychology (external), it can be the case that one belief rather than another is uniquely involved in the production of a particular behaviour. Once more the eliminativist conclusion looms because no such unique causes of output can be located in the case of widely distributed connectionist networks. But again Ramsey, Stich and Garon have misrepresented folk psychology (external). Obviously, if (external) folk psychological states are loci in rich causal networks then it will not be the case that one belief rather than another is uniquely involved in the production of a particular behaviour. Many, many beliefs will causally bear on each one of an agent's actions.

Contrary to Ramsey, Stich and Garon, there is a remarkable match between the properties of (external) folk psychological propositional attitudes and the properties of the representational states of widely distributed connectionist networks. Thus, far from it being the case that widely distributed connectionist models of cognition are a threat to folk psychology (external), they are a blessing to it. They give a tantalizing glimpse of how (external) folk psychological mental states might be implemented in the brain.

6.3 Arguing Against Eliminativism. In this Section I want to consider a powerful anti-eliminativist argument advanced by Frank Jackson and Philip Pettit. Their argument has the following general structure.

(1) It is sufficient for having beliefs and desires that one be in states that satisfy the functional roles specified by folk psychology (external).

(2) We are sometimes in states which satisfy the functional roles specified by folk psychology (external).

Therefore

(3) We have beliefs and desires.\(^\text{29}\)

\(^{29}\) See Jackson & Pettit 1990a.

\(^{30}\) See Jackson and Pettit 1990a: 36.
Premiss (1) is, in effect, an endorsement of commonsense functionalism, and I urge that we accept it without further ado. The argument rests, then, on premiss 2. As Jackson and Pettit point out, if the argument sketched above is to carry weight against the eliminativist, premiss 2 must be 'peculiarly unlikely to be undermined by future progress in neuroscience'. How could premiss 2 be thus peculiarly resistant to revision? Clearly, the less premiss 2 ventures, the less likely it is to be overthrown. Thus, if Jackson and Pettit can demonstrate that in fact premiss 2 ventures very little, they will have provided a strong anti-eliminativist argument. Jackson and Pettit point out that premiss 2 claims only that the folk-specified roles are occupied; it makes no claims whatsoever about the occupants of the roles. If folk psychology (external) ventured an opinion on the nature of the occupants of the folk-specified roles, then it would be open to direct contradiction by neuroscience. But by restricting itself to a functionalist level of description, folk psychology (external) is saved from head-on confrontation with neuroscience. Thus they write that premiss (2)

is entirely in terms of the functional roles played by internal states. If the folk hypothesis is wrong, it is wrong by having the functional roles wrong. There is no other way for it to be wrong. But having the functional roles wrong is a very hard matter to conceal precisely because they are functional roles. The hypothesis that metals are good conductors of heat and electricity was established much earlier and much more securely than the explanation of the fact in terms of the atomic theory, and this was precisely because being a good conductor is functionally defined in terms of playing a certain role between observable inputs and outputs.

There is an obvious objection to be made here. The functionalist account of being a good conductor is spelt out entirely in terms of inputs and outputs, but the commonsense functionalist account of mental states appealed to by premiss (1) quantifies over internal states as well as over inputs and outputs. How can we be sure that neuroscience will not reveal that the internal states postulated by folk psychology (external) do not exist?

Jackson and Pettit are aware of this objection. They reply by admitting that folk psychology (external) is committed to there being internal causes of behaviour, but insist that the postulated internal causes are such that there is little room for doubt about their existence. They offer the following analogy:

---

32 Jackson and Pettit 1990a: 41-42.
Suppose that I punch 4 followed by 7 on my pocket calculator, and that 11 then appears on the screen. I discard the chance hypothesis and adopt one which includes the clause: my calculator stores for subsequent utilization the number \( n \) when \( n \) is punched. This is a functional hypothesis that goes beyond its observational base both in respect to internal causal links and in respect to hypothetical inputs and outputs. Nevertheless, after a relatively few trials, the hypothesis would be overwhelmingly credible, although there might still be much doubt about how the calculator does the storing.\(^{33}\)

How satisfactory is this response? Jackson and Pettit have pointed out that we can be confident that something inside the pocket calculator plays the store \( n \) role because it is absolutely incredible that the calculator could succeed in generating the correct input-output profile without storing \( n \). As long as the correct input-output role is generated we can be sure that the role is being occupied, and no investigation of the inner workings of the device is likely to prove otherwise. Similarly, as long as we generate the correct input-output profile – as long as we produce the correct behaviour in circumstances – we can be confident that the relevant folk-specified roles are occupied, and no neuroscientific investigations of our inner workings is likely to prove otherwise.

But there is a catch. As we saw when discussing Block's robot example in Section 5.2, folk psychology (external) is not interested simply in input-output profiles; it is, rather, interested in how those input-output profiles are mediated. In particular, it insists that they are mediated by a rich network of beliefs and desires. The calculator example is thus misleading. It is true that neuroscience could not discover that there is nothing in the brain which mediates our input-output profile; but neuroscience could discover that there is nothing in the brain which can plausibly be identified with the rich causal network of beliefs and desires specified by folk psychology (external).

The problem with Jackson and Pettit’s argument is that it defends the wrong kind of functionalism. It shows that input-output functionalism is peculiarly resistant to elimination; it does not show that commonsense functionalism is peculiarly resistant to elimination. Think again of the Blockhead. If we accepted input-output functionalism we would have to accept that the Blockhead has the same beliefs and desires as a human chess expert. Notice that the attribution of beliefs and desires to the robot by input-output functionalism would be peculiarly resistant to overthrow by examination of its internal structure. Since nothing we discovered about the robot’s

\(^{33}\) Jackson and Pettit 1990a: 42.
internal structure would change our opinions about its input-output profile, nothing we discovered about its internal structure would alter our attribution — qua input-output functionalists — of mental states to the robot.

However, if we accept not input-output functionalism but commonsense functionalism, the robot's internal structure is very relevant to our attributions of beliefs and desire. In Section 5.2 we concluded that the Blockhead had no beliefs and desires because it had the wrong internal structure, even though it had the right behaviour in circumstances, and even though its behaviour was caused by its own inner states rather than by, say, a distant agent. We too might turn out to have the wrong internal structure. Neuroscience may yet reveal that our input-output profile is mediated in a way incompatible with the ascription of (external) folk psychological mental states. In that case the eliminativist will have turned out to be right.

There is a second anti-eliminativist argument hinted at by Jackson and Pettit: the predictive success of folk psychology (external) suggests that it is true and therefore will not be subject to elimination. Intriguingly, the same argument is mentioned by Paul Churchland who remarks that 'the average person is able to explain, and even predict, the behavior of other persons with a facility and success that is remarkable'. Folk psychology (external) does enjoy considerable predictive success, particularly when contrasted with alternative paradigms such as neuroscience and connectionism. I take these sort of considerations to be the best anti-eliminativist argument available. The predictive success of a theory provides considerable — although defeasible — evidence of its veracity.

Nevertheless, we can ask just how successful folk psychology (external) is. At present we simply do not know — almost none of the relevant empirical research has been done. Note that determining the predictive power of folk psychology (external) requires first determining the domain of that theory. As we saw when discussing Paul Churchland's claim that folk psychology (external) is silent about many aspects of what might broadly be called 'mental life', folk psychology (external) is concerned with predicting and explaining a certain class of behaviour. Folk psychology (external) cannot reasonably be judged to be a predictive failure just because it does not yield accurate predictions outside that domain. However, characterizing that domain in a way that allows us to explore folk psychology (external)'s predictive success will not prove easy.

34 See Jackson and Pettit 1990a: 41.
35 P.M. Churchland 1981: 68.
36 In Section 2.3 we saw an example drawn from Nisbett & Ross 1980 in which folk psychology (external) failed to predict behaviour influenced by the so-called position effect.
6.4 Eliminativism And Folk Psychology (Internal). The discussion during the last three Sections has focussed exclusively on folk psychology (external). But what about folk psychology (internal)? Are its posits candidates for elimination? Very little is presently known about folk psychology (internal), and thus it is extremely difficult to pass any reasonable judgement about the likelihood of its theoretical posits being eliminated. Nevertheless, some quite surprising conclusions can be reached by reflecting on the ways in which folk psychology (internal) might be represented in the brain.

Folk psychology (internal) is a largely tacit, internally represented, theory of human psychology involved in the prediction of our own and others' behaviour. (In Section 1.3 we saw an analogy with folk mechanics, which is a largely tacit, internally represented, theory of the movement of middle-sized physical objects.) How is folk psychology (internal) represented? There are at least five possibilities.

1. As a list of declarative propositions.
2. As a list of declarative propositions plus a list of (non-declarative) instructions for manipulating the list of declarative propositions.
3. Just as a list of (non-declarative) instructions.
4. By a connectionist network that maps smoothly onto a list of declarative propositions.
5. By a connectionist network that fails to map smoothly onto any list of declarative propositions.

Let us consider proposals (1) to (5) in turn, and examine the consequences of each one for the eliminativist's claim.

(1) Folk psychology (internal) is represented in the brain as a list of declarative propositions. This is perhaps the most intuitive of the five proposals. The idea is that folk psychology (internal) is represented in the brain as a list of law-like statements. Thus, folk psychology (internal) might be represented as a list of propositions of the form:

If X is in mental state M1 and experiences sensory input S1 then X will move into mental state M2 and output behaviour B1.

Clearly, if this is the way folk psychology (internal) is represented in the brain, then the eliminativist has a target. The law-like propositions could turn out to be radically false, thus raising the possibility that the posits of folk psychology (internal) do not exist. We have here a direct analogy with eliminativism about the posits of folk psychology (external).

(2) Folk psychology (internal) is represented in the brain as a list of declarative propositions plus a list of (non-declarative) instructions for manipulating the list of declarative propositions. On this view folk psychology (internal) consists of law-like

---

37 This Section is based in part on Stich & Ravenscroft 1994.
propositions of the sort described above, plus a set of rules or instructions for applying the law-like propositions to particular cases. If this is how folk psychology (external) is represented, the eliminativist still has a target, for again it could turn out that the law-like propositions are largely false.

(3) Folk psychology (internal) is represented in the brain as a list of (non-declarative) instructions. On this view, the representation of folk psychology (internal) includes no declarative propositions; rather, it consists only of rules for forming predictions and explanations in certain circumstances. If folk psychology (internal) is indeed represented in the form of (non-declarative) instructions, then the eliminativist may have no target. Lists of instructions make no claims about the world and thus cannot be false – although they may be useful or otherwise.

It might be objected that the list of instructions presupposes that certain propositions are true of the domain, and that those presuppositions can be true or false. The notion of presupposition is, however, notoriously difficult to unpack. Moreover, it is not clear that there will exist a univocal set of propositions that can be identified with what is presupposed by folk psychology (internal). If there is not, the eliminativist will have no clear target.

(4) Folk psychology (internal) is represented in the brain by a connectionist network that maps smoothly onto a list of declarative propositions. If folk psychology (internal) is thus represented then the eliminativist will have a determinate target: the set of declarative propositions represented by the connectionist network. Those declarative propositions will be true or false. If they are radically false, eliminativism about the posits of folk psychology (internal) will have been vindicated.

(5) Folk psychology (internal) is represented in the brain by a connectionist network which fails to map smoothly onto any determinate list of declarative propositions. If folk psychology (internal) is indeed represented in the brain by a connectionist network that fails to map smoothly onto any determinate list of declarative propositions, then the eliminativist will have no definite target. There will be no determinate set of declarative sentences which could turn out to be largely false, and thus there will be no clear sense in which eliminativism about the posits of folk psychology (internal) could be true.

Whether eliminativism about the posits of folk psychology (internal) is an intelligible doctrine thus turns out to be very dependent on certain empirical facts about representation in the brain. If it turns out that folk psychology (internal) is represented in the brain solely as a list of rules or instructions, or if it turns out that it is represented by a connectionist network that does not map smoothly onto a determinate set of declarative propositions, then eliminativism about the posits of folk psychology (internal) may not even be a coherent doctrine.
Before closing this Chapter I want to deal briefly with a remarkable anti-eliminativist argument put forward by off-line simulation theorists Robert Gordon and Alvin Goldman.38 Recall the discussion of off-line simulation theory in Chapter 2, above. According to off-line simulation theory, we predict and explain human behaviour by taking our own decision making apparatus off-line and feeding into it "pretend" beliefs and desires. Gordon and Goldman argue that if off-line simulation theory is true then the eliminativism debate is entirely misguided. How so? Well, if off-line simulation theory is true then we do not predict and explain human behaviour by accessing an internally represented folk theory of the mind. Indeed, if off-line simulation theory is true it would be entirely otiose to posit an internally represented folk theory of the mind. But if there is no such theory then there cannot be a debate about the existence or otherwise of its posits. It follows that eliminativism is a singularly benighted doctrine: it worries about the existence of the posits of a nonexistent theory!

The distinction between folk psychology (internal) and folk psychology (external) allows us to see the mistake in Gordon and Goldman's argument. It is the case that if off-line simulation theory is true then there is not such thing as folk psychology (internal). But the truth or falsity of off-line simulation theory leaves untouched the issue of whether folk psychology (external) exists or not. And if it does exist then the eliminativist will indeed have a legitimate target: the posits of folk psychology (external).

CHAPTER SEVEN

NATURALISM AND IRREALISM

7.1 Introduction. In the previous Chapter we examined a variety of arguments both for and against eliminativism. Those arguments were, by and large, found wanting. There exists, though, an eliminativist concern not addressed in the last Chapter. This concern turns on fears that the intentional properties will resist naturalization. Non-natural properties, the argument goes, are not instantiated at the actual world. It follows that, if the intentional properties cannot be naturalized, they are not actual.\(^1\) Fodor makes the point thus:

> the deepest motivation for intentional irrealism derives . . . . from a certain ontological intuition: that there is no place for intentional categories in a physicalistic view of the world; that the intentional can't be naturalized.\(^2\)

Stich and Laurence have recently surveyed existing accounts of naturalism and argued that, on every available account, either the intentional properties turn out be naturalistic or, if they do not, intentional irrealism does not follow.\(^3\) If Stich and Laurence are right, Fodor and those attracted to his position have been unduly alarmed. But, as Stich and Laurence themselves point out, their argument is convincing only to the extent that their survey of accounts of naturalism is complete.\(^4\) In the next Section I sketch an account of naturalism not considered by Stich and Laurence. That account turns on the notion of global supervenience. Whilst Stich and Laurence consider a global supervenience approach to naturalism, I argue that the global supervenience approach I favour comes closer to capturing what we might call 'the naturalist intuition'. Stich and Laurence's general point, however, remains: even

---

1 See for example Dretske 1988; Fodor 1987; Schiffer 1982.
on the favoured account of naturalism, intentional non-naturalism does not entail intentional irrealism.

But perhaps intentional non-naturalism, together with some other true, non-question-begging, premiss or premisses, entails intentional irrealism. In Section 7.3 I consider whether the success of science provides the required premiss. Finally, in Section 7.4, I draw a general moral, concluding that the naturalistic status of the intentional properties has little bearing on the eliminativism issue.

7.2 Global Supervenience.5 The most appealing contemporary accounts of naturalism are based on the notion of global supervenience. A property is naturalistic on these accounts if and only if it globally supervenes on a suitable set of base properties. Typically, the base properties are taken to be the physical properties.6 Stich and Laurence define global supervenience as follows:

A class of properties, \( S \), **globally supervenes** on a class of properties, \( B \), if and only if, all possible worlds which are \( B \)-doppelgangers are also \( S \)-doppelgangers.7

Taking our base class to be the set of all physical properties, the intentional is naturalistic on this account if and only if every logically possible world which is a physical duplicate of the actual world is also an intentional duplicate. As Stich and Laurence point out, though, the intentional is not naturalistic on this account of naturalism, for there exist logically possible worlds which are physical but not intentional duplicates of the actual world. Consider two cases. (1) Say that the world conforms to the naturalist intuition and that, moreover, my (actual) desk has no mind. Then we can readily imagine a world which is a physical but not intentional duplicate of this one – a world in which epiphenomenalism is true and my desk has a mind. (2) On the other hand, say that materialism is false at the actual world and that, moreover, the (actual) Prime Minister has an epiphenomenal mind. Then once again we can imagine a world which is a physical but not intentional duplicate of this one – a world in which epiphenomenalism is false and the Prime Minister has no mind. Note, though, that whilst the intentional properties turn out not to be naturalistic on the account of naturalism under consideration, nothing untoward follows. The existence

---

5 This Section is based in part on Ravenscroft 1993.
6 Which properties are the physical properties? I explore this issue below.
7 Stich & Laurence 1993: 103.
of logically possible worlds which are physical but not intentional duplicates of the actual world tells us nothing about the existence or otherwise of intentional properties at the actual world. That is, intentional non-naturalism thus defined does not entail intentional irrealism.

The global supervenience account of naturalism presently being considered is, however, defective. It fails to capture the naturalist attitude emphasized in the works of naturalists such as Armstrong, Devitt, Jackson, Lewis, Smart and Sterelny. For the naturalist holds that the world is, as a matter of contingent fact, naturalistic; she does not take naturalism to be a necessary truth. Yet, on the global supervenience approach considered by Stich and Laurence, naturalism is upheld as a necessary truth: any world which is a physical duplicate of this one is an intentional duplicate. We might hope therefore to capture the naturalist intuition by restricting the set of worlds appealed to in the global supervenience condition. Thus Lewis proposes:

Among worlds where no natural properties alien to our own are instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates.

Lewis defines a property as a class of actual or possible entities. Since every class of entities defines some property, the sharing of properties simpliciter is of no great interest. Lewis therefore defines a special class of properties: 'the ones whose sharing makes for resemblance, and the ones relevant to causal powers'. These are the natural properties. A property is alien to our world if it is not instantiated at our world. According to Lewis, then, the actual world is naturalistic if and only if any world which contains no natural properties alien to the actual world, and which is a physical duplicate of the actual world, is a duplicate simpliciter of the actual world. Lewis' approach to naturalism handles the problematic cases considered above. Call the worlds which are physical duplicates of the actual world and which instantiate no natural properties alien to the actual world, Lewis worlds. Again consider two cases.

---

8 See Armstrong 1968, 1978; Devitt & Sterelny 1987; Jackson forthcoming; Lewis 1983a; Smart 1963, 1978; Sterelny 1990. Note that Jackson's commitment to naturalism is not global. As we saw in Section 5.4, Jackson is not a naturalist about qualia.

9 This point has been variously made. See for example Lewis 1983a and Jackson forthcoming.

10 Lewis 1983a: 363. See also Jackson forthcoming. Jackson deploys the perspicuous label minimal duplication for this way of capturing naturalism. Note also that Lewis calls his position 'materialism' rather than 'naturalism'. Unlike some writers (for example Armstrong 1978), I make no distinction between naturalism and materialism.

11 See Lewis 1983a: 344.

12 Lewis 1983a: 347.
(1) The actual world conforms to the naturalist intuition. Then epiphenomenalism and other forms of dualism are false at the actual world and so epiphenomenal and other dualist properties are alien properties. It follows that epiphenomenal and other dualist properties are not instantiated at any Lewis world and thus that neither epiphenomenalism nor any other form of dualism is true at any Lewis world. Every Lewis world will therefore be a intentional duplicate of the actual world, as required.

(2) Epiphenomenalism or some other form of dualism is true at the actual world. Then there exist Lewis worlds which are not intentional duplicates of the actual world – worlds at which epiphenomenalism and other forms of dualism are false. The actual world is therefore not naturalistic, again as required.

Before assessing Stich and Laurence's claim in the light of this new account of naturalism, it is worth spending some time thinking about the subvenient base properties to which this account appeals. Which properties are the physical ones? An obvious suggestion is that the physical properties are those properties quantified over by physics, but we can immediately see that this suggestion is problematic. Physics has advanced dramatically over the last few hundred years, and to a large extent that advance has involved the introduction of new predicates. By taking the properties over which contemporary physics quantifies as the subvenient base, the naturalist may be shortchanging herself: the properties on which intentionality supervenes maybe intuitively physical but not yet uncovered by physics. It is tempting to respond to this problem by taking the subvenient base to be the properties quantified over by ideal or complete physics. But this response, too, is problematic, for the intentional non-naturalist can reasonably complain that ideal or complete physics might quantify over minds or other bearers of intentional properties: perhaps intentionality is part of the fundamental structure of the physical world. If that turns out to be the case, the intentional properties will indeed supervene on the physical ones, but in a way that should be of no comfort to the naturalist.

J.J.C. Smart and David Lewis have responded to the problem of identifying the subvenient base by claiming that those areas of contemporary physics relevant to the subvenient base of intentionality are essentially complete. Both support their claim by citing Feinberg's 1966 paper 'Physics and the Thales Problem'. The physics of blackholes and the interior of stars may still be evolving, but the physics of 'mild conditions' (the phrase is Feinberg's) is pretty much settled. Since intentionality – or at least our intentionality – occurs under mild conditions, the relevant subvenient

---

13 See Smart 1978 and Lewis forthcoming.
14 Feinberg 1966.
properties are just those properties which contemporary physics has posited to deal with mild conditions; that is, those properties which are thoroughly understood. Needless to say, the properties posited by contemporary physics to deal with mild conditions are properties with which the naturalist feels entirely comfortable. If intentionality globally supervenes on those properties, the naturalist's case is won.

Attractive as this response is, I do not think that it is adequate. How is the claim that the physics of mild conditions is essentially complete to be substantiated? Smart and Lewis rest their case on the word of a single physicist writing in the *Journal Of Philosophy* over twenty five years ago. At the very least a more general survey of contemporary physicists is required. Moreover, it is not clear that the opinions of contemporary physicists count for a great deal when attempting to answer a question such as this. It is the history of physics that alert us to the difficulties of gauging the discipline's completeness, not the writings of the present generation of physicists. Think of the sense of certainty present in 18th Century classical Newtonian writings. It is only with the benefit of hindsight that that certainty seems misplaced.

A more promising way to characterize the subvenient base is found elsewhere in Lewis' writings. In 'New Work for a Theory of Universals’, Lewis characterizes the subvenient base in terms of properties 'not too different' from the properties quantified over by present-day physics.¹⁵ The intentional is naturalistic, on this view, if and only if it globally supervenes (in the restricted sense outlined above) on properties suitably similar to the properties recognized by today's physics. At the heart of Lewis' account of naturalism, then, lies an appeal to an unanalyzed notion of similarity. Such appeals are a recurrent theme in Lewis' work.¹⁶ Lewis argues that we ought not balk at such appeals. Progress can be made on various fronts (characterizing naturalism; giving the semantics for counterfactuals) by appealing to similarity. It is true that we are left with the problem of analyzing the notion of similarity, but that is a problem we were faced with anyway.

Lewis' position can be strengthened by observing that there is surprising agreement between naturalists and non-naturalists about the sense of similarity at work. Or, to put the point another way, there is remarkable agreement between naturalists and non-naturalists about which properties can be legitimately allowed into the subvenient base. It remains true that the subvenient base properties cannot be characterized by a short list of necessary and jointly sufficient conditions; that is, it remains true that we cannot give a classical conceptual analysis of the notion of

---

¹⁶ See for example Lewis 1973.
physical property. But, as stressed in earlier chapters, there are very few properties for which classical conceptual analyses can be given. We cannot give a classical analysis for the property *bovine*, but nevertheless there is near universal agreement about which individuals instantiate that property. Similarly, we cannot give a short list of necessary and jointly sufficient conditions which characterizes the physical properties, but there is near-universal agreement about which properties are the physical ones. Most importantly, there is near-universal agreement on both sides of the naturalist/non-naturalist divide.

It is time to get back to Stich and Laurence's claim. Recall that Stich and Laurence argue that, on every available account of naturalism, either the intentional turns out to be naturalistic or, if it does not, intentional non-naturalism does not entail intentional irrealism. On the account of naturalism endorsed above, does intentional non-naturalism entail intentional irrealism? Certainly not. Say that epiphenomenalism is true at the actual world. It follows that certain non-physical properties (the epiphenomenal ones) are instantiated at the actual world, and that therefore there exist Lewis worlds which are not intentional duplicates of the actual world – Lewis worlds at which epiphenomenalism is false. On the assumption of epiphenomenalism, then, the actual world is non-naturalistic according to the account of naturalism under consideration. But epiphenomenalism is a theory of intentionality: if epiphenomenalism is true at the actual world then intentional irrealism is false. Intentional non-naturalism therefore does not entail intentional irrealism. Stich and Laurence's general claim has been vindicated.

On reflection, this result is not at all surprising. Old-fashioned naturalists like Armstrong and Smart articulated naturalism as a doctrine opposed to epiphenomenalism and other forms of dualism; they did not articulate naturalism as a doctrine opposed to intentional irrealism. Their concern was with the nature of intentionality, not its existence. Imagine a simple tree structure. The left branch comprises intentional irrealist positions; the right branch realist positions. The right branch bifurcates into naturalist versus dualist (including epiphenomenalist) accounts of intentionality. Naturalists like Armstrong and Smart were interested in distinguishing the naturalist accounts of intentionality from the dualist ones. That is, their interest focussed exclusively on the righthand side of the tree. They were not concerned with separating positions on the right side of the tree from positions on the left. It is not surprising, therefore, that intentional non-naturalism does not entail intentional irrealism. At best it entails a disjunction: either intentional irrealism is true or some form of dualism is true.
The last remark suggests a strategy for arguing to intentional irrealism from intentional non-naturalism. I said that intentional non-naturalism entails that either dualism or intentional irrealism is true. This suggests that intentional non-naturalism, plus some other premiss which entails the falsity of dualism, would entail intentional irrealism. I explore that strategy in the next Section.

7.3 Arguing For Irrealism. If we were convinced that the world was thoroughly naturalistic, we could argue from intentional non-naturalism to intentional irrealism quite straightforwardly:

(1) Global naturalism is true; that is, all the properties instantiated at the actual world are naturalistic.

(2) Intentional properties are not naturalistic.

Therefore,

(3) Intentional properties are not instantiated at the actual world; that is, intentional irrealism is true.

Note that premiss (1) entails the falsity of dualism; if all the properties instantiated at the actual world are naturalistic then epiphenomenalism and other forms of dualism are false. The argument from (1) and (2) to (3) is thus an instance of the strategy proposed at the end of the last Section. Moreover, the argument from (1) and (2) to (3) is valid, and so the plausibility of its conclusion rests entirely on the plausibility of its premisses. There is, though, considerable tension between premisses (1) and (2). For it might seem that premiss (1) is secure precisely to the extent that premiss (2) is challenged; that global naturalism requires the naturalization of the intentional. From this perspective the argument just given is rather curious. For it seems that we should not accept premiss (1) unless we already have good grounds for rejecting non-naturalistic accounts of intentionality, but premiss (2) asserts that the intentional properties are non-naturalistic.

Note, though, that the tension between (1) and (2) only exists if irrealism is denied; that is, if (3) is rejected. For if we accept irrealism we can endorse both the claim that every property which exists at the actual world is naturalistic and the claim that the intentional properties are non-naturalistic: this world is naturalistic but
instantiates no intentional properties; other worlds instantiate intentional properties but are not naturalistic. We have therefore an inconsistent triad comprising statements (1), (2), and the negation of (3). There are three philosophically interesting ways of resolving the inconsistency. Thus the anti-naturalist about intentionality typically asserts (2) and the negation of (3), and argues to the negation of (1); the naturalist who remains realist about intentionality accepts (1) and rejects (2) and (3); and the naturalist who endorses intentional irrealism argues, as we have seen, from (1) and (2) to (3). What can be said in favour of the irrealist resolution of the inconsistency between (1), (2), and the negation of (3)?

Recall that we are concerned with the worry, aired by Fodor and others, that intentional irrealism follows from intentional non-naturalism. So let us, for the sake of argument, grant premiss (2) (intentional properties cannot be naturalized) and cast around for support for premiss (1) (all actual properties are naturalistic). How can premiss (1) be established? One suggestion is that the triumph of the sciences makes it overwhelmingly probable that all the actual properties are naturalistic. After all, science has proven to be immensely successful at explaining the world, and yet it only quantifies over naturalistic properties. An idea something like this seems to lurk behind Schiffer's concern about the failure to naturalize intentionality. Thus he writes that:

We are physical objects in a physical world; our bodies, collections of molecules, move, and amid the myriad products of these movements are marks and sounds. These physical phenomena have physical explanations, forthcoming, in principle, from physics, at the most fundamental level, to neurobiological science at more specialized levels. So much for the unassailable.

At the same time we are apt, pretheoretically, to suppose that some of these marks and sounds have semantical properties, and those who produce them have psychological states, notably beliefs, desires, and intentions. . . [But] we should not be prepared to maintain that there are semantical and psychological facts unless we are prepared to maintain that such facts are completely determined by, are nothing over and above, physical facts.17

The idea under consideration, then, is this. The argument from (1) and (2) to (3) is valid. We are accepting premiss (2) for argument’s sake. The plausibility of the conclusion rests, therefore, on the plausibility of premiss (1). What can be said in favour of the claim that all the actual properties are naturalistic? Suggestion: The success of science, which is thoroughly naturalistic, strongly supports the claim that all actual properties are naturalistic. Something like this idea, I claim, lies behind the worries aired by Fodor and the rest.

I do not think that premiss (1) can be successfully defended by appealing to scientific triumph. Let me begin to air my discomfort by averting to the history of science. William Harvey is best known for his work on the circulation of the blood. Indeed, he is often upheld as a model scientific investigator. Less well known is his work on theories of generation, or, as we would say today, theories of sexual conception. Harvey’s theory of conception was, in effect, Aristotelian. He believed that the function of copulation was to transfer information from the male to the female—information which the female then used in the development of the embryo. Of course, present-day accounts of conception also postulate the transfer of information from the male to the female during copulation—genetic information. But we presently believe that that information is carried by a physical structure—DNA. Harvey, on the other hand, believed that the information was conveyed—that the uterus was informed, so to speak—by non-naturalistic means. Note that this was not just idle speculation on Harvey’s part, induced perhaps by Aristotle’s influence. Maybe Aristotle did indeed influence Harvey’s thought, but Harvey also conducted controlled experiments and made extensive observations. (Indeed, his methodology was precisely that advocated by his contemporary, Francis Bacon.) Harvey conducted careful experiments on deer, and detected no material in the uteruses of deer that had recently copulated. He did not, of course, deny that semen passed from male to female during copulation, but he could find no trace of semen in the female’s uterus. Harvey concluded that conception involved not the transfer of material, but the transfer of Aristotelian form. Moreover, he observed that the texture of the uterus was very similar to the texture of the brain—the organ of information reception par excellence.18

Harvey provides an example, then, of a major scientist offering an argument to the best explanation that involves non-natural properties. Of course we now know

---

18 My discussion of Harvey is based on Gasking 1967. See especially Ch. 2.
that Harvey was wrong and that material is passed from the male to the female during conception. Nevertheless, Harvey's example helps to establish the important point that science is not committed to naturalistic explanations — that science does not have to respect the metaphysical intuitions of the naturalist. There can be no guarantees in advance that science will not quantify over non-natural properties. We can, if we like, build into our definition of 'science' the claim that science only quantifies over natural properties; that is, we can use the term 'science' in a way that makes it analytic that science does not quantify over non-natural properties. However, note that if we insist on using 'science' that way we are forced to rule that Harvey's investigations of generation were not science. That is, we are forced to be strongly revisionist about the history of science.

It might be argued that the history of science does in fact support premiss (1); that the history of science is a long progression from non-naturalistic theories (like Harvey's) to naturalistic ones (like contemporary accounts of generation). But is contemporary science naturalistic? The non-naturalist about intentionality may well argue that premiss (1) is question-begging. Psychology is a science. It quantifies over various properties. Are those properties naturalistic or not? The defender of premiss (1) cannot simply assert that the properties over which psychology quantifies are naturalistic, for the properties over which psychology quantifies include the intentional properties, and the naturalistic status of the intentional properties is precisely what is at issue.

To defend premiss (1) by appealing to the triumph of science is clearly to resort to an inductive argument. Modern science does not encompass every phenomenon, nor does it quantify over every property. Defending premiss (1) by appealing to the triumph of science thus involves generalizing from those properties science does quantify over to those it does not. The naturalist can attempt to extend the inductive generalization to cover the psychological properties; that is, she can argue that the properties over which every other science quantifies are naturalistic and so the properties quantified over by psychology are naturalistic too. But the non-naturalist can reasonably resist such a move by claiming that the phenomena addressed by psychology are of a fundamentally different kind to the phenomena addressed by the other sciences, and thus the induction fails. And of course this is precisely the strategy the dualist adopts, arguing that intentionality and consciousness are radically different kinds of properties from inertia and charm.

We have been considering an argument from intentional non-naturalism to intentional irrealism, using as an additional premiss the claim that all the actual
properties are naturalistic. We have seen that that argument, whilst valid, is not very convincing. The claim that all properties are naturalistic rests on an inductive generalization that the non-naturalist need not endorse. Stich and Laurence's claim that Fodor et al have been unnecessarily alarmed by intentional non-naturalism has been vindicated. Intentional non-naturalism does not entail intentional irrealism, nor is there a persuasive argument from intentional non-naturalism, plus some other plausible, non-question-begging premiss, to intentional irrealism.

7.4 Concluding Remarks. We are committed to the existence of the properties over which physics quantifies. Physics is enormously successful, and it seems entirely reasonable to accept the properties it posits. What else, if anything, do we have to allow into our ontology? The naturalist insists that nothing further is required: if we fix the physical properties we fix everything. The non-naturalist, on the other hand, thinks a richer ontology is needed. According to the non-naturalist, the physical properties are inadequate: fixing the physical properties leaves some business unfinished. In particular, fixing the physical properties does not fix the intentional properties. The non-naturalist thus allows into her ontology epiphenomenal or other dualist properties in order to account for intentionality. This is a familiar strategy. One of Descartes' motivations for accepting an ontology that embraced non-naturalistic properties was his conviction that naturalistic properties alone could never account for our ability to understand language, and an enduring reason for endorsing epiphenomenal properties is, of course, the apparently insurmountable difficulty of accounting for qualia in naturalistic terms.

Since we are committed to allowing the naturalistic properties into our ontology, we might accept the following methodological principle: A theory that quantifies exclusively over naturalistic properties is to be preferred, ceteris paribus, to one which in addition quantifies over non-natural properties. This principle is simply an application of Occam's razor: we are already committed to the naturalistic properties; do not add further properties unnecessarily.

Note that both naturalists and non-naturalists might agree to this principle. The non-naturalist does not regard herself as violating Occam's razor; on the contrary, she regards herself as quite justified in adding new properties. Without postulating epiphenomenal properties, she argues, intentionality remains forever unexplained. The naturalist, on the other hand, thinks that no non-naturalistic properties need be postulated: the battery of properties provided by physics is sufficient to account for intentionality. That both the naturalist and the non-naturalist are prepared to subscribe
to the proposed principle is telling. It reveals that Occam’s razor is incapable of settling the naturalist/non-naturalist debate in its present form. There is a substantive issue at stake here, not a mere quibble about who has the cleanest ontology.

These considerations suggest that the irrealist strategy which we have been scrutinizing in this Chapter is radically mistaken. That strategy involves arguing, in one way or another, from the alleged non-naturalism of intentionality to intentional irrealism. But the strategy should be turned on its head. The intentional properties will stand or fall on whether or not they earn their keep in theorizing about the mind. If our best theories of the mind turn out to be intentional then we should allow intentional properties into our ontology; if not, we should repudiate them. If they do earn their keep – if we are committed to intentional realism – then the philosopher of science can pursue the interesting question of whether or not the postulated intentional properties are naturalistic. That will be no easy task, for it amounts to determining whether or not the intentional properties supervene on the physical properties. And, in Terence Horgan’s words, the subvenient basis for intentionality is likely to be ‘complex and baroque’.

But, difficult though it might be, the task pursued by the philosopher of science is a secondary one. It is pursued after intentional realism has been established, not before it.

The naturalism issue is a red-herring. Nothing of significance stands or falls by it. If our best theories of mind turn out to be intentional theories then we should be intentional realists, whether or not a physical subvenient base for intentionality can be located. To think otherwise is to return – ironically – to Cartesian first philosophy; it is to let our a priori metaphysical scruples determine the direction of empirical science. Better to let the science speak for itself.

CHAPTER EIGHT

THE SCIENTIFIC IMAGE

8.1 Introduction. This thesis has been an extended attempt to articulate the nature and status of folk psychology. I began by distinguishing between folk psychology (internal) and folk psychology (external). Very little is at present known about folk psychology (internal), so the focus has largely been on folk psychology (external). I have argued that commonsense functionalism provides the correct meta-theory for folk psychology (external); that is, (external) folk psychological mental states are loci in a causal network involving sensory inputs, mental states, and behavioural outputs. This conception of mental states raises the possibility of eliminativism: if the folk-specified causal roles – or roles sufficiently similar to them – are not occupied in the actual world, then there are no such things as (external) folk psychological beliefs and desires. Chapters 6 and 7 were devoted to examining pro- and anti-eliminativist arguments, and concluded that the threat of eliminativism has been greatly exaggerated. The anti-eliminativist arguments on offer are unconvincing, and the predictive success of folk psychology (external) gives us some grounds for confidence in its theoretical posits.

It is one thing, though, to hold eliminativism to be unlikely, and quite another to claim that folk psychology (external) will form the basis of a scientifically respectable psychology. As we saw in Section 6.1, there are properties which cannot seriously be regarded as candidates for elimination but over which no science quantifies. Being a warm coat is such a property. In this last Chapter I want to address the relationship between folk psychology (external) and scientific psychology. In Section 8.2 I survey three arguments to the effect that folk psychology (external) will not find a place amongst the respectable sciences. Two of those arguments proceed by identifying a property of (external) folk psychological propositional attitudes and then arguing that no state with that property could be quantified over by a respectable scientific psychology. The third

---

1 What is a respectable (or as I will sometimes say, ‘serious’) scientific theory? I cannot give necessary and jointly sufficient conditions for the concept of a respectable science. All I can offer is some examples which I take to be paradigmatic: relativity theory; classical thermodynamics; the theory of evolution.
proceeds by locating a property of (external) folk psychological generalizations and then arguing that no theory that relies on such generalizations could be scientifically respectable. I urge that none of these arguments is convincing, and thus that there is nothing to prevent folk psychology (external) – perhaps enlarged and/or regimented – from forming the basis of a scientific account of the mind. Indeed, in Section 8.3 I offer an extended example of contemporary scientific psychological theorizing which is essentially (external) folk psychological.

8.2 Folk Psychology (External) As Science (I).
Could folk psychology (external) form the basis of a scientific theory of the mind? There are at least three (non-eliminativist) arguments in the literature which suggest that it could not. Two of the arguments proceed by locating properties of (external) folk psychological propositional attitudes which, it is alleged, are incompatible with the scientific enterprise. The third proceeds by locating a property of (external) folk psychological generalizations and then arguing that generalizations which exhibit that property have no place in science. Let us take each argument in turn.

First argument. As we saw in Section 4.5, Fodor has objected to content holism on the grounds that if content is holistic, intentional psychology would be unworkable. Thus he has written that

[p]resumably an event (e.g., the production of behaviour by some organism) would fall within the domain of [intentional] psychology in virtue of instantiating one of its generalizations. And presumably such generalizations would apply to an organism at a time in virtue of the intentional state(s) that the organism is in at the time. The way it ought to go is that the theory says things like: 'From any organism x that believes such and such and desires so and so, you get behaviors of type . . . . blah.' You can, therefore, use the theory to predict that this organism x will give behavior of the type . . . . blah if you can identify this x as believing such and such and desiring so and so. This is just a long form of the truism that one way that intentional psychologies achieve generality is by quantifying over all the organisms that are in a specified intentional state.

But now . . . . if we follow Meaning [ie content] Holism . . . . it's going to turn out de facto that no two people (for that matter, no two time
slices of the same person) ever are in the same intentional state. (Except, may be, by accident.) So no two people will ever get subsumed by the same intentional generalizations. So intentional generalizations won't, in fact, succeed in generalizing. So there's no hope for an intentional psychology.2

For Fodor, that there is no hope for an intentional psychology pretty much entails that there is no hope for a scientific psychology, because he takes it that all hope for a scientific psychology lies with intentional psychology. We might disagree with Fodor about this; nevertheless, it is true that if folk psychology (external) is to be the basis of a respectable scientific psychology, it must be possible to form useful generalizations about people who hold (external) folk psychological beliefs and desires.

In Section 4.5 I endorsed Fodor's asymmetric dependency theory of content. Whilst Fodor claims that his theory of content is strictly atomistic, I urged that Fodor's theory is in fact holistic. The holism to which Fodor's theory is committed is, however, quite benign. In particular, it does not prevent us forming classes of agents who hold some specified belief or have some specified desire. We are therefore free to form useful generalizations with the predicates of folk psychology (external).

Second Argument. At least in some contexts folk psychology (external) individuates propositional attitudes widely. Thus, as we saw in Section 3.4, in some contexts folk psychology (external) wants to distinguish between Oscar's belief about the stuff in the glass in front of him and Twin Oscar's belief about the stuff in the glass in front of him. Recall the example. Oscar lives on Earth and is in causal contact with H2O. Twin Oscar lives on Twin Earth which is a molecule-for-molecule duplicate of Earth except that the stuff that flows from the tap, is essential for life, and falls from the sky as rain, is XYZ not H2O. Both Oscar and Twin Oscar have a belief which they would express by saying that the stuff in the glass in front of them is water. Do Oscar and Twin Oscar express the same belief with that form of words? I argued that, in some contexts at least, the answer is No. Oscar's belief is about H2O; Twin Oscar's about XYZ. It follows that at least sometimes (external) folk psychological propositional attitudes are individuated widely; that is, at least

---

some (external) folk psychological propositional attitudes supervene in part on states outside the agent's skin.

Fodor, amongst others, has argued that the wide individuation of the propositional attitudes presents a problem for a serious intentional psychology. It is obvious that, if the propositional attitudes are individuated widely, then they do not supervene on brain states. After all, Oscar and Twin Oscar have, by hypothesis, identical brain states but distinct (wide) propositional attitudes. 'But', claims Fodor, 'it's arguable that any scientifically useful notion of psychological state ought to respect supervenience; mind/brain supervenience (and/or mind/brain identity) is, after all, the best idea that anyone has had so far about how mental causation is possible'. The point is that psychology is in the business of predicting behaviour and therefore needs a taxonomy which distinguishes properties or entities on the basis of their causal powers. From the perspective of predicting behaviour, Oscar's belief about the stuff in the glass in front of him is identical to Twin Oscar's. It follows that we should treat their beliefs about the stuff in the glass as identical. That in turn entails individuating beliefs narrowly, and yet as we have seen that at least some (external) folk psychological propositional attitudes are individuated widely. (External) folk psychological belief individuation thus seems at odds with the hopes of a scientific intentional psychology.

Fodor's response to this problem is to turn revisionist: he is quite happy to jettison folk psychology (external)'s commitment to the wide individuation of the propositional attitudes and accept narrow individuation. However, it is not at all clear that the wide individuation of entities is impermissible in the sciences. Think of the concept of fitness in evolutionary biology. Pete the possum lives in environment E₁ and has a fitness quotient (on some scale or other) of F in E₁. Twin Pete is a molecule-for-molecule replica of Pete but lives in environment E₂. Can we assume that Twin Pete has a fitness quotient of F in E₂? Certainly not. Fitness supervenes not simply on the intrinsic properties of the organism but on the properties of the organism and its environment, and is therefore a wide property. Nevertheless, fitness is a perfectly respectable scientific property that figures in causal explanations – we explain the spread of a gene through a gene pool in terms of the fitness that gene confers upon individuals in the relevant environment.

We thus have three claims that are in tension. (1) Fodor's stricture that 'what you need in order to do science is a taxonomic apparatus that distinguishes between

---

3 See in particular Fodor 1987, Ch. 2. See also Burge 1979, 1986.
4 Fodor 1987: 30.
things insofar as they have different causal properties, and that groups things together insofar as they have the same causal properties. 5 (2) The claim that differences in wide properties do not, ceteris paribus, make for differences in causal properties. And (3), the observation that fitness is both relevant to causal powers and a scientifically respectable property. Since fitness is a wide property, by (2) differences in fitness should, ceteris paribus, make no difference to causal properties. Thus, by (1), fitness should play no part in a respectable science. Yet, by (3), fitness is both relevant to causal powers and scientifically respectable.

It is clear that we should reject (2): differences in wide properties can make for differences in causal powers. Pete and Twin Pete differ only in the wide property of fitness. Nevertheless, Pete has causal powers that Twin Pete lacks. In particular, Pete has causal powers that will manifest themselves (ceteris paribus) in his having a larger number of offspring than Twin Pete.

How is it, then, that differences in wide properties seemed causally irrelevant in the Oscar/Twin Oscar example? Wide properties appeared to be irrelevant to causal powers in that example because the environments in which Oscar and Twin Oscar exist are in effect identical. By hypothesis, they vary only in that one contains H_2O and the other XYZ, and we are told that all of the run-of-the-mill causal properties of H_2O and XYZ are the same. So it is not surprising that we were happy to disregard Oscar and Twin Oscar's wide properties. The example is set up so that differences in such properties can have no causal effects. In normal circumstances, though, differences in wide beliefs typically entail differences in narrow beliefs which in turn entail, ceteris paribus, differences in behaviour. That is, to use the terminology introduced in Section 4.6, the language of thought contains very few homonyms. It follows that a scientific intentional psychology will in normal circumstances fruitfully quantify over wide propositional attitudes.

Third argument. Say that S wants a cappuccino and believes that the Calypso Cafe serves the best coffee on campus. As (external) folk psychologists we readily predict that, ceteris paribus, S will walk over to the Calypso. The ceteris paribus clause is, obviously, carrying a good deal of the predictive burden. Without it the inference from 'S wants a cappuccino' and 'S believes that the Calypso Cafe makes the best coffee on campus' to 'S will walk over to the Calypso' is less than plausible: perhaps S will drive to the Calypso; perhaps she has quarreled with the Calypso management and vowed never to return; perhaps she believes that the

5 Fodor 1987: 34.
Calypso is closed for the vacation. Moreover, the details of the *ceteris paribus* clauses of (external) folk psychological explanations are notoriously difficult to fill in. In the example given we can make a start with 'S will not drive to the Calypso', 'S has not quarreled with the Calypso management and vowed never to return', and 'S does not believe that the Calypso is closed for the vacation'. But obviously this is only the beginning.

Folk psychology (external) is thus up to its ears in *ceteris paribus* clauses that are, in practice, ineliminable. It might seem that any serious science cannot be thus dependent on *ceteris paribus* clauses. The presence of extensive *ceteris paribus* clauses saves a theory from honestly facing the tribunal of empirical testing: a mistaken prediction does not count against the theory; it merely shows that *all was not equal*. Folk psychology (external) is thus very unlikely to take its place amongst the sciences. Or so it might be argued.

This objection to folk psychology (external)'s potential status as a science is, however, overstated. First, notice that folk psychology (external) is not alone in appealing extensively to *ceteris paribus* clauses. As Fodor has pointed out, the perfectly respectable science of geology makes extensive use of *ceteris paribus* clauses. 'Meandering rivers erode their outer banks, *ceteris paribus* is an entirely acceptable geological generalization'. In Fodor's view, only the laws of basic physics are free from *ceteris paribus* clauses – all the laws offered by the so-called 'special sciences' are hedged about with *ceteris paribus* clauses. Indeed, any non-idealized, real world, application of physics will use *ceteris paribus* clauses extensively. When the physicist predicts that an electron in an electro-magnetic field will accelerate at such-and-such a rate, the prediction assumes that the electron is not subject to other forces, for example, gravity.

In the case of physics, physical properties suffice to spell out the *ceteris paribus* clauses. Thus, in the case just given, the *ceteris paribus* clause can be spelt out in terms of other physical forces that might act upon the electron, for example, gravity. As Fodor points out, physics is unique in this respect. In the special sciences the *ceteris paribus* clauses must often be spelt out in terms of properties which do not lie within the domain of the science in question. Thus, whilst it is true that meandering rivers erode their outer banks unless someone builds an erosion-resistant concrete wall, the property *is an erosion-resistant concrete wall* is not a geological property.

---

6 See Fodor 1987: 4-5.
7 See Fodor 1987: 5-6.
What matters is not that the *ceteris paribus* clauses of a science can be eliminated by reference to properties quantified over by that science (only in physics will that happen), but that in principle the *ceteris paribus* clauses can be eliminated by reference to the properties of *some science or other*. And there is no reason to think that this is not the case with folk psychology (external), anymore than with any other special science. (External) folk psychological predictions can thus honestly face the tribunal of empirical evidence. It is true that predictive failure may be due either to the poverty of the theory or to an errant *ceteris paribus* clause. But, in principle at least, the *ceteris paribus* clause can be detailed, thus allowing us to correctly apportion the blame.

8.3 Folk Psychology (External) As Science (II). In this Section I sketch an example of contemporary psychological research which can reasonably be described as based on folk psychology (external) in that it very much takes for granted the sorts of mental states posited by folk psychology (external). The example demonstrates, moreover, how a scientific psychology that exploits the posits of folk psychology (external) can be constrained by, and constrain, other scientific endeavours— in this case anthropology and evolutionary biology. The example is based upon Robert Frank's remarkable investigation of the strategic role of the emotions, and also draws on Leda Cosmides and John Tooby's explorations of the Wason selection task.⁹

Central to Frank's work is the observation that we would do better—*ceteris paribus*—if under certain circumstances we were disposed to behave in ways contrary to our interests.¹⁰ Let me dispel the air of paradox surrounding this remark by giving three examples, all drawn from Frank's work.

*First example.* We would do better—*ceteris paribus*—if we were disposed to seek compensation (or, as I shall often say, *revenge*) even though the cost of seeking compensation exceeded the gain:

Suppose Smith grows wheat and Jones raises cattle on adjacent plots of land. Jones is liable for whatever damage his steers do to Smith's wheat. He can prevent damage altogether by fencing his land, which would cost

---

⁸ This Section is based on Ravenscroft (in preparation).
¹⁰ Derek Parfit discusses a series of very similar claims in his *Reasons And Persons*. See Parfit 1984: 1-17.
him $200. If he leaves his land unfenced his steers will eat $1000 worth of wheat. Jones knows, however, that if his steers do eat Smith's wheat, it will cost Smith $2000 to take him to court.

These court costs notwithstanding, Smith threatens to sue Jones for damages if he does not fence his land. But if Jones believed Smith to be a rational, self-interested person, this threat is not credible. Once the wheat has been eaten, there is no longer any use for Smith to go to court. He would lose more than he recovered.

If Smith could make a binding agreement to go to court in the event of damage, however, his problem would be solved. Knowing a damage suit was inevitable, Jones would then have nothing to gain by not fencing his land, and there would thus be no need for Smith to incur the costs of suing. 11

Second example. We would do better – ceteris paribus – if we were disposed to refuse an unfair deal, even though we stood to gain by it.

Smith and Jones face the opportunity of a profitable joint venture. There is some task that they alone can do, which will net them $1000 total. Jones has no pressing need for extra money, but Smith has important bills to pay. It is a fundamental principle of bargaining theory that the party who needs the transaction least is in the strongest position. The difference in their circumstances thus gives Jones the advantage. Needing the gain less, he can threaten, credibly, to walk away from the transaction unless he gets the lion's share of the take, say $800. Rather than see the transaction fall through, it will then be in Smith's interests to capitulate.

Smith could have protected his position, however, had he been able to make a binding commitment not to accept less than, say, half of the earnings. 12

---

12 Frank 1988: 48-49.
Third example. We would do better – ceteris paribus – if we were disposed to cooperate, even when it would be in our interests to cheat.

Two persons, Smith and Jones, can engage in a potentially profitable joint venture, say, a restaurant. Their potential for gain arises from the natural advantages inherent in the division and specialization of labor. Smith is a talented cook, but is shy and an incompetent manager. Jones, by contrast, cannot boil an egg, but is charming and has shrewd business judgment. Together, they have the necessary skills to launch a successful venture. Working alone, however, their potential is much more limited.

Their problem is this: Each will have opportunities to cheat without the possibility of detection. Jones can skim cash from the drawer without Smith's knowledge. Smith, for his part, can take kickbacks from food suppliers.

If only one of them cheats, he does very well. The non-cheater does poorly, but isn't sure why. His low return is not a reliable sign of his having been cheated, since there are many benign explanations why a business might do poorly. If the victim cheats also, he, too, can escape detection, and will do better than by not cheating; but still not nearly so well as if both had been honest.

Once the venture is under way, self-interest unambiguously dictates cheating. Yet if both could make a binding commitment not to cheat, they would profit by doing so.13

In each of the cases just described, it is in the agent's interests (ceteris paribus) to be disposed to behave in a fashion contrary to her interests. In the cases just considered, Smith's being disposed to behave in ways contrary to his interests deters Jones from behaving in ways contrary to Smith's interests. It must be stressed that Frank is not claiming that it is in the agent's interests to behave in ways that are contrary to her interests. Crucially, the agent's dispositions to behave in ways

---

13 Frank 1988: 47-48. The case Frank describes is, of course, an example of the Prisoner's Dilemma.
contrary to her interests typically remain unactualized, and thus the agent's actual behaviour is rarely contrary to her interests.

How can an agent dispose herself to behave in ways contrary to her interests? Think of the first example given above. Going to court will only make things worse for Smith: he stands to lose a further $2000 on top of the $1000 of lost wheat. Once the damage is done, what will motivate him to seek revenge? Smith faces what Frank calls a commitment problem. Self-interest demands that he commit himself to behaving in a certain way should certain circumstances arise, but once those circumstances have arisen he is no longer motivated by self-interest to behave in that way. Moreover, the agent's commitment must be publicly observable if it is to have the required effect. Thus, in the first example discussed above, Jones must know that Smith is determined to sue irrespective of the cost; in the second example, Jones must know that Smith is prepared to walk away from a deal simply because it is unfair; and in the final example, both Smith and Jones must know that the other is disposed not to cheat.

How can commitment problems be solved? Occasionally artificial devices can be introduced to solve commitment problems. Thus Thomas Schelling considers the case of a reluctant kidnapper who wants to set his victim free but cannot do so for fear that, once free, the victim will report the kidnapper's identity to the police. The victim faces a commitment problem: she needs to commit herself to not informing the police of the kidnapper's identity, but both she and the kidnapper know that once she is released she will no longer be motivated to protect the kidnapper. Schelling suggests that the victim could commit a blackmailable offense and provide the kidnapper with incontrovertible evidence of that offense. The victim is now committed to keeping the kidnapper's identity secret, for if she reveals it the kidnapper can reveal her crime to the police.14 Ingenious though this scenario may be, life rarely presents us with commitment problems which can be solved in this way. We require much less artificial solutions.

Frank's proposal is that the emotions play a fundamental role in solving commitment problems. Commitment problems arise when our interests demand that we bind ourselves to a course of action which, were it undertaken, would contravene our interests, and it is a commonplace of human behaviour that we pay scant attention to our interests when experiencing powerful emotions. Consider again the examples given above. If Jones knows that Smith is the sort of person who will

---

become angry – *irrationally* angry – if his wheat is damaged by his neighbour's cattle, then he is very unlikely to let his cattle stray onto Smith's land. For Jones knows that when Smith is very angry he is likely to be blind to his own interests and sue Jones. Similarly, if Jones knows that Smith becomes so enraged when offered unfair deals that he always refuses them, then Jones will not attempt to foist an unfair arrangement onto him. Finally, if both Smith and Jones realize that the other is the sort of person who would be overwhelmed by guilt if he were to cheat, then both can continue to cooperate. In other words, at least some emotions are dispositions to behave in ways contrary to our interests and they thus help us overcome commitment problems.

Frank's theory strongly suggests an evolutionary account of the emotions. Those of our Pleistocene forebears who experienced the sorts of emotional responses we have been discussing, and who successfully conveyed those responses to their conspecifics, would, *ceteris paribus*, have a reproductive advantage over those that did not. We would therefore expect there to exist marked similarities in emotional responses across cultures, and in fact a number of anthropologists and psychologists have argued that human emotional responses are significantly pan-cultural. Thus, for example, Paul Ekman investigated the emotional responses of New Guinea highlanders and found that not only did they produce facial expressions (fear, hope, surprise, etc) which a North American could readily identify, they also readily identified the emotions expressed by faces drawn by North Americans. In a second study, Ekman showed a film to both Japanese and American audiences and secretly videotaped their facial expressions. Both audiences produced the same expressions in response to the same parts of the film. The work of Ekman and others thus strongly supports the evolutionary account of the emotions derived from Frank's work.

Notice that the emotions which concern both Frank and Ekman are those posited by folk psychology (external). It is thus very natural to see the theory of the emotions on offer as deepening and extending folk psychology (external). Notice also that Frank's theory enjoys relations of mutual support with both contemporary anthropology and evolutionary biology. We have thus already established that folk psychology (external) can form the basis of a serious scientific investigation of the mind which enjoys relations of mutual support with other sciences. Cosmides and Tooby's investigations of the Wason selection task show us how to further extend

Frank’s theory whilst retaining its essentially folk psychological nature. It is to their work which I now turn.

As mentioned above, overcoming commitment problems requires that others be aware that the agent is disposed to behave in ways contrary to her interest: Jones must know that Smith is disposed to seek revenge; disposed to turn down unfair bargains; disposed to cooperate. Thus we have reason to posit cognitive mechanisms dedicated to determining whether or not our conspecifics are disposed to seek revenge, refuse unfair deals, or cooperate. We can expect that those of our Pleistocene ancestors which had such mechanisms would have done better, ceteris paribus, than those which did not. If Frank’s claims about the emotions are right, we would expect such mechanisms to take as input information about emotional dispositions. They may, however, takes as input other sorts of information as well. Is there any independent evidence of the existence of such mechanisms?

It appears that cheater detection mechanisms exist in other species. Thus consider the vampire bat. Vampires do in fact exist. They are a species of South American bat which live exclusively on mammalian blood and which exhibit food sharing behaviour. If a vampire hunts unsuccessfully for as little as two consecutive nights it can die of hunger. Co-operative strategies have therefore evolved. Vampires which have hunted successfully regurgitate blood for the benefit of the unsuccessful. As we might expect, vampires share food with close relatives. But they also, on occasion, share food with non-relatives. Clearly, an effective food sharing arrangement greatly benefits all concerned: the recipient avoids death by incurring a debt which is easily discharged; the donor at least partly insures itself against future disaster. Nevertheless, opportunities to cheat arise. A particular vampire will be better off if everyone else shares food with it, but it shares food with no one. Thus we would expect evolution to favour bats which can distinguish the cheaters from the co-operators. And it turns out that vampires are reasonably good at keeping track of which of their conspecifics have accepted food but not returned in kind when the opportunity arose.16

Cosmides and Tooby have, however, provided a much more direct argument for the existence of a cheater detection mechanism in humans. Their evidence comes from a rather surprising quarter – studies of so-called 'Wason selection tasks'. In a Wason selection task, the subject has to seek evidence in support of a

16 The situation in which the vampires find themselves is, of course, an iterated prisoner’s dilemma. They resolve the dilemma by detecting cheaters and refusing to "play" with them. My description of vampire bat behaviour is based on Wilkinson 1988, 1990.
conditional claim 'If p then q'. For example, the subject might be shown four cards lying on a table. The subject is told that each card has a letter on one side and a number on the other, and is then asked which of the cards must be turned over in order to establish whether the conditional 'If there is a vowel on one side of the card then there is an even number on the other side' is true.\textsuperscript{17} Figure 8.1, below, represents the four cards.

If we think about the task carefully we can see that we only have to turn over cards 1 and 4. To check whether the conditional is true we must look for cases in which the antecedent is true whilst the consequent is false. Card 1 is clearly a case in which the antecedent is true so we must turn it over and check that the consequent is also true. Card 4 is a case in which the consequent is false and so we must check the reverse side to see if it makes the antecedent true. We do not, though, have to check cards 2 and 3. 'D' does not make the antecedent true so it is pointless to investigate the truth value of the consequent by turning over the card. Similarly, '4' does not make the consequent false so again it is pointless to turn over the card.

![Figure 8.1](image)

Most people get this task wrong. In fact, less than 25\% of subjects typically get the right answer.\textsuperscript{18} (Turning over only card 1, or cards 1 and 3, are the most common incorrect responses.\textsuperscript{19}) Indeed, it can be very difficult to get some subjects

\textsuperscript{17} The example is taken from Johnson-Laird & Wason 1970.
\textsuperscript{18} Cosmides & Tooby 1992: 181.
\textsuperscript{19} Cosmides & Tooby 1992: 181.
to see that turning over cards 1 and 4 is the right answer. A natural response to these results is to conclude that humans are poor at conditional reasoning. However, further investigation reveals that the results vary with subject matter: how well people do at the Wason selection task depends upon what the conditional is about. With certain conditionals performance improves remarkably. Thus, 75% of subjects perform correctly when asked to seek evidence for the conditional 'If a person is drinking beer, then he must be over twenty years old'.

What has this got to do with mechanisms for detecting cheaters? Cosmides and Tooby have suggested that we have specialized cheater detection modules hardwired into our brains. As we have seen, there is good evolutionary grounds for positing such a module. But Cosmides and Tooby go on to account for our capacity to perform well on some Wason selection tasks but poorly on others in terms of the hypothesized cheater detection module. They take a cheater to be an individual who satisfies the antecedent but not the consequent of the conditional 'If X has received the benefit of co-operation then X has paid the cost', and claim that we are good at precisely those Wason selection tasks that involve assessing a conditional of this nature. Obviously, costs and benefits are relativized in that what is a benefit to one party in a social exchange is a cost to the other. If Cosmides and Tooby are correct, performance on Wason selection tasks should reflect this difference of perspective. In an elegant series of experiments Gigerenzer and Hug have shown that they do.

I have only briefly sketched Cosmides and Tooby's position and the empirical evidence for it. Much more could be said. Moreover, it is clear that characterizing the function of the cheater detection mechanism as simply assessing the conditional 'If X has received the benefit of co-operation then X has paid the cost' is too crude. Thus I mentioned above that 75% of subjects performed well when asked to seek evidence for the conditional 'If a person is drinking beer, then he must be over twenty years old', but it is not clear how this relates to the costs/benefits model of cheating in social co-operation situations. Exactly what are the costs and who pays them when a sixteen year old violates the drinking laws?

Cosmides and Tooby have provided independent evidence for the existence of a cheater detection module. On the basis of Frank's work we also predicted the existence of cognitive mechanisms for detecting conspecifics who are likely to seek revenge or insist on a fair deal. In a small pilot project I investigated people's ability

---

22 For details see Cosmides & Tooby 1992: 190-193.
to reason about conditionals set in contexts which involved either the possibility of revenge or the possibility of unfair dealing. The subjects (12 Rutgers University graduate students) were given two Wason selection tasks. In the first task they read a short story about Martin who is very proud of his new car. They were then asked to seek evidence in support of the conditional 'If someone ran into Martin's car in the supermarket carpark, then Martin would make them pay'. In the second task the subjects read a short story about Sally who wants to sell her car to raise money for an overseas holiday. She thinks that a fair price for her car is $5000. The subjects were then asked to seek evidence in support of the conditional 'If Sally was offered less than $5000 then she did not sell her car'. Over 90% of subjects reasoned correctly about both tasks, whereas less than 25% of subjects typically reason correctly about the Wason selection task described earlier in this Section which involved the conditional 'If there is a vowel on one side of the card then there is an even number on the other'. We therefore have some evidence for the existence of modules dedicated to recognizing conspecifics who are likely to seek revenge or insist on a fair deal. However, it cannot be stressed too much that these results are very preliminary and that more carefully controlled experiments on larger samples are called for.

To sum up, Frank's work has yielded a powerful account of the emotions and a fruitful line of research on human reasoning mechanisms. The resulting theory enjoys relations of mutual support with cognitive psychology, anthropology, and evolutionary biology. Moreover, in principle the research sketched could also come to enjoy relations of mutual support with the neurosciences. If they exist at all, the posited cognitive mechanisms must be implemented in the brain, and thus research in the neurosciences should benefit from, and benefit, the sort of psychological research we have been examining. But the states Frank's theory quantifies over are simply those of folk psychology (external): they are neither connectionist nor neurobiological. We thus have before us an example of a serious scientific investigation of the mind that is naturally viewed as building upon and deepening folk psychology (external).

8.4 Concluding Remarks. For Wilfred Sellars, one of the central concerns of philosophy is to unify the disparate manifest and scientific images of ourselves and our world. However, to many it has seemed that the manifest and scientific images of ourselves are on the verge of a total and permanent separation with catastrophic consequences. The manifest image of ourselves is that of folk psychology
(external). It is an image of agents whose behaviour is largely the product of their beliefs and desires. Eliminativism and non-intentional paradigms in the cognitive sciences threaten that image. The threat is particularly painful because we cannot but see ourselves in (external) folk psychological terms.

We can now see that such a gloomy prognosis is uncalled for. No doubt parts of the manifest image will be overhauled; no doubt research in the cognitive sciences will yield unsettling and disturbing results. (Who could read the literature on blind sight and not be unsettled? Who could hear the results of Stanley Milgram's experiments and not be disturbed?) But the threat of a lasting division between the manifest and scientific images of ourselves has been wildly overstated. The key to their unification is commonsense functionalism. Commonsense functionalism allows us to identify (external) folk psychological states with brain states, and once that identification is achieved our world is made whole again.
BIBLIOGRAPHY


- (forthcoming): 'The message is: there is no medium'.


