FUNCTIONALISM, QUALIA, AND CONTENT

by

Janet Marchel Levin

B.A., University of Chicago (1972)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

DOCTOR OF PHILOSOPHY IN
PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1980

The author hereby grants to M.I.T. permission to reproduce and
to distribute copies of this thesis document in whole or in part.

Signature of Author

Ned Block
Thesis Supervisor

Accepted by

Judith J. Thomson
Chairman, Departmental Graduate Committee

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

OCT 6 1980

LIBRARIES
FUNCTIONALISM, QUALIA, AND CONTENT

by

JANET MARCHEL LEVIN

Submitted to the Department of Philosophy on August 9, 1980, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Philosophy

ABSTRACT

In my dissertation, I first discuss, and defend, the motivations for accepting a functionalist theory of the mind, like David Lewis', that is derived from our common sense generalizations about the characteristic causal roles of mental phenomena in the production of behavior. I differ with Lewis, however, about what it is for a theory to be derived from common sense, and offer an account of how far a functionalist theory may diverge from commonly accepted generalizations without changing the subject.

Next, I evaluate the prospects of a common sense functionalist theory for providing an account of certain intuitively (and explanatory) important properties of mental phenomena, namely, the qualitative properties of sensations, and the representational properties of beliefs and desires. In the recent philosophical literature, for example, it has frequently been argued that the former are more resistant to a functionalist account than the latter. I argue that, on the contrary, beliefs pose as much of a problem as sensations for common sense functionalism, and that, further, the characterization of the content of beliefs provides a problem for any sort of functionalist theory that asserts that beliefs and desires may be defined in terms of their characteristic causal roles.
Acknowledgements

I would like to thank my advisers, Ned Block and Jerry Fodor, for their extremely helpful comments and criticisms of all the drafts of this dissertation. I have also greatly benefitted from their own writings, as well as from their valuable comments on mine. I would also like to thank Norton Batkin and Dan Osherson, who gave me much constructive criticism and encouragement during the early stages of my work. I have benefitted tremendously from numerous conversations with Hartry Field, especially about functionalism and the language of thought, and I have learned a great deal from his writings as well. Thanks also go to Stephen Schiffer for stimulating conversations about the problem of qualia. I would like to give special thanks, however, to Michael Friedman, who, as colleague and friend, provided the support and encouragement that helped me to get started—and finished.
BIOGRAPHICAL NOTE

Born, Pittsburgh, PA on November 4, 1950

Received B.A. from University of Chicago, June, 1972.

Visiting Lecturer, University of Pennsylvania, 1977-1979

Instructor, University of Southern California, 1979-1980
# Table of Contents

Chapter I. Introduction ........................................ 7

0. Opening remarks ............................................ 7
1. Varieties of functionalism .................................... 7
2. Common-sense beliefs about the mind ......................... 11
3. Functionalism and our common-sense beliefs .................. 21

Chapter II. Functionalism and behaviorism ..................... 28

0. Opening remarks ............................................ 28
1. Definitional behaviorism ..................................... 29
2. Radical behaviorism ......................................... 41

Chapter III. Functionalism and the common-sense theory of mind ..... 67

0. Opening remarks ............................................ 67
1. Common-sense theories of mind: Lewis' view .................. 72
2. The meanings and denotations of theoretical terms ............. 90
3. Common-sense theories revised ................................ 101

Chapter IV. Functionalism and the problem of qualia ............. 110

0. Opening remarks ............................................ 110
1. The problem of absent qualia ................................ 113
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Characterizing belief and desire</td>
<td>159</td>
</tr>
<tr>
<td>0</td>
<td>Opening remarks</td>
<td>159</td>
</tr>
<tr>
<td>1</td>
<td>Non-relational theories of belief and desire</td>
<td>161</td>
</tr>
<tr>
<td>2</td>
<td>Relational theories of belief and desire</td>
<td>168</td>
</tr>
<tr>
<td>3</td>
<td>The language of thought: A neural individuation of internal sentences</td>
<td>180</td>
</tr>
<tr>
<td>4</td>
<td>The language of thought: A functional individuation of internal sentences</td>
<td>197</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions</td>
<td>213</td>
</tr>
</tbody>
</table>

Footnotes                                      221

Bibliography                                  231
I. Introduction

0. Opening remarks

Functionalism, in one or another of its many forms, has been claimed by increasing numbers of philosophers to provide an explanatorily and intuitively satisfying account of the nature of mental states. The critics of functionalism, however, and the depth and variety of their criticisms, have increased along with the advocates of the doctrine.

My purpose here is to evaluate both the support and the opposition that this doctrine has recently received. I believe, as do its advocates, that the general functionalist strategy of defining mental states represents an extremely promising approach to a satisfying theory of mind, and I think that there are extant functionalist theories that, with a bit of modification, appear to keep the promise. This dissertation is intended, first, as an argument for that claim. Second, I believe that there has been a general misdiagnosis of the difficulties for functionalist theories, and that the most pressing problems for a functionalist account of mental states arise where they have been least suspected. In the latter half of this dissertation, I argue for this claim as well.

1. Varieties of functionalism

Though there are many varieties of functionalism, one thesis common to them all is the claim that an entity has,
for example, pain, belief, or desire if and only if occurring in it is a state of a type that has certain characteristic relations to sensory stimulations, other mental states, and behavior. That is, if and only if the entity is in a state of a type that has a certain functional role. If some such thesis is correct, we may regard mental state terms--'pain', 'belief', 'desire'--as capable of being functionally defined.

My use of the word 'defined' should not be taken to imply that functional definitions of mental state terms must preserve their ordinary meanings. What counts as the standard of correctness for functional definition is a substantive issue among functionalists which requires discussion.² I use the word instead to signal my interest only in those varieties of functionalism that may be construed as type-identity theses. On some views,³ satisfying a functional definition is a necessary, but not sufficient, condition for a state, event, or property to count as a mental state, event or property of a particular kind. Here, however, I shall restrict my attention to theories that attempt to provide necessary and sufficient conditions, since I am interested in whether a purely functional account can be given of what mental states are.

What, then, will a functional definition look like? In giving a programmatic sketch of such a definition, I shall take the opportunity to introduce some terminology
that will be used throughout the dissertation. Using 'pain' as an example, a functional definition may have the following form:

\[(P) \text{ x is a pain if and only if x is a token of a type of state that bears, in some entity at some time, such-and-such relations to stimulations, other mental states, and behavior.}\]

This sort of definition asserts an identity between the class of pairs and the class of tokens of state-types that, in some entity, at some time, have certain relations to one another, stimulations, and behavior. Or, to use the terminology introduced above, each pain is a token of a state-type that itself has a certain property, namely, a particular functional role.

My primary interest, however, in the chapters to come, will be in the attempt to identify the properties alluded to in (P). Thus, the functional definitions that I shall evaluate will more frequently be of the form:

\[(P') \text{ The property of being a pain } =_{df} \text{ the property of being a token of a state-type that has, in some entity or other, a certain functional role.}\]

If some such definition is correct, then the property of being a pain, or, as I shall sometimes say, the abstract universal pain, will be a functional property of certain particulars--those tokens of state-types that have the appropriate functional role. Functional roles, of course,
are properties of universals, namely, the state types that have those roles. Finally, the abstract universal pain is not to be confused with the (also abstract) state-type which, in an individual, has the functional role in question. I shall call the state-type that has the appropriate functional role in some entity the realization of pain in that entity: it is an important feature of functionalism that pain may be realized, in different entities, by different types of states. 5

Since it was intended as programmatic, my sample definition of pain has virtually no content. How then do we begin to exchange program for substance, and determine the characteristic relations in terms of which 'pain' is to be defined? Again, this is a major issue among functionalists. Many theories of mental phenomena that count as broadly functionalist have little else in common: some are intended to provide meaning analyses of ordinary terms such as 'pain', 'belief', and 'anger', while others are intended as frameworks for theory construction in psychology. Others still fall somewhere in between.

Any of these theories, however, in order to be regarded as the theoretically and intuitively satisfying theory of its philosophical advocates' claims, must meet certain conditions: It must give an account of at least most of those properties that mental states are intuitively believed to have, and must seriously address the questions about mental
states that have been discussed by traditional philosophers. Otherwise, functionalists could be taken to have changed the subject, to have provided an account of a set of phenomena distinct from the sensations, beliefs, and desires that are appealed to in ordinary conversation, and that have been discussed in the philosophical literature since Descartes.

I do not want to suggest that an adequate functionalist theory must solve all the traditional problems of mind, or make all our common sense beliefs about mental states consistent. Some problems may be more easily dissolved, and some commonly held beliefs explained away. A theory of which this is a consequence, however, owes an explanation of why these beliefs ever seemed plausible, or these problems ever taken seriously. What then are these questions and assumptions to which an adequate theory must do justice, and can any functionalist theory possibly succeed?

2. Common sense beliefs about the mind

There seem to be three different kinds of beliefs about mental states which provide rough conditions on the adequacy of a theory of mind. The first may be classed as ontological: It is commonly believed that people indeed have mental states and processes, which, in turn, have properties of certain kinds. For example, common sense would balk at accepting the claim that there were no pains,
and equally at the claim that pains did not feel a certain way. Another requirement for a theory that can be classed as ontological regards its scope: We want a theory that permits the attribution of mental states to all and only those entities we intuitively believe to have them. Common sense would most probably counsel the rejection of a theory that implied, for example, that people in California did not have pain, or that rocks and waterfalls did.

It is also assumed that the appeal to beliefs, sensations, and emotions is of value in explaining a person's behavior and meaningful utterances. An adequate theory, then, should say enough about the relation of mental states and behavior to permit a plausible account of why these explanatory appeals are (or seem) successful. I shall henceforth refer to this assumption about mental states as explanatory.

The third assumption is that there is a great, and perhaps in principle difference, between the knowledge that people can have of their own mental states and the mental states of others; that certain ways of knowing about the existence and properties of mental states are available only to the person whose mental states they are. This seeming epistemological asymmetry, then, is the third phenomenon that must be explained (or explained away) by an adequate theory of mind.
Functionalism has been heralded by its proponents as meeting these requirements far better than either dualism or physicalism, two theories of mind that are frequently regarded as its alternatives. Dualism, broadly speaking, is the thesis that mental phenomena are different in kind from physical phenomena. There are, however, a variety of dualistic views. On the view held by Descartes, mind and body were distinct substances, while other views hold that both mental and physical states and properties are distinct, irreducible states and properties of a single, neutral, kind of stuff.

This view, of course, asserts that there are mental states, and accounts for the special properties that they are believed to have. For Descartes, mental substance was just defined as that substratum of which mental states, with all their commonly acknowledged properties, were modifications. On the other dualistic views, mental properties such as having a throbbing pain, or a yellow after-image (just like weighing 150 pounds, or having arms and legs) were assumed to be perfectly unproblematic properties had by those capable of a mental life.

But dualism, however construed, does not combine well with the explanatory assumptions described above. There are two distinct but related difficulties for the doctrine. First of all, a dualistic explanation of human behavior would require the postulation of entities that are radically,
and in principle, different from those needed to explain
the nature and interrelations of any other phenomena. But
general methodological considerations about what counts as
a good explanation suggest that such theories be avoided
if at all possible. Second, a dualist must give some
details about how it can be that behavioral regularities
are successfully explained by appeal to mental states. An
intuitively obvious answer, and one that was given by
Descartes, is that beliefs, sensations, and emotions cause
people to behave in various ways. But from Descartes on,
dualists have owed an explanation of how mental and physical
entities can causally interact (without, for example, flout­
ing the law of conservation of energy) as it prima facie
seems they must if the appeal to mental states in the
explanation of behavior is taken seriously.

Of course, dualists have the option of attempting to
explain away the semblance of causal interaction between
mental and physical states. There have been two dualistic
strategies for doing so. The first is epiphenomenalism,
a doctrine that states that physical phenomena (such as
environmental stimulations) simultaneously cause other
physical phenomena (the occurrence of events in the nervous
system?) and mental phenomena such as sensations and beliefs
as well. But although mental events are caused by physical
phenomena, they cannot, in turn, cause physical events, and
thus flout the laws of energy conservation. Further, the
causation of mental events by physical events does not provide a counter-example to such generally accepted laws since the physical events that have mental effects have physical effects as well.

Another attempt at meeting the explanatory assumptions is the doctrine of psycho-physical parallelism. This theory attempts to explain away the appearance of mental-physical interaction by stating that mental and physical phenomena are in fact governed by distinct laws, and do not interact; we just think interaction occurs because beliefs, sensations and emotions--entities of the mental realm--just happen to occur after the physical events that we mistakenly identify as their typical causes.

However, these attempts to explain away the appearance of mental-physical interaction compound the conflict between dualism and the first of our explanatory desiderata: Epiphenomenalism goes counter to those general methodological constraints by postulating a different kind of causation, and parallelism, besides giving rise to further questions about what accounts for the parallel, is explanatorily redundant. Further, both these views require the rejection of our deeply entrenched belief that some mental phenomena (e.g., beliefs, desires, intentions) do indeed cause physical phenomena (e.g., behavior). Dualism, then, remains a view that should be resisted if at all possible.
What about physicalism, then? Physicalism, the doctrine that mental states are just certain kinds of physical states, includes a number of different theories as well. The strongest is the psycho-physical type-identity thesis, in which each kind of mental state or property (pain, anger, desire) is held to be identical with a particular kind of physical state or property (for example, brain state B). Or, to put it extensionally, the class of pains is identical with the class of physical states of a certain kind. The weakest view claimed to be physicalistic is known as token-physicalism, or the token identity thesis. Whereas the type-identity thesis requires that each instance of a particular kind of mental state or property be an instance of the same physical state or property, the token-identity thesis requires only that each instance of a mental state or property be identical with some physical phenomenon or other.

The former, type, identity thesis, is often regarded as implausibly strong. If mental states are particular kinds of neural states, then it follows that no entity without a brain sufficiently like ours can be said to have beliefs, desires, sensations, and emotions like ours. This rules out Martians, silicon-based life forms, robots, and other denizens of science fiction as capable of sharing our mental life, though intuitively, we assume this to be, at the very least an open question. Such a consequence, in Ned Block's terms, is excessively chauvinistic. It goes against our
common sense beliefs about the proper scope of a theory of mind. On the other hand, token physicalism may plausibly be regarded as too weak. Without some specification of something physical that all instances of pain or anger have in common, materialism has arguably failed to yield a bona-fide physicalist account of the nature of pain or anger. Indeed, in providing necessary but not sufficient conditions for an entity to be, say, a pain, the token-identity thesis gives no adequate answer to the question of what it is to be a pain: it provides no conditions for identifying pains in all entities at all times. One might argue, on this basis, that the major challenge for physicalism is to come up with some plausible intermediate view that provides a physically acceptable account of the properties that all instances of a particular kind of mental state must share, that is yet sufficiently abstract to avoid the charge of chauvinism.

It is not clear, however, that any version of physicalism will be compatible with certain epistemic assumptions that I characterized earlier, as underlying an important problem area in traditional writings about the mind. The two epistemic assumptions I mentioned were first, that people can know about the existence and properties of their own mental states better than they can know about the existence and properties of other events and objects in the world, and second, that they can know about their own mental states better than others could know about them. Indeed,
the second assumption is often strengthened to the claim that, in principle, no possible evidence that an agent had a particular mental state (or didn't have it) could override the agent's sincere claim that he didn't have that mental state (or did). But are such assumptions compatible with any physicalistic thesis? The difficulty here is that if headaches, for example, were simply physical phenomena like head injuries, then our relative abilities to know whether you have a headache should be similar, in principle, to our relative abilities to know whether you have a head injury. But then my claims about the properties of your headache might well, if I were an expert, be taken to outweigh yours.

Physicalists, however, have claimed that these epistemological beliefs about mental states can be partially accounted for, and partially explained away. According to philosophers such as Armstrong and Smart, our access to our own mental states may be merely privileged, rather than in principle private. This would not rule out the identity of mental with physical states, however, since we unproblematically regard our access to various of our own bodily processes and changes as privileged as well. Further, the weight given to first-person testimony about one's own mental states may be merely an acknowledgement of this privilege. This, however, is not unique to the case of mental states; a person's own testimony is regarded as close
to definitive not only on certain bodily matters, but on other routine information such as name, address, and birth-date.

Whether or not this response is ultimately satisfactory, it remains more satisfactory than the dualist's attempt to explain away the semblance of mental-physical interaction; unlike the dualist's explanations, the physicalist's account of our epistemological beliefs about mental states does not run counter to any other commonly held beliefs about the mind.

However, there are yet other difficulties for physicalism: It has been argued that no version of physicalism can account for the distinctive properties that sensations, beliefs, and emotions are commonly thought to have. Suppose that pains and after-images are indeed physical states of some sort. Nevertheless, it is argued, physicalism need not be true. The problem is that mental states, especially sensations, are described as having certain phenomenal properties—hurting, feeling warm, looking yellow—which, do not appear to be reducible to physical properties.

A promising move for a physicalist, however, is to claim that there is an alternative way of describing these sensations which, unlike our ordinary, phenomenal vocabulary, does not imply that they must have other than physical properties. If this is so, then there would be no logical impediment to a physicalist theory of sensations. This was
Smart's strategy in "Sensations and Brain Processes", in which he suggested that statements such as "I am now having a yellow after-image" be given what he called "topic-neural translations" of the form, "Something is going on in me now like what goes on when I am attentively looking at a lemon." Smart, unlike some physicalists who advocate such redescription, requires that his translations preserve meaning. But even if we require only that Smart's new predicates be true of individuals when and only when the phenomenal predicates were true of them, Smart's "translations" are inadequate. The "like" locution is the source of the difficulty; it is too vague. After all, what occurs in me when I am looking at an egg, or a football, is in some sense like what occurs in me when I am looking at a lemon. But that the difficulty resided in the vagueness of "like" is disheartening, as it suggests that the internal occurrences must be compared as to likeness in various respects, and the prospects of doing so without reverting to phenomenal descriptions were thought by many to be quite dim.

Prima facie, then, we have a dilemma. In order for there to be beliefs, desires, sensations, and emotions, it must be said what they are. That is, either dualism or physicalism must be accepted. But the acceptance of dualism makes it hard to account for the explanatory features of our commonly held view about mental phenomena, while the acceptance of physicalism makes it difficult to give an account of the distinctive, perhaps essential properties of
mental states, and of our special relations to our own. Thus it appears that the common sense assumptions about mental states may not all be compatible. Indeed, some may regard these remarks as leading to a reductio of the claim that there are beliefs, sensations, desires, and emotions. 9

3. Functionalism and our common sense beliefs

A number of philosophers, however, have diagnosed 10 this seeming incompatibility as arising from an overly restrictive view of the ontological alternatives. According to these views, the acceptance of functionalism will permit us to say that people have sensations, beliefs, desires, and emotions, without committing ourselves either to dualism or to physicalism, nor thus in turn to their undesirable consequences.

After all, consider the example of a functional definition of 'pain' that I presented earlier. According to (P') pain is a property that state-tokens have solely in virtue of being tokens of states that have a certain functional role. The states that have this functional role in different entities however, may well be different physical types. Indeed, not only could widely disparate types of physical states realize pain in different entities, but pain could be realized, in principle, by non-physical states as well. Of course, if non-physical states could not, as a matter of physical law, assume the interrelations that are definitive
of mental phenomena, then the acceptance of functionalism would in fact preclude a dualist ontology. But the nomological impossibility of such interrelations would take further argument; it is not a consequence of functionalism itself. According to functionalism, what is essential to something's being a belief, sensation, or emotion are certain of its functional properties; its other properties, including its internal composition, are purely accidental. Functionalism, then, can be said to assert that the question about the nature of mental states has little to do with the stuff of which they consist.

This claim about functionalism, to be sure, is strictly speaking true, but I believe that it masks some advantages that the doctrine has for those who are sympathetic to physicalism, broadly construed. First of all, functionalism can be viewed, (and has been viewed by, for example, David Lewis, in "An Argument for the Identity Thesis"), as a doctrine that promises to yield the topic-neutral redescriptions of mental states and properties desired by Smart. (I shall give details as to how functional theories purport to do this in Chapter III.) Further, these topic-neutral functional definitions are promising candidates for the solution of what I labelled earlier as a major challenge for physicalism: I claimed earlier that the psycho-physical type-identity thesis requires that we withhold attributions of mental states from certain entities that, intuitively, seem
to have them, while a token-identity thesis does not give a physically acceptable account of what it is to be a mental state of a certain kind. Functionalism, however, is a doctrine that, along with certain other assumptions, could provide an intuitively plausible theory of mental states that would be acceptable to those who want to deny that either pains themselves, or the property of being a pain, are irreducibly mental. Functional definitions, after all, provide a topic-neutral property that all pains must have in common, namely, the property of being tokens of states with a certain functional role. If it can be shown that only physical state-types can have this functional role, then functionalism, plus this assumption, yields a theory that will be acceptable to those with the physicalistic sympathies I described: All pains will be tokens of physical states and they will be pains by virtue of having some physical property which, in turn, has a topic-neutral property--a certain functional role. Thus the property of being a pain will be as ontologically benign as the property of occurring at time t, given its topic-neutrality plus the further assumption that all events are physical events.

My own interest in functionalism has its source in the hope that the doctrine can indeed help to provide a physically acceptable account of mental states and properties. However, I shall not address the issue, in this
dissertation, of whether the hopes of the physicalist that I described above are likely to be fulfilled. My purpose here is to determine whether functional definitions of mental states (whether or not they admit of non-physical realization) do justice to our common sense assumptions and beliefs about the mind.

Functionalist theories assert that there are mental states and processes, and they seem to account for our general commonly held belief that mental phenomena have efficacy in the production of behavior. Indeed, beliefs, sensations and emotions are to be defined in terms of their characteristic, behavior-producing interactions. However, I have not yet given the details of any specific functionalist theory; it may turn out that the specific roles of these mental states in various functionalist theories are not what we commonly believe the roles of beliefs, sensations, and emotions to be. If so, then despite appearances, functionalism will be found not to honor the common explanatory assumptions about the mind. I address this issue in more detail in Chapter III.

What about the epistemic assumptions? Functionalism, even along with the assumptions that make it a physicalistically acceptable theory, fares at least no worse than the identity thesis in accounting for the common beliefs about our epistemic relations to our own mental states. The strategy of the physicalist, as I mentioned earlier, is to deny the truth
of the stronger claim that we have in principle private
access to our own mental states, but concur that we have
de facto privileged access. According to the physicalist,
however, the truth of the weaker claim vindicates our common
sense beliefs about the special evidential nature of our
sincere claims about our own mental states. But the truth
of the weaker claim is compatible with physicalism, and thus
with (physicalistic) functionalism.

Some functionalists, however, have argued that the
doctrine can account for the stronger epistemic assumptions
as well. Since mental states are to be defined in terms of
their interrelations, a state such as pain could be (par-
tially) defined as that state which invariably and
exclusively gives rise to, among other things, the belief
that one is in a state of a certain kind. If this is so,
then functionalism would be compatible with a very strong
"in principle" claim about a person's privileged access to
his own mental states: indeed, functionalism would be
compatible with the thesis that the belief that one is in
pain is necessary and sufficient for being in pain. This
would account for the overriding evidential force of a
person's sincere reports of the occurrence of pain.

It is not clear, however, that functionalists need
go to such lengths, since many philosophers have suggested
that the strong privileged access thesis is suspect for
reasons independent of its alleged incompatibility with
physicalistic views. In the recent philosophical literature, there have appeared criticisms (made by, among others, Quine, Sellars, Reichenbach, Wittgenstein, and Goodman) of all views that claim that certain kinds of statements are immune to revision, or correction in light of certain evidential or theoretical considerations. Thus the assumption that the sincere statements one makes about one's own mental states must withstand all counterevidence would be undermined by the truth of this general view.

I shall not discuss the relative merits of these responses, but either, I believe, lends support to a major working assumption of this dissertation, namely, that if functionalism appears to be incompatible only with the stronger versions of what I called our common epistemic assumptions, then the doctrine is worth taking very seriously as a theoretically and intuitively satisfying theory of mental states.

Unfortunately, however, there are thought to be other difficulties for the doctrine. Indeed, most of the criticisms of functionalism have not concerned its alleged incompatibility with our epistemic assumptions; rather, critics have claimed that functional definitions of mental states neither capture the important properties of sensations, beliefs, and emotions, nor provide a theory of mental states that, intuitively, has the proper scope. In Chapters
IV and V I attempt to defend functionalism against these charges.

However, my claim that the strong epistemic assumptions about mental states can be, in part, ignored because they are subject to internal, independent difficulties has a parallel which may be disturbing. It has been argued, likewise, that our common sense explanatory assumptions about mental states are benighted, and that there are other interpretations of the apparent success of our mentalistic explanations of behavior. Some have argued that the common view that mental states are causally efficacious in producing behavior was conceptually confused, others, that it was empirically misguided. Either consequence, of course, would be extremely damaging to any functionalist theory. Thus I shall devote the next chapter to an appraisal of the doctrines that, respectively, involve such claims, namely, definitional and radical behaviorism.
II. Functionalism and Behaviorism.

0. Opening Remarks

As suggested before, behaviorism is a doctrine which asserts that human behavior can be explained without appeal to the causal efficacy of beliefs, sensations, and emotions. As with functionalism, however, there are many varieties of behaviorism, each with different reasons for rejecting the explanatory appeal to mental states.

My major concern is with two kinds of behavioristic views. The first is definitional behaviorism (sometimes called "philosophical" or "logical" behaviorism), which was first and most prominently advanced by Ryle in The Concept of Mind. The second is radical behaviorism, a view that runs throughout the work of Skinner and is given a name in his About Behaviorism. I focus upon these views, first, because the reasons each offers for rejecting mentalistic explanations were at one time widely accepted, and should be seriously discussed. Second, definitional and radical behaviorism can be seen as the antecedents of two different approaches to functionalism, each of which accepts the appeal to mental states in the explanation of behavior for much the same reason for which it was rejected by the relevant behavioristic theory. Thus an examination of these reasons will help to make clear just what is at issue among the different functionalist theories that will be appraised in Chapter III.
1. **Definitional behaviorism**

Definitional behaviorism was classically presented not as an account of the nature of mental states and properties, but rather as an account of the meanings of **mentalistic terms**. According to Ryle, the **proper** meanings of terms such as 'belief', 'sensation' and 'emotion' do in fact involve the explanation of why people do what they do. But it is not the case, Ryle argues, that all legitimate answers to questions such as these must involve the citation of a **cause**. Indeed, Ryle claims that the assumption that the appeal to mental phenomena in the explanation of behavior is a species of **causal** explanation is a mistake that has been made by nearly every philosopher since Descartes. The mistake has important consequences; it, plus the assumption that mental phenomena were not observable in the same ways as physical phenomena, leads inexorably to Cartesian dualism. That is, we assume that there are causally efficacious states and processes that are responsible for behavior. They can't be **physical** states and events, since they are not observable in the same ways. Thus they are entities of a different nature altogether. The crucial mistake, according to Ryle, is in our assumption that mental phenomena **cause** behavior. This assumption, he says, involves a special kind of mistake--a **category mistake**.
It has never been quite clear what Ryle regards to be definitive of a category-mistake. He states, in *The Concept of Mind*, that a sentence involves a category-mistake if it includes the conjunction or simultaneous predication of terms from **different logical categories**. Some examples of these sentences are "She came home in a taxi and a flood of tears," and (Ryle's) "She lowered her standards by raising her glass, her courage, her eyes, and his hopes." Unfortunately, this merely pushes the question back one step to the question of what is a logical category.

An attempted answer to this question has often involved the following criterion: Two words belong to the same **logical category** if and only if one can be substituted for the other in all sentences, preserving the logicality, or meaningfulness, of the sentence. Again, this merely staves off the question, since the distinction between meaningful and meaningless sentences is in want of clarification itself. There have been some recent attempts to rescue the notion of a category-mistake from technical difficulties (and some conceptual difficulties) in its formulation which are independent of the questions about a criterion of meaningfulness. Here, however, I want merely to indicate what Ryle thought the **category** that included mentalistic terms such as 'belief', 'sensation', and 'emotion' to be.

According to Ryle, such terms denoted neither states, nor events, nor properties, but rather **behavioral dispositions**.
And, given Ryle's views about dispositions, a term that denotes a behavioral disposition functions not to denote an enduring property of a person, but merely as an **abbreviation** for a set of hypothetical statements about how a person would behave given various environmental situations. Thus the **meaning** of a sentence that included mentalistic terms could be captured and exhausted by the list of hypotheticals of this sort. That is, each sentence could be given a **translation** into a statement or conjunction of statements which made reference only to environmental stimulations and behavior.

Thus the view is **behavioristic**, since the causes of human behavior were to be stated with reference to stimulations and behavior alone; the view is **definitional**, in that these sentences that involve reference solely to stimulations and behavior are alleged to **preserve the meanings** of the mentalistic sentences of which they are translations.

For example, a sentence such as "He is afraid of the water" might include in its Rylean translation the following sentences: "When swimming is suggested, he shudders", and "When he goes to the beach, he stays away from the water." The sentence, "She wants to catch the bus" might be partially rendered as "When she sees the bus coming, her pace quickens."

Now what is the motivation for believing that translations such as these could possibly preserve the meanings of
(or even give acceptable new meanings to) the mentalistic sentences of which they are translations? After all, terms such as 'pain', 'anger', and 'desire' are commonly used as terms that purport to denote entities distinct from, and causally efficacious in, the production of behavior. Ryle, however, does not merely overlook this feature of our ordinary use of mentalistic terms, but explicitly rejects it. So, we may ask, why does he claim to be giving an analysis of our ordinary mentalistic concepts, rather than merely changing the subject?

Ryle's position, I believe, could be defended as follows: According to our traditional views about the mind, appeals to mental phenomena are ordinarily made in the explanation of behavior. This is the source of the "explanatory" assumptions discussed in the previous section. The "ontological" assumptions—the assumptions that mental phenomena were distinct from behavior—are not independently motivated, but follow from an overly restrictive view of explanation. Although Descartes, among others, thought that an appeal to mental phenomena in the explanation of behavior counted automatically as an appeal to their efficacy in the production of behavior, this was not a correct analysis of the explanatory role of mentalistic terms. It stemmed, in fact, from the category mistake of treating intelligent behavior as the effect of various special processes rather
than, in Ryle's view correctly, as the manifestation of various dispositions to behave. This treatment of mental phenomena, according to Ryle, has had precedent:

Sentences embodying these dispositional words have been interpreted as being categorical reports of particular but unwitnessable matters of fact instead of being testable, open hypothetical and what I shall call 'semi-hypothetical' statements. The old error of treating the term 'Force' as denoting an occult force-exerting agency has been given up in the physical sciences, but its relatives survive in many theories of mind. ... (p. 117)

But there is equal precedent, as suggested in this passage, for abandoning this treatment of mentalistic terms. Thus Ryle may be taken as offering another interpretation of the explanatory role of our ordinary appeals to mental phenomena, namely, that they be construed as shorthand for the appeal to behavioral dispositions.

In short, what Ryle could claim to be rejecting is not the traditional explanatory roles of our mentalistic terms, but rather a confused and mistaken theory of what these terms denote. This is consistent with his claim, in the introduction to The Concept of Mind, to be "rectifying the logical geography of the concepts we already possess." But although we may in this way justify Ryle's attempt to distinguish the traditional use of mentalistic terms from the traditional theories about what they denote, it is far from clear that the attempt is successful. Thus, let us examine
Ryle's attempts to translate our ordinary mentalistic statements into equivalents that refer only to behavioral dispositions, and determine whether such sentences could in fact fulfill the purposes for which they were traditionally desired. That is, is it indeed possible to give purely behavioral translations, or for that matter purely behavioral truth-conditions, for statements that include terms such as 'belief', 'sensation', and 'emotion'?

Since there has been extensive discussion of definitional behaviorism in the philosophical literature (see, for example, Fodor, Putnam, Chisolm, Cornman, Lewis, and Geach), I shall not restate in detail the well-known criticisms of the program here. However, I would like to record the major points of issue with Ryle's proposals.

As has been pointed out frequently in the literature, it is unlikely that Ryle's translations preserve even the desired explanatory features of our ordinary mentalistic statements. For example, consider the statement that Sarah believes it is raining. How could this be translated into a statement that makes no reference to mental phenomena, but solely to stimulations and behavior? Consider, as a first attempt, the following:

(R) Sarah believes it's raining if and only if she takes an umbrella when she leaves for work.

Clearly, the statement on the right of the biconditional would not yield the implications that we desire statements
about beliefs to have. The statement on the right, plus "Sarah leaves for work" implies "Sarah takes an umbrella", while we would not expect the statement on the left plus "Sarah leaves for work" to have any such implication. Thus the statement on the right does not provide a necessary condition for Sarah to believe that it is raining. Nor will it provide a sufficient condition, since it is implied by statements that we would not think implied the correlative statement about belief. For example, "Sarah is about to return her co-worker's umbrella to him at the office" will imply that Sarah carries an umbrella to work, but not that she believes that it is raining. (The anomaly remains even when we replace talk of implication with talk of empirical truth-conditions or licensing of prediction). It is argued further, by the authors mentioned above, that these difficulties are not merely a product of the crudeness of our example, but are in principle difficulties with the definitional behaviorists' approach. That is, it is argued that no finite, adequately stated disjunction of statements about Sarah's behavior, or dispositions to behave in various circumstances, would provide a set of either necessary or sufficient conditions for Sarah to believe that it is raining. Briefly, the counter-examples to the necessity of these disjunctions involve our imagining conditions under which a person is, intuitively, in a particular mental state--say, a pain, or a state of desire--without having the
behavioral dispositions we associate with the occurrence of that state. Sometimes the examples involve further supposition that the beings in question eventually manifest a completely disjoint set of behaviors when in that state.

The counterexamples to the sufficiency of the disjunctions appeal to the intuition that a person could indeed manifest all the behaviors typically associated with a particular mental state, and nevertheless not be in that mental state. Radical examples of such situations involve our imagining a well-trained actor to be simulating pain in appropriate circumstances, for a considerable time, again, perhaps a lifetime.

Yet the behaviorist may claim that it is the very radicality of these examples that casts suspicion upon them. That is, it may be taken to suggest that the definitional behaviorist might merely need a clause in his behavioral definitions that insures that normal conditions obtain. The addition of this simple *caeteris paribus* clause might be hoped to insure the necessity and sufficiency of the behavioral conditions in all but these special circumstances; indeed, the specification of the behavioral conditions plus the built in escape clause designed for these special cases would yield behavioral definitions of our mentalistic terms outright.

But, as the critics of definitional behaviorism point out, it looks as if the circumstances to be covered by a
cateris paribus proviso are not so special. The very feature that makes the radical counterexamples to the behavioral definitions plausible are importantly similar to those operative in the simpler cases, and, in fact, similar to those that guarantee the plausibility of the behaviorists' more successful attempts. In Putnam's case of the Super-Spartans, we were to imagine a race of people whose dispositions to manifest pain behavior were overridden by the greater desire to act as if impervious to pain. Similarly the case of the actor involves, at least implicitly, the suggestion that there is some motivation for his continuing to act as if in pain. That is, the stories about these behavioral anomalies become plausible when spelled out enough to yield reasons for the anomalies, and these reasons, invariably, involve the reference to other mental states of the agent in question.

But so it does in circumstances less extreme as well. In the case of Sarah's belief that it will rain, it looks as if any cluster of conditionals about carrying umbrellas or looking at the sky can be overridden by an alternative explanation of those behaviors by reference to different motivations and desires. And most important, it does so even in the circumstances in which the behaviorist' definitions seem to work best; that is, we imagine, in these cases, that the behavior in question is in fact produced by the mental states by which such behavior is normally
produced. That is why it looks as if no reference to anything other than environmental stimulations and behavioral events must be made in these cases. But the relative success of these definitions, and the need for a caeteris paribus clause in the other situations, can be explained by the same phenomenon, namely, that the description of 'normal' situations in which behavior occurs requires implicit, but ineliminable, reference to the mental states of the agent. In short, definition is not possible without appeal to a caeteris paribus clause that mentions normal conditions, but such conditions cannot be specified without reference to mental phenomena. Thus, the reference to mental phenomena, instead of being eliminated, merely remains hidden. Thus why not look for characterizations of mental states that can bring these "hidden variables" to light?

Defenders of certain kinds of functionalist theories claim that their account of mental phenomena does just this. Yet functionalism, when regarded as a theory of meaning for mentalistic terms, preserves an important insight of the definitional behaviorists, namely the view that the attribution of mental phenomena to an organism is conceptually connected, in some sense, to the specifications of the organism's dispositions to behave. Moreover, it preserves the definitional behaviorists' conditions that no mentalistic vocabulary appear in the translations of our ordinary
statements about mental phenomena. At the same time, however, this view has the added advantage of preserving the ordinary intuition that mental phenomena are distinct from, yet efficacious in the production of, behavior. 8

Functionalist views meet all these desiderata by not insisting that the necessary and sufficient conditions for the occurrence of any particular mental phenomenon be given in purely behavioral terms. Rather, each phenomenon will be defined by its effects upon behavior given various environmental stimulations and the presence of various other mental states. Whereas the definitional behaviorists attempted to define mental phenomena singly as dispositions to behave in certain ways given certain situations, functionalists permit the simultaneous definition of mental phenomena as those states and processes that together produce behavior of various sorts. In a functionalist theory then, will occur statements such as "(For each individual) there are states and processes that interact to produce such-and-such behaviors..." Particular mental states, such as pain, or desire, will be defined as "the state such that (in that individual) there are other states and processes with which it interacts to produce such-and-such behavior." 9 Thus in the translation of sentences that include particular mentalistic terms such as 'pain' or 'desire', will appear quantification over, though no mention of, other mental
phenomena. So although the lifting of the restriction that each mental state be given a purely behavioral definition precludes the reduction of mental phenomena to behavioral dispositions, it does not preclude the reduction of mental phenomena to otherwise characterizable states (or dispositions) of a system. The rejection of definitional behaviorism does not involve the recognition of irreducible mental phenomena, but merely phenomena not reducible to behavioral dispositions.

Thus functionalism preserves the intuition that we could not ordinarily attribute mental states and properties to persons unless we were committed to certain claims about their past, present, or future behavior. Yet it does not require that mental phenomena be singly definable in terms of stimulations and behavior. Moreover, it preserves the intuition that they are distinct from behavioral events, and are causally efficacious in their production, an intuition central to our ordinary conception of the nature of mental states.

In this section I have argued that a theory that permits the quantification over other mental phenomena has a better chance than definitional behaviorism of providing definitions that preserve the ordinary explanatory functions of our mentalistic terms, while according with other common beliefs about mental states. However, I have not yet answered the
question of whether such a "definitional functionalism" will be good enough. This issue will be discussed in Chapter III.

Further, I have not yet shown that no behavioristic theory can provide an adequate explanation of behavior. That is, perhaps it was not the behaviorism of Ryle's view that provided the problems, but merely the requirement that the behavioristic "translations" of our mentalistic terms preserve their ordinary meanings. In turning to a behavioristic theory--Skinner's--that obeys no such requirement, I shall be interested both in its content as an explanation of behavior, and in the justification it provides for rejecting our common sense mentalistic explanations wholesale.

2. Radical Behaviorism

With respect to our ordinary mentalistic vocabulary, radical behaviorism (at least as propounded by Skinner) advocates not meaning-preserving translation, but theoretical replacement. That is, this view calls for the replacement of the "mentalistic theory," comprised of our ordinary, unreflective appeals to mental phenomena in the explanation of behavior, with behavioristic theory. It does not require the sentence-by-sentence translation of the so-called mentalistic theory into statements which make reference only to behavioral dispositions and patterns of
reinforcement; in fact, it does not even require the provision of behavioristic truth-conditions for ordinary statements about beliefs, desires, and pains. Rather, an entire theory, or method of explanation (conceptual scheme, if you will) is to be dropped in favor of another which is alleged to have met with greater empirical success. Skinner allows, and in fact desires it to be the case, that most mentalistic sentences be given behavioral truth-conditions which he sometimes even calls "translations"; he has not taken arguments for the irreducibility of the intentional idiom to heart. Those sentences for which translations cannot be provided, however, do not count as counterexamples to a radical behavioristic view; these sentences, Skinner insists, will merely be dropped from use after the adoption of the newer, empirically successful theory.

It is fairly clear then that the arguments for the acceptance of radical behaviorism depend primarily on the relative empirical success of the theory. Thus let us quickly sketch enough of the theory we need in order to evaluate its chances for empirical success.

According to Skinner, the occurrence of behavior in a certain situation can be explained solely by reference to the behaving organism's history of reinforcement and the genetically determined traits peculiar to its species. The latter properties are widely regarded to be purely biological; thus no reference to mental phenomena need be made in this
part of the explanation. But Skinner's further, more substantive claim is that the former part of the explanation—the appeal to an organism's history of reinforcement—may be stated without appeal to any mental phenomena as well. This is the issue upon which most criticisms of behaviorism as a psychological theory are focused, namely, can a theory which permits no reference to mental phenomena irreducible to behavioral dispositions yield laws which correctly predict and explain the observed behavioral regularities?

Note that both Skinner's goals and the criticisms of his behavioristic theory are different from the goals and criticisms of the definitional behaviorists. As mentioned before, criticisms directed at his failure to provide translations, or even purely behavioral truth-conditions, for each ordinary mentalistic statement would miss the point: this is not his purpose. Rather, the determination that such a theory meets with predictive failure, or seems unlikely to arrive at behavioral laws would be regarded as definitive criticism of the view. The determination that functionalism, with its laws that include quantifiers which range over non-behavioral, internal events of a system, provides a better codification and explanation of observed behavioral regularities would count as a relative vindication of functionalism. Thus this will be my strategy here. But since Skinnerian behaviorism is not well enough developed to really
test for predictive success, and since there are no well-developed theories to which it could be compared, I will try to indicate, instead, some of the internal difficulties that beset Skinner's attempt to provide a framework for an explanation of behavior free of reference to any internal, non-behavioral events.

It may be claimed that I put Skinner's position here too strongly, since he makes a point, especially in his later writings, of stating that a behaviorist may recognize the existence of "private" events such as sensations and feelings which are not identifiable with what Skinner calls "covert" behavior. (Covert behavior, according to Skinner, includes events which use the same nerve and muscle structures and conform to the same laws as overt behavior, but are performed on a scale too small to be observed by others.) It is nevertheless the case that on Skinner's view neither reference to nor quantification over such private events may appear in legitimate behavioral laws; according to Skinner, these events are, rather, the "collateral" effects of behavior, i.e., they have no role in its production.

Here Skinner's view is in consort with the program of methodological behaviorism, a program that he rejects as empirically and philosophically unsophisticated. The methodological behaviorists, like Skinner, recognized the existence of private events, but believed that the proper
task of psychologists was to look for functional relations among behavioral and environmental events alone, and not to include these private events in their determinations. Also like Skinner, they believed that it was unnecessary to give a quantitative account of their occurrence. Unlike Skinner, however, they believed that these private events did have efficacy in the production of behavior; they were considered to be causally efficacious variables mediating stimulations and behavior. For the methodological behaviorists, it was impossible and unnecessary to give a more quantitative account of the occurrence of "private" events: impossible because the events were indeed private, and could not be intersubjectively observed; unnecessary because of the transitivity of cause and functional relation: if A causes B and B causes C, then we may legitimately claim that A is causally related to C.

Skinner criticizes the methodological behaviorists, however, for their unreflective acceptance of the methodology of the logical positivists. He says (AB P·16), "Methodological behaviorism and some versions of logical positivism ruled private events out of bounds because there could be no public agreement about their validity." This is incorrect, according to Skinner; private events that make a contribution to the production of behavior should not be ruled out of consideration for methodological reasons alone;
reports of events that are not intersubjectively observable may be questioned as to their reliability, but their status as reports with at least some degree of empirical significance need not be denied. Skinner continues:

[Radical behaviorism] does not insist upon truth by agreement and can therefore consider events taking place in the private world within the skin. It does not call these events unobservable, and it does not dismiss them as subjective. It simply questions the nature of the object observed and the reliability of the observations. [But this does not mean] that what are felt or introspectively observed are the causes of behavior.

Thus, though Skinner claims to believe that it is not impossible to give a more quantitative account of "private" events, he argues, as do the methodological behaviorists, that it is unnecessary to do so. His is not an argument from the transitivity of causal or functional relation, however, but rather depends upon the simple claim that these private events are in fact not efficacious in the production of behavior.

Interestingly enough, Skinner gives no independent argument, methodological or otherwise, for this latter claim. Its plausibility, he suggests, derives solely from a positive appraisal of the empirical results of a behavioristic theory. If the theory can account for the behavioral regularities, considerations of simplicity, etc., it will force the acceptance of the claim that the private events are not efficacious. Thus the legitimacy of the methodology
of behaviorism may be justified only by the success of the theories produced by that methodology. So Skinner is a more thoroughgoing empiricist than the methodological behaviorists, even though the methodology of the latter group was proposed in the interests of making the study of the causes of behavior more scientific. Hence, a priori arguments about the inappropriateness of the behaviorist methodology will not be arguments against Skinner, unless the inappropriateness of the methodology is proved by a failure of empirical results. Thus, as I mentioned before, the justification of the acceptance of functionalism over behaviorism depends upon a demonstration that functionalism has the better chance of empirical success. I have spent time contrasting the goals of methodological behaviorism with those that Skinner alleges to be his goals because I believe that most of Skinner's difficulties arise from his not taking his own criticisms of methodological behaviorism to heart. Another way of describing Skinner's difficulties is to say, as Fodor does in Psychological Explanation, that Skinner regards behaviorism to be the only alternative to dualism, which he believes to be no explanation at all. But Skinner's classification of mentalistic explanations with dualistic explanations arises from his reluctance to ascribe causal efficacy to events that are unobservable and thus not experimentally manipulable. And this reluctance links him quite closely
with the methodological behaviorism that he professes to reject. This becomes clear, however, on examination of Skinner's view.

Skinner believes that the occurrence of any behavior on the part of an organism can be explained by appeal to the Law of Effect. The Law of Effect (or at least one statement of it) states that any (type of) behavior of an organism that gets reinforced will recur with greater frequency among the subsequent behaviors of that organism. (This version of the Law of Effect does not mention particular causal links between behaviors and events of reinforcement; it is a statistical law stated in terms of the frequency of events.)

The criticisms of the attempt to explain behavior by appeal to this law can be divided into two types. First are the criticisms that regard the empirical significance of the principle: Does the Law of Effect have substantive empirical force, or is it true merely because its crucial constituent terms are interdefined? The second kind of criticism regards the projectibility of the limited, but significant experimental success of the behaviorists. Does the laboratory success of certain operant experiments depend upon the in practice (but not in principle) control of the mental states of the participating organisms? That is, even if the Law of Effect does have empirical content, will it yield lawlike connections between the environment and
the organism's behavior, or must the true "laws of behavior" include reference to non-behavioral states and events as well? That is, is the Law of Effect a law? These criticisms are related, in that an adequate answer to either will depend upon the possibility of giving independent characterizations of various experimental parameters without the use of intentional concepts--without reference to the goals, purposes, or mental states of the behaving organism.

Given, then, that a stimulus, a type of behavior, and a reinforcer are to be characterized by some common non-intentional feature shared by every instance of the type, must the Law of Effect, as stated above, be either circular or false? Fodor, Chomsky, and Dennett, among others, have argued that this must be the case. First is the difficulty of giving an independent specification of the class of reinforcers. Skinner admits that there can be no general characterization of a reinforcing stimulus--no necessary and sufficient conditions for something to count as a reinforcer, even for a particular type of organism. At one time, an organism will "work for food"; at another, not food, but exercise.

As Skinner says in *Science and Human Behavior*, (p. 72), "The only defining characteristic of a reinforcing stimulus is that it reinforces. [And] The only way to tell whether or not a given event is reinforcing to a given organism under certain conditions is to make a direct test." But
if a particular event cannot be specified as reinforcing independent of the situation in which it is so defined by virtue of the fact that certain behaviors that it follows recur, then the so-called law, "The presence of a reinforcer tends to increase similar behavior" will not be falsifiable: How could a reinforcer not tend to increase behavior since reinforcers are defined as just those situations that do?

Skinner responds, however, that the use of the Law of Effect to specify what reinforces a particular organism at a particular time need not entail that the concept of a reinforcer is empirically vacuous. What is required, instead, is a piecemeal, bootstrapping, assignment of values to the experimental parameters in question. For example, from common sense, evolutionary considerations, and introspection on the part of the experimenter, good guesses may be made about what will be reinforcing to most organisms under certain easily specified conditions; obvious guesses may include food, sex, water, and exercise. In experiments in which these are assumed to be reinforcers, whatever behaviors in fact increase in frequency will be counted as instances of the same behavioral type. Then, in other experimental situations, other events or phenomena may be tested for their reinforcing properties by having their occurrence made contingent upon the performance of some instance of the already identified behavioral type, and
determining whether instances of that type increase in frequency. That is, information gathered by the use of the Law of Effect in one experimental situation may be used to make "trans-situational" predictions.

The feasibility of trans-situational prediction, however, depends upon two further conditions. As Skinner points out in the passage quoted above, "the only way to tell whether or not a given event is reinforcing to a given organism is to make a direct test." But this requires the ability to specify whether or not the organism is indeed being tested under the relevant conditions, and whether the test succeeds or fails. It is not obvious, however, that the behaviorist permits himself the use of a vocabulary rich enough to make such specifications.11

The first difficulty involves the specification of the response class, or behavioral type (or in Skinner's vocabulary, the "operant"). As noted earlier, each behavioral type must be specifiable independently of the experimental situation in which it is initially identified by use of the Law of Effect. Otherwise it cannot serve as evidence for identifying reinforcers in new situations. And, in keeping with the behaviorists' program, the behaviors must be described as bodily motions of such and such physical type.

The question to the behaviorist, then, may be put this way: Is there a unique physical property shared by
all and only all those behaviors that increase in frequency when followed by a reinforcer in a particular experimental situation? This question, in turn, may be broken down into two: namely: Is there one such physical property, and is there more than one? Mentalists' criticisms of Skinner tend to address the **first** question, and Skinner tends to respond, in defense, to the **second**. This, of course, contributes to misunderstanding and polemic, and is the result, I think, of the extremely bad examples chosen by the critics of behaviorism to illustrate Skinner's difficulties.

For example, Dennett, in "Intentional Systems" and Fodor, in *Psychological Explanation*, give examples designed to show that behavioral types cannot be specified in physicalistic language, but only with reference to the goals and desires of the organism. Consider, they say, an experimental situation in which a rat is reinforced by food pellets for pressing a metal bar when a light flashes green. According to the behaviorists, the behavioral type that gets reinforced must be a type of **bodily motion** on the part of the rat. But then one should expect a rat that is so reinforced to continue to produce the same pawing motions even if the bar were moved to a different position in the cage. But they go on to say, this prediction is experimentally falsified; rats reinforced for pressing a bar continue to **press the bar** after it is moved. Thus the "behavior" that was in fact reinforced must be described,
if explanatory adequacy is at issue, as the rat's trying to depress the bar. But this description includes reference to the purposes of the rat, and is therefore behavioristically illegitimate.

Dennett, in "Intentional Systems" expresses equal pessimism about the prospect of giving a non-intentional description of the class of stimuli to which a reinforced response will generalize. After all, he speculates, a pigeon trained to peck a key upon the presentation of two horizontal lines of equal length will not respond when presented with the Muller-Lyon illusion. Again, the proper description of the class of discriminative stimuli must make reference to lines that the pigeon perceives to be of equal length, a description not compatible with the behaviorists' methodological restrictions.

Skinner, of course, could produce an easy answer to these objections, straightforwardly extracted from *Science and Human Behavior*. In the first case, what the rat gets reinforced for is not the performance of a particular pawing motion, but rather, for pressing the bar. This type of behavior may be given a characterization that is somewhat more abstract than the former, but is nevertheless physical, namely, as contact of the rat's paw (or body) with the bar (or piece of metal in the cage). Similarly, in Dennett's case of the Muller-Lyon illusion,
the discriminative stimulus for the pigeon may be characterized as two horizontal lines of equal length, one above the other, alone, or against just certain kinds of relatively uniform backgrounds. Presumably, Dennett would not expect a rectangle formed by joining two lines, one above the other, of equal length, to count as a discriminative stimulus for a pigeon reinforced as I have described above. But the description I offered of the relevant stimulus class rules out both the rectangle and the Muller-Lyon illusion, yet is expressible in physicalistic terms.

In spelling out the possible more abstract characterizations of a behavioral type or a stimulus class, Skinner often concentrates upon a methodological issue that may have interest, but is irrelevant to the point behind these criticisms offered by Dennett and Fodor, etc. Because they are relevant to the particular examples, however, the more important issues get obscured.

This issue involves an answer to what I earlier characterized as the second question addressed to the behaviorist, namely, what if there is more than one acceptable description of a stimulus or a behavioral type. Skinner responds, correctly, that this will matter no more to the discovery of behavioral regularities than it does to the discovery of laws in any science. In setting up a situation in which reward is made contingent upon the performance of behavior, the experimenter is always making the
reward contingent upon the performance of behaviors that have a number of different properties. In the case of the rat in the experiment described above, for example, reward in fact follows the rat's pawing in a certain way, and its coming into contact with the bar, and also its coming into contact with a piece of metal. However, further experimentation can most likely discriminate the cases. The issue of whether the rat was reinforced for pawing or for pressing could be resolved by moving the bar to the opposite end of the cage, and seeing whether the rat continued to paw or to press. If the further question arises as to whether it was bar pressing or metal pressing that was reinforced, another distinctly shaped piece of metal could be put into the cage. If the rat continues to press the bar rather than distribute its presses between the two pieces of metal, then it will be the less abstract description, in this case, which is relevant. Skinner would say that it was "up to the organism" to determine the proper level of abstraction for a correct description of the response class, just as it is up to the experimenter to devise situations that distinguish among the many such descriptions that could, on the basis of a few examples of the relevant behavior, be invoked. And if there are no such experiments to be devised, the behavioral scientist is still no worse off than the geologist who classifies emeralds as green rather than grue. (Although Skinner himself doesn't argue in quite this way, these
arguments, as I said before, are easily extractable from *Science and Human Behavior*. For example, in explaining why it is that a pigeon reinforced for pecking a spot on the wall continues to do so during extinction, Skinner says that it is not necessary to say that the pigeon is looking for the spot. Rather, "although we have undertaken to condition the pecking response, we have in fact strengthened many different kinds of precurrent behavior which bring the bird into positions from which it sees the spot and pecks it. These responses continue to appear, even though we have removed the spot, until extinction occurs." (p. 89)

But, as mentioned before, this does not address the question of whether there can be even one such behavioristically legitimate description of a stimulus or a behavioral type. The point behind the examples discussed above is that the effects of reinforcement (if any) will be to strengthen a class of responses that have no physical element in common, however abstract, but may well be picked out by a description that is irreducibly goal-ascribing, or intentional. Thus the behaviorists' strictures against the use of intentional vocabulary in characterizing behavioral types will render the Law of Effect useless for the trans-situational predictions that were to save it from circularity.

Perhaps, then, we can embellish the argument offered by Dennett and Fodor.12 Above, I described an experimental situation in which Skinner could claim to have described a
rat's response, at a level that will yield correct predic-
tions, as the contact of some part of the rat's body with a
bar. But how predictive will this description, in principle,
be? Consider, for example, a situation in which the rat had
access to some instrument that could be used magnetically or
electronically to depress the bar. (And suppose that the
rat had been previously trained to use it.) The above des-
cription of the behavioral type would not include these
responses although, intuitively, we would predict them to be
strengthened as well. Thus the description in question cuts
the response class too narrow. Attempts to broaden the
class, however, will encounter difficulties in the other
direction. For example, we could describe the relevant
response as the rat's having some causal role in the depres-
sion of the bar. But suppose that, unbeknownst to the rat,
there was a mechanism in the cage that depressed the bar
when the rat ran into the corner and faced the wall. Then
the "causal role" description would cut the predictively
relevant class too broad since the rat's running into the
corner and facing the wall would thereby be counted as a
relevant response. Examples such as these make it plausible
that the behavioral type must be described as the rat's
trying to depress the bar. It is here, then, that the
irreducibility of the intentional idiom causes difficulties
for Skinner. Although he need not provide meaning preserving
translations for our ordinary mentalistic terms, he must nevertheless provide physicalistic descriptions of explanatorily relevant classes of behavior. But in cases like the above, it looks as if the explanatorily relevant classes of behavior must be intentionally described.

However, even if the response class could be defined independent of intentional concepts, there remains a second difficulty for the prospect of trans-situational prediction, namely: How is it to be ascertained that an organism is in fact being tested under certain given conditions? Can these conditions be specified by using only behavioristically legitimate vocabulary, or does the success of certain experiments in the laboratory depend upon the organism's in fact being in a state that can only be characterized by use of mentalistic concepts? If this is the case, then behavioristic predictions may be expected to have severely limited success outside the laboratory, since the regularities discovered in the laboratory are not lawlike.

Consider the circumstances under which behavioristic experiments upon laboratory animals are performed. Because of controlled conditions, the variations in the organism's internal states are minimized. Animals enter the Skinner box at 80% body weight; in most cases this assures that they are hungry, a condition usually essential for determining whether food will be reinforcing for the organism at the time. But hunger is an internal state, a feeling.
Skinner cheerfully admits that he insure that the experimental animals are hungry, but claims that hunger is merely a usual (though not invariable) collateral effect of food deprivation. The important functional relation is the amount of food deprivation in the animal to the degree of reinforcement it receives from food, a relation in which hunger itself plays no role. The functional relation itself, he suggests in *About Behaviorism*, was selected for in the evolutionary history of the species. And so for similar connections between the reinforcing properties of drink, sex, exercise, and sleep, and their respective deprivations. Thus a physicalistic specification of the relevant experimental conditions can be given, and trans-situational prediction may proceed.

But this way of specifying experimental conditions clearly will not suffice for less "obvious" cases of reinforcement. Consider typical human reinforcers, such as approval, music, art, driving race cars, keeping a job, or getting good grades. It seems quite unlikely that a physicalistic description can be given of the state of a person to whom these things, at the time, are reinforcing. Skinner could claim (and does suggest in a different context) that these events became reinforcing because they were associated with the evolutionarily selected reinforcers: Perhaps an individual listened to music when eating, or
looked at paintings when sexually aroused. But then in order for these associations to be useful in facilitating trans-situational predictions, the organism must be in the physical state of deprivation usually sufficient for hunger, thirst, or sexual arousal at all those times at which looking at a painting or listening to music is predicted to be reinforcing. This, however, seems quite implausible.

It is not necessary, however, to retreat to a discussion of peculiarly human reinforcers to make the critic's point against Skinner. As our own cases probably show, food deprivation is neither necessary nor sufficient for feelings of hunger, and the reinforcing properties of food depend upon whether we are hungry. If the experimental animal's desire for food, were not a de facto controlled independent variable, then the behaviorists' predictions would fail.

Thus, an organism's internal states such as the desire for food, have an effect upon the behavior produced. If this is so, then terms denoting such internal states should appear in the behaviorists' hypotheses, in order for the psychologist to further quantify and refine the calculations about the effects that such internal states have upon behavior. But the reference to internal states such as the desire to get food or the belief that one can get it by working is not permitted in behavioristic theories. Thus the behaviorists' methodological restrictions leave them without the theoretical resources needed to predict behavior
that results from the interaction of stimuli with different internal states. This objection parallels the "cateris paribus" objection to the definitional behaviorists, discussed on p. 37.

Skinner, however, constantly refers to internal states such as hunger as states with no causal efficacy. For example, in About Behaviorism (pp. 52-53) he says, "The behavior occurs because appropriate mechanisms have been selected in the course of evolution. The feelings are merely collateral products of the conditions responsible for behavior. This is very different from saying that these things reinforce because they taste or feel good ... [it] does not mean feelings are causally effective."

But in our examples, it seemed as if the best way to get lawlike descriptions of behavior was to include reference to internal states in the descriptions; to relate eating behavior to hunger, rather than to deprivation, and crying out to pain, instead of bodily injury. One could even give an evolutionary justification for this move: The most evolutionarily efficient way to get an animal to avoid injury, for example, is to select for one connection between pain and avoidance behavior, to insure that anything that causes pain gets avoided. This way, specific connections between specific kinds of bodily injury and avoidance behavior need not be forged.\textsuperscript{13}
Why is it then that Skinner is so intransigent about entertaining the causal efficacy of internal states? There seem to be three major reasons. First is his seeming conflation of mentalism with dualism, a phenomenon that Fodor, in Psychological Explanation, identifies as common among behaviorists. On such a view, mental states such as hunger and pain had better not be recognized as causal links between stimulations and behavior, since there are problems in accounting for the causal interaction of physical and non-physical states. Skinner does profess worry about this issue (See AB, p. 11) a worry based upon the very non-sequitur criticized by Fodor. But this worry seems at least somewhat spurious, since Skinner is willing to say that such feelings are collateral effects of bodily conditions. Thus it seems that it is not so much the causal relations of feelings to physical states that Skinner is most interested to deny, but rather, the need for reference to these feelings in the laws of behavior.

This interest, in turn, is based upon certain views about explanation that belie his more recent liberalism about the existence of "private events" and link him to the views of the methodological behaviorists that he professes to reject. Important to these views is Skinner's identification of the explanation of behavior with its in practice prediction and control. Thus, even if the appeal to feelings and other internal states did not commit one to
dualism, Skinner would nevertheless be reluctant to admit their explanatory role, given that there is no way to directly manipulate and control for the presence of these states in the laboratory. Skinner says, (AB, p. 11):

The theory of knowledge called Physicalism holds that when we introspect or have feelings we are looking at states or activities of our brains. But the major difficulties are practiced ... we cannot anticipate what a person will do by looking directly at his feelings or his nervous system, nor can we change his behavior by changing his mind or his brain.

That is, phenomena cannot be explanatorily relevant unless they can be directly experimentally manipulated; otherwise, according to Skinner, there is no way to verify their contribution to behavior. In this way Skinner, as I suggested earlier, is subject to his own criticisms of the methodological behaviorists (as well as to criticisms of positivistic conceptions of explanation in general). In About Behaviorism, he criticizes them for "rul[ing] private events out of bounds because there could be no public agreement about their validity." (P. 16) But Skinner himself is guilty of this as well.

The assumption that mentalism is equivalent to Cartesian dualism and the assumption that the explanatory appeal to unobservable events is methodologically illegitimate have been roundly, and in my view correctly, criticized in the philosophical literature of the last twenty or so years.
(See Reichenbach, Sellars, Putnam, Hempel, etc.) But there is a third assumption Skinner makes about explanation that reflects interestingly upon a number of issues discussed here. This assumption is evinced most clearly in the following passage from About Behaviorism: (p. 17)

An organism behaves as it does because of its current structure, but most of this is out of the range of introspection. At the moment we must content ourselves...with a person's genetic and environmental histories. The environment made its first great contribution during the evolution of the species, but it exerts a different kind of effect during the lifetime of the individual, and the combination of the two effects is the behavior we observe at any given time...To the extent that either can be changed, behavior can be changed.

Skinner is claiming here, and he may well be correct in claiming, that there will be no behavioral differences among organisms unless there are either environmental or genetic differences as well. But does it follow that all (or even any) lawlike descriptions of behavior must make reference solely to the genetic structure and environmental history of the behaving organism?

Fodor, Putnam, Dennett, and Davidson argue, each in different contexts, that this need not, and cannot, be the case. The following example, I hope, will illustrate the plausibility of these arguments. A Skinnerian may ask, "Given that a current internal state of a finite state
automaton can be specified as a function of the history of the system up to that time, why can't all reference to the current internal state be "reduced" to talk of the history and input alone?" One answer is this: The successive elimination of reference to each internal state must stop, unfortunately for the behaviorist, with the description of the initial state of the system. Thus since irreducible reference to even one internal state is required for the specification of the current state of the system, the behaviorists' attempt at reduction must fail.

This consequence, however, would not be automatically damaging for the more liberal Skinner of About Behaviorism. Here, Skinner could claim that one could, in principle, give a genetic account of the initial state of the organism and that the genetic properties of an individual can be specified in terms of the evolutionary properties of the species. Thus, the specification of the current state of the organism, though irreducible to behavioral predicates alone, will be specifiable in terms of behavioral and evolutionary vocabulary, and will therefore be acceptable to a radical behaviorist.

But there nevertheless remains a problem for the radical behaviorist view, namely, that the behaviorists' specifications will not capture the important regularities in the behavior of the system, if it is the case that
numerous different histories could be responsible for the occurrence of the same internal state. In this case, the reference to the internal state plus current input, but not the reference to the history plus current input, will yield the proper predictions about the system's subsequent behavior. The behaviorist, using behavioral and evolutionary terminology alone, will not have the theoretical resources to state the regularities in question. And although the different histories that result in what the mentalist refers to as the same internal state could be classed together, by the behaviorist, for the purposes of prediction, there will be no behavioral property that they have in common. Thus the explanatory power of behaviorism would be parasitic upon the descriptive categories introduced by the mentalist. Functionalism, on the other hand, is better equipped to state the laws of behavior. Thus it seems quite clear that the superiority of a "scientific" functionalist to a "radical" behaviorist view would be empirically vindicated, issues about meaning-preservation aside. This, then, concludes our discussion of whether functionalism, at least so far, is the best candidate for the codification of our ordinary intuitions of the role of mentalistic terms in the explanation of behavior.¹⁴
III. Functionalism and the common-sense theory of mind

0. Opening Remarks

In the last section I argued that functionalism, when regarded as a thesis about the meanings of mental state terms, adhered closer to our ordinary usage than its behaviorist alternative. I argued as well that functionalism, when regarded as an empirical hypothesis, provided a better explanation than its behaviorist alternative of the important observed regularities of human behavior.

As we saw above, both the definitional behaviorists and the radical behaviorists believed their claims to be relevant to the philosophical questions about mind. The definitional behaviorists believed that their "analyses" of the concepts of sensation, belief, and desire were the first steps of a program to translate all mentalistic terms into terms of dispositions to behave. These "analyses", to be sure, were intended primarily as analyses of our third-person attributive uses of mental concepts, and elements even of these uses--those involving the causal efficacy of mental states--were abandoned as well. But, as discussed before, the definitional behaviorists believed they could give a philosophical argument as to why these attributive uses were primary, the others having been derived from suspicious metaphysical and epistemological doctrines. Although these behaviorists placed restrictions upon what could
legitimately count as part of the meanings of our mentalistic terms, they were quite serious about giving behavioral translations of these terms that exhausted their so-called legitimate meanings.

The success of this program depended upon the success of these translations; otherwise the behaviorists' program would be seen not as giving the proper analyses of important and often used concepts, but merely as changing the subject. It was fairly clear, however, that any plausible behavioristic translations of even the simplest of our ordinary mentalistic attributions had neither the same implications nor even the same truth-conditions as the original sentences in question. This result, given the goals of the definitional behaviorists themselves, meant failure. I claimed that a "definitional functionalism" came much closer to the goal of preserving the ordinary explanatory functions, and thus the third-person attributive meanings, of our mentalistic terms. Two further questions about such a functionalist theory, however, remain to be addressed: First, does a "definitional functionalist" theory come close enough to preserving the ordinary attributive meanings of our mentalistic terms, and why is the preservation of ordinary meaning even a requirement for an adequate functionalist theory?

Radical behaviorists such as Skinner, on the other hand, had both a different behavioristic theory, and different views about the relation of such a theory to our
ordinary mentalistic attributions. According to Skinner, our ordinary mentalistic attributions need not be translated one by one into terms of behavioral dispositions. Rather, Skinner believes that the concepts of a certain theory about the acquisition of behavioral dispositions through the contingencies of reinforcement would (or, could in principle) replace the concepts presently used in ordinary mentalistic attribution, preserving our ability to predict and explain human behavior.

Translations may be given when they can be, according to Skinner; they can be given when some concept of behavioristic theory has sufficiently the same use in the (correct) explanation of behavior as did the old term of the mentalistic vocabulary. Otherwise, the mentalistic terms are to be abandoned. If there is no behavioristic term whose use is sufficiently like the use we make of "depression", "anxiety", or "joy", then these familiar terms are to be classed with terms such as "phlogiston": they are to be considered as untranslatable terms of a discredited theory. Locutions such as "She did it because she wanted to" are to be assigned the status of "The sun rose this morning," mere manners of speaking that are picturesque relics, again, of a theory no longer in use.

Because Skinner's theory seemed clearly inadequate as an empirical hypothesis about the production of behavior,
and because it has methodological difficulties of its own, it was clear that this particular attempt to replace our ordinary mentalistic attributions by explanations couched in a different set of concepts would fail. Thus I have not yet sufficiently discussed an important issue, namely, when is a theory to be considered, in general, as replacing our ordinary mentalistic concepts, and when as merely, in Ryle's words, "rectifying their logical geography"? Further, does "rectification" require that the ordinary meanings of the mentalistic terms be preserved? And finally, could, and if so, under what conditions could, the replacement of our ordinary mentalistic concepts count as a philosophically interesting answer to the question of what mental phenomena are? If not, does such theoretical replacement entail that there are no mental states?

These questions, however, are just as well addressed to functionalist as to behaviorist theories. Functionalist theories of the mind, which permit internal states and events to appear in laws governing behavior, have a much better chance than behaviorist theories in accounting for the behavioral regularities of humans. Even theories that permit reference to internal states, however, can turn out to be as instrumentalistic about beliefs, desires, and sensations as a Skinnerian theory purported to be. No theoretical entities of a functionalist theory designed to provide the best explanation of behavior need be identifiable with mental phenomena as ordinarily construed.
In such a case the result would be a functionalist theory that counted as an empirical hypothesis in the domain of psychology and provided a negative answer to the question of whether there are such things as beliefs, desires, and sensations. But what is the difference between such a theory and one that asserts that there are beliefs, desires, and sensations, but that their properties are at least somewhat different than we ordinarily thought them to be? Or, what is the difference between our having many false beliefs about beliefs, desires, and emotions, and there being no such things at all?

One suggestion for when a functionalist theory is merely an elucidation rather than a replacement of our ordinary concepts is that it be about (at least most of) the items traditionally regarded as mental phenomena. But this just moves the question one step back, to the question of what it is for a theory to be about beliefs, desires, and sensations: Must it provide an analysis of these ordinary mentalistic concepts, or rather, may it merely specify entities with which beliefs, desires, and sensations may be (theoretically) identified, or to which they may be reduced? However this question is answered, another remains, namely, what information must a functionalist theory include in order to provide an analysis of these concepts, or make plausible the attempted identification of beliefs, desires, and sensations with the items denoted by its theoretical terms?
One requirement (for either project) posed by a number of philosophers is that the functionalist theory remain sufficiently close to our common sense views about the kinds of mental states there are, and their efficacy in the production of behavior. David Lewis is a foremost advocate of the importance of a common sense derivation for a functionalist theory; he has offered a fairly detailed account of the content of our common sense views, and a suggestion for when a theory is to be counted as sufficiently close to them. Thus I shall concentrate upon Lewis' account of mental phenomena for the remainder of this chapter.¹

1. Common sense theories of mind: Lewis' view

Lewis believes that implicit in our ordinary attribution of mental states to persons in the explanation of behavior is a common sense theory. This theory may be made explicit by following a number of steps. First, we are to collect all of our commonly believed generalizations about the causal efficacy of mental phenomena in the production of behavior. An example of this kind of generalization would be:

(1) Pain is (typically) the result of bodily injury which (typically) leads to the desire for its cessation and the avoidance of situations similar to the one which produced it.
We may also include, as well as generalizations about the causal roles of mental phenomena, further commonly believed generalizations about the genus and species relations among them. For example,

(2) Headaches are a kind of pain.

The next step is to conjoin these generalizations, transforming sentences so that the mentalistic terms appear as names. (For example, rather than "If someone is hungry, then ...", use "Hunger results in ...") Replacing each name with a different variable, and existentially quantifying, the result is the Ramsey-sentence of what Lewis calls our implicit common sense theory of the mind. The Ramsey-sentence, as should be clear, contains no mentalistic vocabulary, and specifies only relational properties of the mental phenomena to be defined. People (or other systems) have the mental states in question if and only if they have a unique set of states and processes that interact in the way prescribed by the theory.

For Lewis, however, only those functionalist theories obtained by including such commonly believed generalizations will be close enough to common sense to provide answers to the philosophical questions with which we began. Otherwise, the theory might best be regarded as a replacement for our ordinary mentalistic explanations of behavior. As Lewis says in "Psychophysical and Theoretical
Identifications", "Hence it is analytic that either pains, etc., do not exist or most of our platitudes about them are true."

Yet Lewis requires more than the mere inclusion of these sorts of generalizations in a theory that is to count as sufficiently close to common sense. He also prohibits the inclusion of any other generalizations, even about the causal roles of mental phenomena and the relations among them, that derive from sources other than our common beliefs. Lewis says in "Radical Interpretation", "The concepts of belief, desire, ... are common property. The theory that implicitly defines them had better be common property, too. It must amount to nothing more than a mass of platitudes of common sense, though these may be reorganized in perspicuous and unfamiliar ways. Esoteric scientific findings that go beyond common sense must be kept out, on pain of changing the subject." (p. 335)

However, there is a real question about whether the generalizations derived from common sense will yield a theory rich enough to conform to our ordinary assumptions about mental phenomena—even about their role in the production of behavior. Two sorts of difficulties that concern us here have been raised in articles by Ned Block and Richard Warner. First, common sense may not yield enough information to allow us to distinguish among mental phenomena that we intuitively regard as distinct.
Second, it is easily imaginable that common sense is flatly wrong about the characteristic causes of particular sensations, beliefs, and emotions. If so, then a Lewis-style functionalist theory will not provide even necessary conditions for a state to be a mental state of a particular kind.

These claims about common sense are made plausible by various additions that psychological, psychoanalytical, and psychophysical theories have made to our body of knowledge about mental phenomena. Consider, for example, the elaborate and, from the point of view of common sense, surprising, account of the typical causes of states such as depression, anger, and anxiety yielded by various psychological theories. We may recall as well that the regimens of the typical 19th century mental hospitals were based upon certain "common sense" beliefs about the typical behavioral effects of certain mental disorders. Here are cases in which our common sense beliefs about the causes and effects of mental states were false, but were corrigible by appeal to psychoanalytic theory.

There is need as well for contributions from more experimentally oriented psychological theories. It is quite unlikely, for example, that we could distinguish, in terms of intensity, between two not very obtrusive pains in my left elbow by appeal only to what we commonly believe about their typical causes and effects. However,
information from experimental psychology, especially psychophysics, could help in providing detailed descriptions of the differences among physical stimuli which characteristically produce sensations of various intensities. The use of such information appears to be the only way to draw distinctions among intuitively distinct sensations in terms of their typical causal roles. Thus the addition of such information to common sense theory would increase the plausibility of functionalism.

Further, not only is common sense unaware of or imprecise about such causal differences, but where it rules at all, it is likely to rule incorrectly. For example, according to the psychophysical experimentation done by Fechner and modified by Stevens,\(^3\) the physical magnitudes typically responsible for certain sensations are not those that were commonly and pre-theoretically thought to be the causes of those sensations. In particular, the perceived degrees of brightness, loudness, and pain change as a logarithmic, not linear, function of the change in the physical magnitude responsible for the sensation. Here is another motivation, then, for the addition of theoretical information to our common sense generalizations about mental states.

To be sure, Lewis himself offers a modification of his theory that goes some way in meeting these objections. In "Psychophysical and Theoretical Identification" Lewis
suggests that rather than form a single conjunction of all our commonly believed platitudes, we form a "disjunction of all conjunctions of most of them. (That way it will not matter if a few are wrong.)" (256) Thus, if one of the disjuncts is uniquely realized, our ordinary terms will denote, and their meanings will be explicated by the conjunction in question. However, if the criticisms of Block and Warner are sound, it is quite implausible that even one of these disjuncts would be uniquely realized since common sense may not be merely wrong, but may require theory to aid in its correction.

Further, there is another difficulty with Lewis' common sense requirement as well, namely, that it may be very hard to distinguish between common sense and more theoretical generalizations about mental phenomena. Are statements such as "People tend to get angry at their children because they are repressing anger at their employers" or "People are often aggressive because they have low self-esteem" part of common sense theory? Permission to include the generalizations of psychology in the functional definitions would obviate the need to draw the distinction.

Why then does Lewis favor the constraints upon common sense theory that I quoted above? That is, why is it the case that the theoretical characterization of a common sense concept be itself a product of common sense? These
conditions are required by Lewis as a consequence of his taking very seriously the claim that a functional theory is to provide implicit definitions of our ordinary mentalistic terms. For example, the definition of 'pain' can be obtained from the Ramsey-sentence of Lewis' common sense theory as follows:

First, the existential quantifier that binds occurrences of the variable that replaced the term 'pain' may itself be replaced by a definite description operator; then the phrase "pain" may be affixed to the left of the entire sentence. Pain will thus be specified as, for example, the (unique) state that bears certain relations to stimulations, behavior, and other states which are quantified over, but not mentioned. But it is not enough for Lewis that the theory yields necessary and sufficient conditions for being a pain that are devoid of mentalistic vocabulary; what Lewis wants is definitions in the literal, ordinary meaning-preserving sense. Thus the information included in the generalizations about pain that appear in the common sense theory must, if this sort of implicit definition is to be served, enter into and exhaust the ordinary meaning of 'pain'.

It is clear that Lewis is treating our ordinary mentalistic terms such as 'belief', 'pain', and 'anger', as theoretical terms that derive their meanings solely from the theory in which they are embedded, and that he believes
this theory to consist of the ordinary explanations of behavior by appeal to mental phenomena which he deems common property.

Lewis argues, following Sellars, that this treatment of mentalistic terms, though genetically incorrect as an account of how our terms acquired their meanings, nevertheless serves to exhibit the meanings they now in fact have. (PTI, p. 257) Lewis, however, does more than merely follow Sellars in treating these terms as theoretical terms in an explanation of behavior: he is making two further claims as well.

First, he is advancing a general view about how theoretical terms acquire and retain their meaning and reference. Second, he is introducing rather severe constraints upon the details of the theory of which terms such as 'belief', 'pain', and 'danger', are to be a part. Although Lewis' views about the meanings of theoretical terms influence his views about what counts as legitimate content for a common sense theory of mind, I will reserve the discussion of this topic for the next section, which is devoted to a general discussion of the meaning and reference of theoretical terms. Here I would like to examine Lewis' condition that the content of our common sense theory be none but "common property", and determine just how close to common property are the generalizations that Lewis himself suggests to be prime examples of the elements of a common sense view.

79
I shall argue here that the generalizations that Lewis permits in the theory are in fact rather far from common property. These generalizations, however, can be seen as revisions of others that do in fact seem to be closer to common property. I do not suggest that the somewhat uncommon generalizations be excised from common sense theory, but rather than the principles of revision used to justify Lewis' examples of our typical common sense generalizations be used, in certain cases, to justify the inclusion of what, in Lewis' view, are more "esoteric" generalizations as well. I shall claim, further, that even the addition of these further generalizations will not produce an adequate theory of mind, and that the addition of still further information will be required. I shall argue, however, that these additions are actually in the service of common sense, where the contents of common sense are determined by constraints counter to, but more intuitively justified than, those offered by Lewis. Thus I hope to show that Lewis' requirement that a functionalist theory of belief, sensation, and emotion adhere to common sense accounts of these phenomena is interesting and important, but that his views about what counts as common sense need revision.

How, then, does Lewis justify his account of our common sense theory of mental phenomena? First of all, the kinds of causal and relational platitudes that Lewis
includes in this theory constitute a small subset of the kinds of things actually said in our ordinary appeals to mental states in the explanation of behavior. Further, some of the platitudes that he takes to be part of our common sense theory never occur in our ordinary explanatory attributions of mental states. Thus there arises a question about what the relation is between what Lewis calls our common sense theory, and the appeals to mental states that people commonly make in the explanation of behavior.

I want to stress right away that Lewis need not be held to what people actually say in their ordinary explanations of why others act as they do. Nor need Lewis restrict himself even to what it is appropriate to say in such circumstances. Although it might be a problem for a philosopher who identified common sense theory with ordinary language that people rarely uttered a certain statement, this need not be a problem for Lewis. He does not require that the elements of common sense theory be commonly used; for Lewis they must be, rather, conventionally believed. That is, we (all) must believe them, believe that others believe them, and believe, in turn, that others believe that we believe them.4

One source of insight, however, into the common beliefs of a community is the ascertainment of what in fact is freely said or assented to by way of ordinary explanation. It is in contrasting what is ordinarily said with
what Lewis claims to be conventionally believed that I hope to extract a set of principles which can be used for zoning, as Lewis implicitly does, a community's common property.

Thus let us look more closely at the so-called common sense explanations of behavior that Lewis has often used as examples of ordinary mentalistic attributions and determine whether or not they are in fact products of unadorned common sense. Generalizations such as the following are typical examples of what Lewis calls the platitudes of common sense.

(1) Having hallucinations can impair one's driving.

(2) If a person wants to catch the bus, and believes that if he runs down the street he will catch it, then he will run down the street.

(3) (P)(Q)(If someone wants Q and believes that only P will bring about Q, then he will want P).5

Statements such as these were claimed by Lewis to be among our common, or conventional beliefs. Obviously, then, there is a great difference between conventional belief and ordinary attribution. For example, (2) and (3) are rarely found, in those forms, as parts of our unreflective explanatory appartus. And although (1) would be ordinarily offered or assented to in explanation, so would other
generalizations about sensation that would not be among Lewis' favored platitudes. For example, Lewis' common sense theory would not include statements such as these:

(4) If you stare into the sun, then look away, there will be a round, yellow blotch before your eyes that is not really there.

This statement appears to assert the existence of sensory particulars which have location ("before your eyes") but are not in physical space. Thus it would be impossible to get a physical realization of a theory that included this statement. But since functionalism purports to stay ontologically neutral, Lewis (and other functionalists) does not want to be committed to the existence of such particulars. Instead, most functionalists offer an alternative to (4) that is to be equivalent in common sense explanatory force; for example,

(4') If you stare into the sun, then look away, you will have a round, yellow after-image.

This statement is taken to refer not to after-images and their peculiar properties, but to experiences or states of persons of certain kinds. Such a state may be referred to as 'having an after-image' as long as 'S has an after-image' is not taken to imply '(Ex)(S has x)'.

Thus, Lewis, in ruling out or providing translations for statements that, on the face of it, are incompatible with functionalism, is diverging from the project of
transcribing into theory all and only the explanatory generalizations of common sense. I do not want to suggest that Lewis is wrong in doing so; on the contrary, I believe that he has very good reasons—all the reasons adduced in Chapters I and II for thinking that functionalism provides a promising theory of mind. Some of these reasons, however, are quite theoretical. My purpose here is not to suggest that theoretical considerations have no place in refining or extending common sense theory, but to make Lewis' theoretical considerations more explicit.

(3) is a statement that Lewis would want to include among his generalizations since it would give the theory great predictive and explanatory power. And again, though (3) is not the sort of statement that would be ordinarily offered in explanation, there may be other considerations that would justify its inclusion in common sense theory. Perhaps one could argue as follows: Though the generalization itself would never be stated, it may be imputed to a community if a sufficient number of its instances were commonly offered or assented to in explanation. Why should belief in the instances be grounds for belief in the generalization?

We could try to argue that the believed instances would not count for the community as explanatory unless the generalization that we wish to impute were believed to be lawlike by the community. But as Davidson points
out in "Causal Relations", this suggests conditions upon the acceptance of singular causal statements that are too strong. Our belief that a certain statement describes a causal relation among various entities does not imply that we believe that the universal generalization of that singular causal statement is a law. What we must believe, rather, is the much weaker claim that there is some law or other that is in fact instantiated by the events described in the singular causal statement. So, this strategy will not provide grounds for the imputation of beliefs of type (3) to the community.

Perhaps, however, the mere fact that people believe the instances of this generalization is enough to ground the attribution to them of belief in the generalization itself. Lewis makes such a suggestion (in a somewhat different context) in Convention, where he distinguishes two different kinds of belief in generalizations. The first is belief in sensu composito, or (roughly) belief that particular instances of the generalization hold, for each instance entertained. However, it seems intuitively unsettling, if not counterintuitive, to routinely ascribe general beliefs to people who hold them merely in sensu diviso. Thus, whether or not such attributions are ultimately justifiable in terms of their theoretical utility, they represent a deviation from common sense.
But not only are generalizations such as (3) deviations from common mentalistic explanation, but so, in some cases, are their instances. Consider, for example, (2). (2), in its full explicitness, would probably never be uttered in explanation of why someone ran down the street. What ordinarily would be said, rather, is:

(2') People run down the street in order to catch the bus.

Using the language of belief and desire, we might ordinarily get, at best:

(2'') People run down the street (in certain circumstances) if they want to catch the bus.

These statements, though more commonly used than statements such as (2), are not couched in terms of the interactions among beliefs and desires. Why then is (2), rather than these latter statements, considered to be among our conventional beliefs?

The reasons for excluding (2') are different from the reasons for excluding (2''). (2'') may be taken as an elliptical rendition of (2) itself. It is elliptical because we assume, in accepting (2''), that the persons whose behavior is being explained are running towards the bus. In other contexts (for example, if the person were running away from the bus, or out into the path of the bus) this would not count as explanation of their actions. This example, however, merely warns that it might be
necessary to add **contextual information** to what is commonly said in order to arrive at the **beliefs** about mental phenomena that are commonly held by the community in question. As such, this is not a very interesting example of the divergence between what is commonly said, and what may be taken to be commonly **believed**.

(2'), on the other hand, yields a more interesting constraint. This is an example of a teleological explanation; here an action is explained by the goals or purposes of the person who performed it. Explanations in terms of goals, that is, **future** occurrences, is by and large rejected as a form of bona-fide causal explanation. Explanation in terms of the beliefs and desires of the agent is favored, since beliefs and desires are taken to be events that occur prior to, or states that are present at the same time as, the action to be explained. Thus while (2'') could be taken to be merely an elliptical rendition of (2), (2'), is a more substantive transformation of (2). But what kind of transformation?

(2) is not a conceptual analysis, nor a meaning preserving translation of (2'). It is, rather, a reconstruction of plausible truth conditions for (2') which are designed to conform to the more well-established schemata for causal explanation in the sciences. The relation between (2) and (2') is thus interesting, I claim, because more is going on here than the mere addition of context to
some commonly offered explanations; more also is going on than the mere generalization of explanations that receive common assent. Considerations external to the body of common sense attributions of mental phenomena, arising from general views about the nature of explanation and of causal efficacy, have been used as constraints upon what may legitimately count as the commonly held beliefs of a community.

Let us, then, attempt to make the principle that legitimized the reconstruction of ordinary teleological explanations more explicit, as follows: "Don't include principles in an explanation of behavior in terms of mental phenomena that flout other well-established principles of scientific explanation, unless absolutely necessary." In itself, this principle is not very controversial; in fact, it makes good sense even as a caveat for a common sense theory. After all, the common sense theory of mind is to be considered a theory; principles that insure its harmonious relation to other theories may be welcome here. Even common sense theory ought to be subject to the constraints placed upon other theories that attempt to explain observed phenomena.

What is problematic about this principle, however, is again that it conflicts with Lewis' other requirement that the information used to define the theoretical terms of our common sense explanations of behavior be common property
in Lewis' restricted sense. What is problematic about this requirement, in turn, is Lewis' hope that it will insure a conceptual analysis, or meaning-preserving translation, of our ordinary mentalistic terms. However, not only will this restriction leave Lewis' functional account without the "esoteric" theoretical resources needed to distinguish among mental phenomena that we intuitively regard as distinct, but, as we have now seen, it could prohibit the inclusion of certain statements about the causal roles of belief and desire that Lewis regards as paradigmatic instances of commonly believed generalizations about mind. But if Lewis permits us to "step off" common property in the latter case, then why not in the former as well? Why is this not support of the view that our common sense mentalistic theory is one among many theories of the world, each influenced by, and revisable in terms of, the others?

It would be open to Lewis to claim, however, that his constraints upon the generalizations that belong to common sense functionalist theories are derived from very general theoretical considerations. These, if not the products of uneducated common sense, need no more to guarantee their emergence than general education; they may be Socratically extracted from individuals who have no knowledge of psychological theory, nor, indeed, of any special science at all. On the other hand, the generalizations from psychology and psychophysics that we are also considering for addition to
the common sense theory would only be available to those who had explicitly studied the subject. So on a slightly looser, but still plausible, view of what it is to be commonly believed, there is nevertheless a distinction between the two kinds of generalizations.

Nonetheless, it looks as if Lewis' theory, given his arguably justified view of what counts as common property, will not if Block and Warner, etc., are correct, will not be realized as it stands. Must we then give up on an attempt to define mental phenomena in terms of a common sense derived functionalist theory, and leave the individuation of mental phenomena and the explication of their nature to cognitive psychology? This conclusion would be unwarranted, not only because there may be plausible conditions other than Lewis' for what counts as a common sense theory, but that there seems to be provision for some of the leeway we need in Lewis' theoretical machinery itself. Lewis himself does not attempt to exploit this feature of his theory, but that is no reason why such a strategy should not be attempted. This requires, however, a somewhat more detailed look at Lewis' general account of how terms introduced by theories acquire meaning and reference.

2. The meanings and denotations of theoretical terms

According to Lewis, the meaning of a theoretical term of any given scientific theory derives from and is exhausted by
its role in that theory. Thus Lewis' functionalist account of the meanings of mentalistic terms is just a special case of a general view about the meanings of theoretical terms. That is, the meanings of terms such as 'electron', 'anti-proton', and 'gravity', as well as 'pain' and 'anger', are determined by constructing the Ramsey-sentence of the theory in which the terms occur, and using it to form as many Ramseyfied definite descriptions as there are terms to be defined. As in the former case, the definiens of each term in a particular theory includes all the non-theoretical terms of the theory plus quantification over entities which interact among themselves and with the other entities mentioned in the theory. Since the definitions take the form of definite descriptions, which themselves include expanded definite descriptions as clauses, each theoretical term denotes if and only if there is exactly one set of entities which satisfies the theory; that is, if and only if the theory is uniquely realized. The theoretical terms in different theories can be said to name the same entities if their Ramsey-equivalents uniquely pick out the same thing.

Let me acknowledge that Lewis' view, as stated here (and in his own writings), may seem incompatible both with functionalism and with the way that I have been talking about functional definition. First, Lewis does not, on the face of it, permit the functional theory to be realized by different entities in different individuals at various
times. His view can be modified, however, as Field shows in "Mental Representation," so that it requires merely a unique realization of the theory in an individual at a time.

Second, Lewis suggests that the terms of our common sense theory—'pain', 'anger', etc.—denote not abstract functional states, as I have claimed, but the state-types that (in each individual) uniquely realize the functional definition. In his writings, Lewis distinguishes between pain and having pain, or being in pain. 'Having pain', for Lewis, denotes a functional property, or abstract functional state shared by all those who, no matter how, realize the common sense theory that includes the term 'pain'. But pain itself is different in different individuals: 'pain' denotes, in each individual, the appropriate component of the unique realization of common sense theory. Thus, what Lewis calls 'having pain' is what I have been calling 'pain'. In this discussion of Lewis' views, those who favor my approach may replace Lewis' question, "Does 'pain' denote and what does it denote (in a given individual)?" with "Is pain realized in that individual, and how?" The answer will be of interest since the identification of pain (in my sense) with a functional state or property is correct only if all and only individuals that, intuitively, are in pain are in states that are appropriate "pain-components" of a unique realization of the theory. Further, my interest
in Lewis' question of what 'pain' (in his sense) means lies in the fact that it can be transformed into the question of what functional state or property 'pain' (in my sense) denotes.

Does Lewis' account of theoretical terms provide a plausible account of their meanings and the conditions under which they can be said to denote? Arguments for the plausibility of this approach often center around cases like the introduction of the term 'phlogiston' by 18th century chemical theory. That this theory of combustion was so far from the facts can be plausibly seen as the reason for why the term 'phlogiston' had no denotation. The 18th-century belief that there was such a substance derived from the introduction of the term by an incorrect theory; the incorrectness of the theory seems to be sufficient explanation for the falsity of the belief in such a substance.

However, this view has counterintuitive consequences as well. First, no newly introduced term of a theory will denote unless the definitional equivalents of all the terms in the theory introduced at that time are uniquely satisfied. Thus the inclusion of any incorrect assertion about any entity in the theory will leave the theory unrealized, and the theoretical terms without denotation. Further, there will be many cases in which those who hold current theories of various phenomena cannot be said to refer to
the same things as did those who held slightly less enlightened theories which used the same theoretical terms, but included some incorrect assertions, since the terms in the earlier theory would not denote at all. For example, given that current atomic theory provides a substantive correction of Bohr's atomic theory, we must take his term 'atom' as having no denotation.

This view, then, appears to have the consequence that if the common sense theory of mind is unrealized as it stands (or realized by more than one n-tuple of entities), then our ordinary terms such as 'pain', 'belief', and 'anger' have no denotation. (Or, speaking as I have been, pain, anger, etc. are unrealized in human beings.) But there is actually a source of greater flexibility in Lewis' view. In "How to Define Theoretical Terms," Lewis asserts that if the Ramsey-equivalents of the theoretical terms of some theory T fail to pick out a unique set of entities, the theoretical terms may nonetheless denote if the theory has "a (unique) near-realization: that is, an n-tuple that does not realize the original theory, but does realize some theory obtained from it by a slight weakening or a slight correction." He continues, "We might want to say that the theoretical terms name the components of whichever n-tuple comes nearest to realizing the theory, if it comes near enough." (p. 432)
As I have mentioned before, the criticisms of Block and Warner suggest that common sense theory is unrealized as it stands, and is corrigeable only by appeal to information outside the bounds of common sense. However, it may nevertheless be true that our common sense theory is not far from being correct. If so, the theory (call it T) may have a near-realization, or more accurately, a nearest realization that is near enough. If so, then our ordinary mentalistic terms could be said to denote; people could be said to have pain, beliefs, and anger.

But could they be said to have pain, beliefs, and anger in the same, common, sense, or must we regard the meanings of our ordinary terms as having changed? After all, T will have a near-realization only if there is a theory T', obtainable in certain ways from T, that is uniquely realized. Suppose we actually obtain such a theory. In such a case, Lewis suggests that T' be taken as the term-introducing theory that provides the meanings for the theoretical terms that initially appeared in T. But the Ramsey-equivalents of the theoretical terms in T' will be different from their Ramsey-equivalents in T, and thus, if Lewis believes that the Ramsey-equivalents of theoretical terms provide implicit definition, these terms will have undergone a change in meaning from the time of their unreflective use in common sense explanation. But could they be said to have pain, beliefs, and anger in the same, common, sense, or must we regard the meanings of our ordinary terms as having changed? After all, T will have a near-realization only if there is a theory T', obtainable in certain ways from T, that is uniquely realized. Suppose we actually obtain such a theory. In such a case, Lewis suggests that T' be taken as the term-introducing theory that provides the meanings for the theoretical terms that initially appeared in T. But the Ramsey-equivalents of the theoretical terms in T' will be different from their Ramsey-equivalents in T, and thus, if Lewis believes that the Ramsey-equivalents of theoretical terms provide implicit definition, these terms will have undergone a change in meaning from the time of their unreflective use in common sense explanation. However, the changes may be so slight or benign that the terms might still be said, informally,
to have retained their previous meanings. Whether this can be so in the case of our common sense theory, of course, depends upon the details of T'.

Now, if Block and Warner are correct in their objections, it looks as if any correction of T that has a chance of unique realization will incorporate information that goes beyond, and in some cases goes counter to, the deliverances of common sense. Such corrections may or may not be substantive enough to insure that T has no near-realization; for now, let us assume that they are not. Nevertheless, given the kinds of corrections they are, it looks as if the ordinary meanings of our mentalistic terms have indeed been substantively changed by T'. Though in this case, common sense theory would be (uniquely) near-realized, this would be to no avail; we would have, by taking the somewhat "esoteric" T' as the term-introducing theory, changed the meanings of our ordinary mentalistic terms, and thus, as Lewis would claim, changed the subject.

It might be claimed, however, that this takes too restrictive a view of what the subject is. I began this chapter by suggesting that a philosophically satisfying functionalist theory be a theory of the beliefs, sensations, and emotions that we talk about in common sense explanation. Lewis' requirement that functional definitions preserve the ordinary meanings of our mentalistic terms was just one way, and a powerful one, of accomplishing
this. But if we drop Lewis' requirements about meaning, we can use some of his theoretical machinery and his insights about common sense to determine whether our ordinary mentalistic terms denote. We may use Lewis' notion of a near-realization to determine the conditions under which a functionalist theory that diverges from a faithful transcription of our common sense beliefs may serve to specify the reference, if not preserve the meanings, of our ordinary mentalistic terms. This will serve the purposes of those, like myself, who are interested not in the preservation of meaning, but of denotation.

So, on this proposal, our ordinary mentalistic terms can be said to denote if there is a uniquely realized T', whose realization is a near-realization of our common sense theory. Or, from another point of view, if there is such a theory T', it will be an adequate, common sense derived, functionalist theory of the pains, desires, and beliefs of common parlance. As I have stressed, however, it looks as if such a T' will include information that flouts, or at least is not available, to common sense. My question, then, for the rest of the chapter, is whether any such T', if it did have a unique realization, could yield a near-realization (or a near enough nearest realization) of T. Does anything in Lewis' discussion of near-realization rule out this possibility?
Unfortunately, Lewis is not entirely clear about what the relation between T and T' must be in order for a unique realization of T' to be a near-realization of T. Sometimes, as in the passage I quoted from "How to Define Theoretical Terms", he claims that T' can include corrections of T. He permits corrections as well in his discussion of near-realization in "Psychophysical and Theoretical Identifications": "The story as told is unrealized, true of no one. But another story is realized, indeed uniquely realized: the story we get by deleting or correcting the little mistake. We can say that the story as told is nearly-realized, has a unique near-realization," (p. 252, my italics). It is not clear, however, just what kind of corrections T' can include, and sometimes, Lewis speaks of T' as being merely a weakening of T. However, Lewis gives, as an example of an unrealized theory, the theory of phlogiston, which he says comes "nowhere near being realized." This suggests that T' may indeed include corrections, perhaps many corrections, as long as they are not too drastic. What counts as overly drastic? Lewis doesn't say, adding only that "The notion of a near-realization is hard to analyze, but easy to understand," ("PPTI, p. 252).

I shall take my question to be, then, the question of whether there is, intuitively, anything about an "esoterically" corrected common sense theory that would prohibit a unique realization of it to count as a near-realization of common
sense theory itself. I shall argue that there is not; that some ways of revising common sense theory, even though they involve the use of information that is not itself part of common sense, remain true to the spirit of common sense theorizing about the effects of mental states upon behavior. A theory that includes just these sorts of revisions, I shall argue, will count as legitimately common sense derived, and may serve, if it is uniquely realized, to specify the denotations of our ordinary mentalistic terms.

But first, I want to point out that, in such a case, we could argue that the meanings of these terms had been preserved as well! If the common sense theory of mind is unrealized as it stands, it is similar to most other theories that have ever been catalogues: few are correct without some revision. Further, this generalization about theories is not a product of esoteric investigation, or philosophical reflection; it is, at least arguably, a dictum of common sense. If so, it would not go counter to common sense to claim that the definition of 'pain' (for example) was not

(P) Pain =df. the state that has the appropriate role in our common sense theory T,\(^{12}\)

but

(P*) Pain =df. the state that has the appropriate role in T or in the revision of T such that a unique realization of it would be the nearest near-realization of T.
Suppose that we actually came up with such a theory $T'$. Then the definition of 'pain' in $T'$ would be:

$$(P') \text{ Pain = } df. \text{ the state that has the appropriate role in } T'$$

Because $(P')$ mentions $T'$, while $(P^*)$ merely picks it out uniquely, it could be argued that $(P')$ and $(P^*)$ differ in meaning. However, since the definientia in $(P')$ and $(P^*)$ denote functional properties—which we know to have just the same realizations—it might be argued that $(P')$ is close enough in meaning to $(P^*)$ to count as merely a trivial change. $(P^*)$, however, is clearly a candidate for explicating the ordinary meaning of pain; though it alludes to a theory that may go beyond common knowledge, it does not explicitly include any information but that provided by common sense. And though some might claim that it goes counter to common sense to insist that this is what people are implicitly asserting when they talk about pain, I would not expect Lewis to raise these objections. Given that he says, in "How to Define Theoretical Terms":

I contend that the theorist who proposed $T$ by asserting the postulate of $T$ explicitly, labelling it as the postulate of a term-introducing theory, has also implicitly asserted the definition-sentences of $T$ (p. 439).

then we might expect him to agree that the theorist is implicitly alluding to (though not mentioning) the product of the almost inevitable revisions of the theory as well.
But whether or not Lewis would agree with this proposal about the ordinary meaning of pain does not matter. My concern is whether our ordinary terms have denotation, and whether it is possible for a functionalist theory that makes use of "esoteric" information ever to specify these denotations. As I said before, I shall argue that it is possible. Unfortunately, I do not have a comprehensive view about what sorts of revisions of common sense theory are substantive, and which are relatively slight. However, I do have a few suggestions. Although I have rejected the requirement that a common sense derived functionalist theory be just a transcription of our common beliefs about mental states, this does not mean that just anything goes. A lot depends upon just how the "esoteric" information is used to modify the body of commonly believed generalizations.

One suggestion is that we be permitted to use scientific information only in order to withdraw certain generalizations about the causes of mental phenomena or the relations among them. There is a precedent for this strategy in Lewis' own account of the content of common sense theory. As noted in the last section, Lewis rules out teleological generalizations on the basis of general theoretical considerations; thus it might be argued that the use of more specific information to rule out certain perhaps more rashly believed generalizations ought to count as well within the bounds of
common sense theory, too. But perhaps the explicit addition of esoteric information, especially when stated in technical vocabulary, ought to mark the revised theory as something other than common sense derived.

However, it seems arbitrary to hold all available scientific information about the causes and effects of mental phenomena to the role of providing censorship alone. It seems counter-intuitive as well. For example, it seems that in the following case the addition of scientific information about the causes of the phenomenon in question would serve the purposes of, rather than run counter to, common sense.

Suppose, for example, that people generally develop an itchy rash—a unique, peculiar sort of itch—after sitting, or so it is thought, under an old oak tree. But it is in fact false that this sort of itch is typically produced by sitting under oak trees; indeed, it is not connected to the oak at all. Instead, a certain chemical secretion from an insect that infests oak trees is responsible for the almost universal allergic reaction falsely attributed to contact with the oak. This would be a case in which scientific information, if permitted into common sense theory, could lead to a correction in what are commonly believed to be the typical causes of a peculiar kind of sensation—the oak itch. (Ned Block uses a similar example in "TF:, an example regarding a peculiar kind of smell, commonly but
falsely attributed to skunks.) Our question is, should the addition of the information about the chemical secretion of the oak fly be permitted, or must the theory, given the scientific debunking of the folk etiology of the oak itch, remain mute as to its actual typical cause?

One consideration is that the discovery would have predictive consequences for common sense theory: The statement describing a connection between oak trees and the allergic itch will not support counterfactuals to the same degree as the statement describing a connection between the chemical secretion and the itch. So far, however, this is besides the point. We expect that the additional information will extend the predictive power of the theory; we are asking whether the resulting theory can nevertheless count as common sense.

It would, I think, for reasons related to, though not the same as, the claim that the predictive power of such a theory would be increased. To prohibit the addition of the esoteric information about the cause of the oak itch would belie the claim that this body of generalizations was a common sense theory of mind and behavior. What Lewis calls common sense theory was introduced (according to Sellars' myth of Jones, which Lewis claims is a useful, though genetically false, reconstruction of our disposition of mentalistic terms) in order to account for why people behaved differently in ostensibly similar environmental
situations, and why people claimed that items which in fact had different colors looked the same; in short, talk about distinctions among internal states was introduced in order to provide a causal explanation of otherwise inexplicable behavioral differences and similarities. But now that we (thanks to the genius of Jones) enjoy full reporting use of the mentalistic vocabulary, we may be expected to ask the same sorts of questions about what we recognize as peculiarly distinctive internal states. Thus to rule out the information that may answer this question is to restrict the very sort of theorizing that led to the formulation of common sense theory itself. Further, as I suggested earlier, in a slightly different context, to restrict the content of common sense theory to what is now considered common property rules out the possibility of theoretical growth and change, phenomena which should be possibilities for any theory. Further, still, it would belie the actual histories of common sense theories themselves, since we see frequent examples, from folk medicine to pop psychology, of the incorporation of various initially uncommonsensical information about causes and effects into the body of these theories.

So far, however, I have tried to justify only the modification of our common beliefs about the causal roles of the mental phenomena we believe to exist. I have not said anything about the legitimacy of tampering with the
common sense roster of mental phenomena or the classifications it recognizes among them. Could common sense theory, for example, be extended to include new theoretical terms that purport to denote entities other than the ones that figure in our common sense beliefs? Could certain terms (and their referential purport) be deleted? Could we revise the generalizations we believe about the genus and species relations among mental phenomena as well?

It seems that none of these sorts of modifications automatically insures that a theory has passed the bounds of common sense. On the contrary, there are examples of each kind that seem quite benign, as well as those that seem to herald substantive theoretical revision. What characterizes certain of those modifications as benign, I claim, is again their being introduced in the Jonesian spirit— in order to explain certain otherwise inexplicable implications of, and inconsistencies among common observations, and to increase the explanatory coherence of common sense theory as well.

Suppose, for example, that the inhabitants of a certain (small) community believed that, typically if not invariably, watching a movie at the Town Arcade made one thirsty. Given the rest of their common sense theory about the typical effects of movie watching and the typical causes of thirst, this generalization stands out as, though not inconsistent with the rest of the theory, nevertheless decidedly odd.
Yet this generalization is firmly and commonly believed. We could imagine, however, a latter day genius like Sellars', who came up with a theory of subliminal stimulation. This would involve the addition to the theory of a new term that purported to denote the (unconscious or semi-conscious) mental state of a person so stimulated, such that the presence of this state was causally responsible for feelings of thirst. This hypothesis, plus the further one that this new state gets evoked by the presentation of images on the screen of that particular theater, helped to provide a coherent explanation for the heretofore bizarre-seeming belief. This addition, again, seems in the interest of common sense, since it adds to the general coherence of our common sense explanations.

More generally, when the mentalistic terms become "old" or "clearly understood" terms of common sense theory, (see Lewis' "HTDTT"), further explanation may be necessary to account for why the entities they denote occur in some of the situations that they do. This may call for the introduction of a new batch of theoretical terms, defined (provisionally) in terms of their relations to each other, behavior, stimulation, and previously introduced mental states. Freud's postulation of unconscious mental phenomena seems to fulfill this model, and parts, at least, of his theories have actually achieved the status of common belief.15

As for the deletion of terms from the theory that we commonly
thought denoted mental states, and the reclassification of states into kinds, the same considerations apply. Although a theory that claimed that there were no pains, beliefs, or desires would not count as an extension of common sense, there may be cases in which the reference to some state may be dropped in the interest of the general coherence of the theory. For example, both psychological information, and a more enlightened perspective upon certain social conditions may lead us to deny that there is such a state as the Victorians called female hysteria, or nerves. What we used to believe was a distinctive phenomenon may now be seen as a loosely connected group of different syndromes, still within the bounds of common sense. These examples, of course, were designed to show that, in certain cases, psychological and even psychophysical information could be introduced into the theory retaining its claim to common sense. I have given no general constraints, however, nor necessary and sufficient conditions for what counts as a common sense theory.

What I have suggested so far is that we should add to or subtract from common sense theory only that information regardless of its source that helps us to preserve, by and large, the distinctions that we commonly believe to hold among kinds of mental phenomena (although, as suggested above, there are circumstances in which we could legitimately reject even these). That is, we are to try to find actual
distinctions in the typical causal roles of those mental states that we commonly believe to have different causal roles. But in order to preserve the common belief that there are distinctions among the causal roles of particular mental states, it is not necessary to preserve all our common beliefs about what those causal roles might be. But the condition that the distinctions, by and large, be preserved is too weak a condition on common sense theories; at best it is a necessary condition for when a common sense theory is benignly modified, rather than replaced. Thus it seems that any extension of a common sense theory will have to preserve quite a bit of the actual content of our common beliefs about the causal roles of mental phenomena, and even more of our common beliefs about how many and what kinds of mental phenomena there are. But, as I have claimed, a theory may undergo modifications more radical than those that Lewis permits, and still remain a product of common sense. Whether or not the addition of such information will produce a theory that is uniquely realized, of course, may be left as an empirical matter.

Yet it is argued in some quarters that the attempt to provide plausible conditions for what counts as common sense will ultimately be fruitless; that even if a "benign modification" of Lewis' common sense theory were to be uniquely realized, it would nevertheless not provide necessary and sufficient conditions for entities to be
mental states of particular kinds, since there are, or could be, entities that satisfy the functional definition, but fail to have the properties distinctive of states of that kind. In the next two sections, I shall examine such arguments, and evaluate their effect on the prospects for an adequate functionalist theory of mind.
IV. Functionalism and the problem of qualia

0. Opening remarks

In the previous chapter I remarked that there is doubt about whether any sort of characterization of mental states in terms of their typical causal roles, be it derived from common sense or from psychology, could possibly be successful. In the next two chapters I want to discuss the motivations for these doubts.

A number of contemporary philosophers concerned with functionalism believe that this doubt can be generated for some mental phenomena, but not for others. It is commonly believed that while functional definitions may provide necessary and sufficient conditions for belief and desire, phenomena like sensations and feelings will be more resistant to a functionalist account. However, in the next two chapters, I shall argue that, on the contrary, functionalism faces great difficulties in giving an adequate account of belief and desire, while the inadequacy of a functional definition of sensations is at best not proved.

In order to clarify this discussion, I shall make a rough classification of mental phenomena on the basis of some of what we take to be their distinctive properties. My interest here is in the distinction between what I shall call the qualitative and intentional properties of mental phenomena. Qualitative properties, or qualia, are those that seem peculiar to pains and itches, after-images...
of yellow, and feelings of warmth: the so-called "felt" properties such as hurting, yellowness, or being hot. In what follows, I shall call those mental phenomena that (seem to) have such properties qualitative states and events. On the other hand, mental phenomena such as beliefs, desires, thoughts, and wishes are usually characterized by another set of properties. It is in virtue of their content that beliefs and desires are said to be foolish or reasonable, true or false. Further, beliefs and desires are representational: beliefs are about people and objects; people and objects figure in our desires as well. Any mental phenomena that share these representational properties shall be classed in what follows as intentional.

Of course, qualitative and intentional phenomena may be said to share properties. For example, my pains and itches, as well as my beliefs and desires, may be said to have relative strengths. Further, it may be the case that some mental states have both qualitative and intentional properties; indeed, emotions and passions are often thought to be of just this sort. Although it may be argued that desires too have qualitative properties, in what follows I shall nonetheless use beliefs and desires as examples of purely intentional phenomena, and pains and itches as examples of purely qualitative phenomena.

What is important to stress, however, is that our ordinary explanations that appeal to mental phenomena make
reference to both their intentional and qualitative properties. Thus any functionalist theory that purports to provide a distillation or a slight modification of our common explanatory appeals to mental states must give an adequate account of both sorts of properties. This can be pursued, as suggested earlier, in either of two ways: first, by showing that these properties are reducible to, or otherwise definable in terms of, various features of the functional roles of the phenomena in question; second, by arguing, alternatively, that the mental states in question really have no such properties, and providing an explanation of why we traditionally thought they did.

Thus the burden of the critics of functionalism is to provide an argument that can stop both lines of defense. And indeed, in both classical and contemporary discussions of mental phenomena, there have most often been two distinct steps in the attempt to show the inadequacy of views that treat the causal properties of mental phenomena as alone essential to their nature. Doubts that such views are adequate are usually generated by a thought-experiment designed to show that an entity that satisfies the functional definition, or occupies the causal role, of a particular mental state could nevertheless fail to have a certain property that mental states of that kind are intuitively thought to have (or vice-versa). Second, it must be argued that our intuitions about the important properties of mental
states are sound, and cannot be outweighed by any theoretical considerations that we in fact have.

1. The problem of absent qualia

Why is it doubtful that functionalist theories provide adequate characterizations of what it is to have an after-image of yellow, or to be in pain? In the contemporary literature on this topic, there has been much discussion of a class of counterexamples to a functionalist account of sensation. These difficulties were first aired in Block and Fodor's "What Psychological States are Not", and were discussed in Shoemaker's "Functionalism and Qualia". Also, in Kripke's "Naming and Necessity," there is an influential argument against the psycho-physical identity thesis, which can be modified to make a point of this sort against functionalism as well. Examples from this class of counterexamples include the following:

First, I could possibly be in a state that is typically caused by my seeing roses and results in my saying "My, how red" (given the appropriate beliefs and desires) but be having the sensation that you typically have when you look at a bunch of broccoli and say, (given the appropriate beliefs and desires) "My, how green." This is the inverted qualia, or more traditionally, the inverted spectrum problem. The difficulty is related to, though not identical with, a problem discussed earlier, namely, whether or not a purely
functional characterization could capture the intuitively recognized distinctions among qualitative phenomena. In the "inverted qualia" case, you and I are imagined to make the same distinctions among red and green objects, and to have the same beliefs about the colors of the objects we see. This a functional characterization could capture. Yet it would not be able to tell us which qualia accompanied which states.

Second, someone could be in a state that is typically produced by being hit with a hammer and that typically results in yelling and rubbing the wound, believing one is injured, and desiring that whatever is going on in one cease, and nevertheless not feel pain; indeed, perhaps not feel anything at all. This is what is called the "absent qualia" problem.

It is concluded by the critics of functionalism that the qualitative properties of pains and other sensations cannot be captured by a purely functional characterization, and that functionalist theories are therefore incapable of providing type-identity conditions for qualitative states. The success of these counterexamples, moreover, would seem to threaten not only functionalism, but materialism in general, since counterexamples against other materialistic theories can be generated in the same way. The argument from the possibility of absent or inverted qualia to the failure of functionalism proceeds as follows: (I shall use pain as an example.)
(1) It is possible that some entity could be in a state that occupies causal role R (the role associated with pain by the functional theory in question) but not be in a state that is qualitatively like pain (say, in particular, a state qualitatively like your last headache).

(2) No state is a pain state unless it is qualitatively like pain.

Thus, (3) an entity could be in a state that occupies R without being in pain. Thus, (4) The property of being in pain is distinct from the property of being in a state that occupies role R.

And since I have previously declared my intention of taking 'pain' to denote the property of being in pain, we may conclude:

(5) Pain is distinct from the abstract state or functional property of being in a token of a state that occupies role R.

Thus the argument, given my identification of pain with the property of being in pain, is valid. Is it sound as well? Few philosophers have sought to challenge premise (2)²; it has been regarded as tantamount to self-evident. (2) can be denied, however, for either of two reasons. One could claim, first, that states with different qualitative characters could all count as pain states as long as they shared certain other properties, perhaps causal roles. To be sure, the argument might go, our pains and other sensations do
have qualitative character, but having a certain qualitative character is neither necessary nor sufficient for being a pain. This response is unsatisfactory, however: Though it avoids the conclusion that pain is not a functional state, it leaves the materialist with the question of whether the property of being qualitatively like my last headache can be given a physical or topic-neutral characterization.

A more radical challenge to (2) would involve the claim that no pains have qualitative character: that there is really no such thing as the qualitative content of sensation. This response is so counter-intuitive, however, that it ought to be a conclusion we are driven to, rather than one that we would willingly embrace.

Indeed, I shall accept (2), and agree that all pains must be qualitatively alike (and similar, perhaps, to my last headache), and try instead to challenge premise (1). What then, is the argument for (1)?

2. Arguments from conceivability

I believe that the most promising way of arguing for (1) derives from Kripke's remarks against the psycho-physical identity thesis at the end of "Naming and Necessity". Kripke's interest there of course, was neither to establish (1), nor to argue against functionalism, but rather to conclude that it is possible for an entity to be in pain without having C-fiber stimulation. His conclusion, however,
depends upon a general, and seemingly powerful, method of determining when claims about the possibility of certain situations are true.

A reading of "Naming and Necessity" suggests that it is possible that \( P \) if:

(*) It seems to us, on careful reflection, as if we can imagine or conceive a world in which \( P \) obtains.

Or, to put it another way:

(a) It seems as if the pictorial, or purely qualitative, content of our imagination is correctly described as a world in which \( P \) obtains, and

(b) There is no alternative description of what we seem to be imagining or conceiving other than a world in which \( P \) obtains.

Kripke illustrates the use of his test with a number of examples. The statements, "Hesperus is Phosphorus", and "Temperature is mean kinetic energy," are necessary truths. Thus their negations cannot be possible, and must not meet conditions (a) and (b) if the test is to be valid. Yet the negations of these truths do meet condition (a): it seems prima facie as if we can imagine or conceive a world in which Hesperus is not Phosphorus, and a world in which temperature is not mean kinetic energy. Kripke suggests, however, that these statements are enjoined from counting as possible by the failure of condition (b): There are ways
of redescribing what we allege to be imagining or conceiving that do not invoke the distinctness of Hesperus and Phosphorus, or the distinctness of temperature and mean kinetic energy, yet do capture the qualitative content of what has been imagined.

For example, we can describe a world in which a planetary body picked out in the way we initially identified Hesperus--the "star" first seen in the evening--is distinct from the planetary body picked out in the way we initially identified Phosphorus--the "star" first seen in the morning. Such a description would capture the qualitative content of what we were imagining when it seemed as if we were imagining that Hesperus was distinct from Phosphorus, as it describes a world in which the planetary body first seen in the evening is not the planetary body first seen in the morning. However, what has been described is a world in which there are two planetary bodies, and not a world in which Hesperus is not Phosphorus.

Similarly, though it may seem as if we can imagine a world in which temperature is distinct from mean kinetic energy, there is an alternative description that captures the qualitative content of what we are imagining. Temperature, that is, mean kinetic energy, was initially identified as the phenomenon that produced sensations of warmth and cold in perceivers. What we may be said to be imagining, then, is a world in which some phenomenon other than
molecular motion is responsible for such sensations. This description does indeed capture the qualitative content of what we had imagined, but it is not a description of a world in which temperature is not mean kinetic energy.

In the first case, the availability of an alternative description turned on the possibility of there being other planetary bodies that could strike perceivers in just the way that Hesperus, that is, Phosphorus, strikes us. In the second case, we depend upon the possibility of there being other phenomena that could strike perceivers in the way that temperature, that is, mean kinetic energy, strikes us. But according to Kripke, the lack of such a possibility in the case of pain thwarts the attempt to provide a redescription of what seems to be a coherently imaginable world in which pain is distinct from C-fiber stimulation. Though it is possible for there to be worlds that are qualitatively like, from the point of view of perceivers, but not actually, worlds in which temperature causes sensations of warmth or cold, and worlds in which Phosphorus is the first planetary body seen in the morning, there can be no world in which there is a phenomenon that is qualitatively like, but not actually, pain, since, as Kripke puts it, "anything qualitatively like pain just is pain." Thus there is no alternative description of what we are imagining when we allege to be imagining a world in which pain is distinct from C-fiber stimulation. Our initial tendency
to describe the content of our imagination as a world in which pain is not C-fiber stimulation cannot be explained away.

Kripke has also expressed the difference between these cases as a difference between the terms 'temperature' or 'Hesperus' and 'pain'. The reason it is not possible to explain away the prima facie conceivability of pain existing without C-fiber stimulation is because there is no contingent, reference-fixing description commonly associated with the term 'pain'. If there were, then perhaps we could, as we did in the cases of 'temperature' and 'Hesperus', describe the content of our imagination as a world in which something other than pain satisfied that description.

Now, my interest is not in Kripke's argument against the identity thesis, but in whether his conceivability test can be used to show that it is possible for an entity to be in a state that occupies the functional role of pain but not feel pain. Nevertheless, I want to look more closely at Kripke's case against the identity thesis, as it sheds light, more generally, upon what he believes the relation between conceivability and possibility to be.

Some critics of Kripke object to his claim that anything qualitatively like pain is pain. This claim is the converse of our premise (2) in the argument against functionalism that we have been considering. I believe that Kripke's claim should be accepted for the same reasons that
we accepted its converse, namely, that it is intuitively plausible, and that its denial merely raises the further question of whether the property of being qualitatively like pain can be given a physicalistically acceptable explanation.

Other critics object to Kripke's entire enterprise of drawing metaphysical conclusions from considerations about what we can imagine or conceive. Why must the scope of our imaginative abilities, they ask, have anything to do with what is, or what could be? It seems to me, however, that a philosopher ought not to dismiss the appeal to conceivability so easily. I believe, with Kripke, that considerations about what we can conceive or imagine (or what we would say about situations that we allege to imagine) are important for metaphysical argument, and I shall argue later that a functionalist, in particular, has special reason to agree.

However, I do not believe that considerations about what we can conceive can provide a quick, definitive "test" for what is possible; their function is much more complicated. Indeed, I shall argue that arguments from conceivability to possibility, when interpreted so they can bear philosophical weight, turn out to be examples of a much less controversial, but also less distinctive, kind of philosophical argumentation.

For the time being, however, I shall accept the view
that P is possible if it meets conditions (a) and (b) above, and ask whether on this assumption it is nevertheless possible to respond to Kripke's argument against the identity thesis. And indeed, some philosophers have attempted to meet Kripke on his own terms, and to argue that even if it is true that anything qualitatively like pain is pain, it is nevertheless false that "Pain is distinct from C-fiber stimulation" meets conditions (a) and (b). Or, to put the point in the "formal mode", some philosophers have argued that even if it is true that there is no contingent description commonly associated with 'pain', the prima facie conceivability of pain as distinct from C-fiber stimulation can nevertheless be explained away.

It has been argued, for example, by Putnam and Boyd, (as reported in Block's "Troubles with Functionalism"), that what looks to be a world in which pain is distinct from the firing of C-fibers can be successfully redescribed: What is being imagined, according to them, is a world in which our instruments for detecting C-fiber stimulation lead us systematically astray, and not a world in which pain is not C-fiber stimulation. (As Block points out, such a redescriptions can be given of any world in which the identity of some macroscopically described entity with some microscopically described entity is in question.)

True enough, there can be no merely qualitative analogue of pain, but there can be a situation that is
qualitatively like, but not actually, a situation in which the firing of C-fibers has been detected. That is, there is a contingent, reference-fixing description associated with the other side of the identity statement--'C-fiber stimulation'--namely, 'the brain phenomenon detected by use of instrument I'. What we are imagining, then, is a world in which some phenomenon other than C-fiber stimulation is so detected, and not a world in which pain is not identical with C-fiber stimulation. The availability of this description, then, prevents the statement, "Pain is distinct from C-fiber stimulation", from meeting condition (b), and we no longer have grounds for believing that it is possible.

Nonetheless, it is open to a supporter of Kripke to say that the mere fact that there exists some redescrip-
tion of what seems to be a world in which pain is distinct from C-fiber stimulation does not give us grounds for doubting the possibility of such a world. A further ques-
tion arises as to whether it is in the least plausible, or intuitively satisfying, to describe the content of what we have imagined in this way. And supporters of Kripke are likely to say that it is most certainly not. Note, however, that this response requires a substantive change in what we took to be the conditions under which a statement P was possible. Instead of our original second condition, that is:
(b) There is no alternative characterization of what we seem to be imagining or conceiving other than a world in which P obtains, we will henceforth judge that a statement P is possible if it meets condition (a) and,

(b') It is not intuitively correct, or natural, to characterize the content of our imagination in any other way than as a world in which P obtains.

As should be clear, this formulation shifts the importance from the existence of an alternative characterization to the intuitive correctness of any alternative characterizations there may be.

Now I agree with the supporters of Kripke that the Putnam-Boyd redescription of what seems to be a world in which pain is distinct from C-fiber stimulation does not seem to capture the content of what we were to have imagined, and thus does not go far in assuaging qualms about the identity thesis. However, it is not so easy to state why this is the case. It would be obviously illegitimate to claim that scientific investigation could not possibly help us to establish the properties, or at least the essential properties, of phenomena such as pain. This certainly is a view of Kripke's, but one that ought to follow from the conclusions reached through the conceivability considerations he offers; they should not be used to help him draw such conclusions. What then is the diagnosis of the intuitive implausibility of the Putnam-Boyd account?
Block suggests, in discussing this example, that the Putnam-Boyd account can be criticized as implausible because it involves the attribution of too much mistake to the inabitants of that world. After all, we are to imagine them as being systematically and invariably deceived by their cerebroscopes.

It seems, however, that the alleged difficulties with this example ought not to be attributed solely to its appeal to widespread mistake, since there are other, non-controversial cases in which it is most natural to explain why it seems that we can imagine the negations of necessary truths by appeal to mistakes of the same sort. For example, consider what could be quite naturally said about the claim to conceive that the Pythagorean theorem is false. What is really going on, we might say, is that we are imagining a world in which we measure the sides of right triangles, square the (fallacious) results, and end up with a number different from that of the square of the hypotenuse. Indeed, Kripke himself uses a similar strategy in "Naming and Necessity" to redescribe what seems to be a world in which Godel's theorem is false: What we are imagining, according to Kripke, is not a world in which Godel's theorem is false, but rather a world in which someone offers a proof of its negation which is mistaken. Why, then, should these redescriptions count as plausible, or natural, while the Putnam-Boyd redescription of what seems to be a world
in which pain is distinct from C-fiber stimulation is rejected as implausible?

Here a supporter of the conceivability test might say that the question of what makes one description more plausible than another is misplaced. The test, after all, was designed to elicit our intuitions about which worlds are possible worlds, and these raw metaphysical intuitions are to be taken as the best evidence of what in fact is possible. That we have the intuitions we do, however, is basic, and needs no further explanation. The very fact that we regard the Putnam-Boyd appeal to mistaken cerebroscopes as intuitively unsatisfactory, but the appeal to mistaken measuring rods and calculations as satisfactory is enough to show that there are relevant metaphysical differences between the cases.

I ultimately want to argue that it is a mistake to regard our judgments about the plausibility of these descriptions as an exercise of raw metaphysical intuition. Some examination of what these intuitions are based on is required in order for these conceivability tests to have any special role in deriving metaphysical conclusions. Otherwise, the use of these tests to support positive claims about possibility will beg questions, since the intuitions we have about what is possible may be influenced by the very theories that the conceivability test is used to judge.
Here, however, I want to ask another question. Suppose, contrary to what I intimated in the last paragraph, that we can and do have raw metaphysical intuitions that provide reliable indications of the way the world is. If this is so, what does the alleged implausibility of the Putnam-Boyd redescription of what seemed to be a world in which pain was not C-fiber stimulation show? It would indeed seem to show that pain is distinct from C-fiber stimulation.\textsuperscript{5} However, one can deny, and can do so on the basis of conceivability considerations, that pain is C-fiber stimulation (or any other kind of neural state) and nevertheless be a materialist. Indeed, this is what it is to be a functionalist.\textsuperscript{6} After all, it is by a conceivability argument (of the sort in which the relative intuitive correctness of different descriptions of an imagined situation are being judged) that functionalism is shown superior to the psycho-physical identity thesis. Functionalism is regarded as more plausible because, intuitively, we would say of an entity who satisfied the functional definition of pain, but did not have C-fiber stimulation, that it was in pain, and thus that pain could not be C-fiber stimulation.

When Kripke's conceivability test is interpreted as a device by which one imagines some situation (in purely qualitative terms), and then chooses the most intuitively correct description of what it is that one has imagined, it looks to be a rather familiar tool used by functionalists as
well as dualists. As such, Kripke's argument against the identity thesis should not count (nor indeed does Kripke think it counts) as an argument against materialism in general. However, when the conceivability test is interpreted in this way, it is no longer possible for a functionalist to object, in general, to the enterprise of drawing metaphysical conclusions from what we can imagine or conceive.

It is now that the question in which I am most interested naturally arises, namely, is it conceivable for an entity to be in a state that satisfies the functional definition of 'pain' without feeling pain? That is, can we, purely qualitatively, imagine a situation that is most plausibly described as a situation in which an entity satisfies the functional definition, but feels nothing qualitatively like our last headache? If so, (then because we can establish property distinctness by establishing the possibility of an object which instantiates one but not the other) conceivability considerations will have provided reason for believing that pain cannot be a functional property. Further, in this case, there will be no out for the materialist. There is no way to account for the greater intuitive correctness of the description of such a world as a world in which the entity feels no pain other than to admit that no physicalistically acceptable account of pain is, after all, possible. Thus, dualism could ultimately be vindicated by even the weaker, "relative correctness"
versions of the conceivability arguments we have considered so far. However, given our current interpretation of how the conceivability test is to work, our intuitions about the relative correctness of the alternative descriptions of what we imagine provide the ultimate court of appeal for a priori arguments about identity. If we have very weak, or no intuitions about a certain case, then there is no longer any basis for an a priori argument about the identity or distinctness of the items involved.

I am going to argue that, in this case, our intuitions tend to support the functionalist. It can be argued, I shall claim, that we cannot, upon reflection, conceive of an entity as satisfying the functional definition of 'pain' without having the relevant quale. At the very worst, I shall argue, it just is not clear whether or not such a situation is conceivable. Either way, however, there will be no a priori reasons, based upon considerations of what can be conceived, for rejecting the identification of pain with a functional property.

Let us see, then, what the attempt to describe such a world involves. First of all, let us get clear about what we are to be imagining when we imagine a world in which an entity satisfies the functional definition of 'pain'. As I have argued in past chapters, the correct functional definition of 'pain' may go beyond our common sense theory of this phenomenon; it may include as yet undiscovered features
of its causal role. The fact that we may be, at least in part, in ignorance of the actual functional definition of 'pain' will be of importance to my argument later on. All that is important here, however, is that we are well aware of some of the important causes and effects of pain. What is to be imagined, in the cases to be considered now, is a world in which an entity has these causes and effects of pain, a world in which the entity satisfies what we believe to be a partial specification of the functional definition of 'pain'. The question, then, will be: Could such an entity fail to feel pain?

3. Humans who cannot feel pain

I now want to discuss a case which seems, initially, to support one's Kripkean sympathies as well as any case can. I shall conclude, however, that a functionalist could argue that, even in this case, it is at best not clear whether we can conceive of a person who is in a state with the appropriate causal role, but does not feel pain. This case involves a person who, at least initially, is by hypothesis unable to feel any pain. It is significant (as I shall discuss later) that this thought-experiment involves the case of a human being, and not a machine or other such entity, but I believe that if we get clear about what can be conceived in the case of humans, the other cases will fall into place as well.
Consider, then, a person born without the capacity to feel pain. Some people are, indeed, unable to feel pain; they, of course, sustain tremendous damage to themselves, since they can break bones, suffer burns, and be afflicted with internal injuries without being aware of it. So far, this provides no problem for the functionalist; though these people are not in pain, they are also not in states with the appropriate causal roles. However, in order for them to have the least chance of survival, such people have to learn to notice their bodily states and judge that they are in injurious circumstances, and do their best to avoid such circumstances at that time and in the future.

Indeed, let us imagine that we have a person who is explicitly taught the consequences of impending bodily injury as follows: "When your hand is sufficiently close to a flame, you are in a state that you should find preferable to get out of (since it leads to bodily damage), and, by withdrawing your hand, you will get out of it." Once the lessons were firmly learned and well-entrenched the person will, by and large, when his hand is sufficiently close to a fire, believe he is in a state that is preferable to be out of, believe that the state is injurious, believe that it is similar to certain states he has been in in similar situations, desire to be out of the state, and make moves to get himself out. Are we now imagining a person who satisfies the functional definition of 'pain' without feeling pain?
It seems not, upon reflection. What we are imagining here is merely the case of a person who is in a state that has some of the typical effects of that state that satisfies the actual functional definition of 'pain'. It is easy, however, to point to differences in the roles of these states. For example, normal pain-feelers must have an unconditional or underived dislike of the state that satisfies the functional definition of 'pain'; it is part of the functional role of pain to produce the unconditional desire to remove oneself from that state. (The desire, of course, is sometimes overridden by others, but its production is nevertheless part of the causal role of pain.) In our case of the non-pain-feeler, however, his desire to remove himself from the "painful" circumstances is highly derived; his aversion to the state is conditional upon his knowledge of its typical consequences.

However, we can plausibly revise the example so as to insure that our non-pain-feeler does have the right underived desires. Such a person can indeed be imagined to have an intrinsic, and not merely conditional dislike of such a state, compatible with our hypothesis that he feels no pain. If the lessons are well enough learned, he may well come to have a habitual aversion, which, though perhaps engendered by his knowledge of the usual consequences of such a state, automatically attaches to his being in the state itself. Would this be a case of a person who satisfies the functional definition of 'pain' without feeling pain?
No, or not yet, at any rate. Even though such a person's "pain" states produced the same beliefs and desires (and behavior) as ours, there is nevertheless a difference in their respective causes. For the person who is born with the inability to feel pain, it will be neither bodily injury itself, nor the stimulation of any sensory receptors, that produces the state which in turn gives rise to the relevant beliefs and desires, but the threat of injury, or the attention paid to situations that would seem injurious, that is causally relevant. Our pain states, but not his "pain" states, would produce the relevant beliefs and desires in us even at times at which we were not paying attention to the part of the body that is, or is about to be, afflicted. Thus there is a functional difference between our state of pain and the state responsible for such beliefs and desires in the person who, by hypothesis, cannot feel pain.

We may think we have imagined a case of an entity's satisfying the functional definition of 'pain' without feeling pain because the state that typically produces the effects described above in normal human beings is the state that satisfies the actual functional definition of 'pain'. But the above description, though it happens to be satisfied by pain in normal human beings, is not the bona fide functional definition of 'pain'. The two definitions do not pick out the same entities in all possible worlds.
What we have imagined in the above case, then, is a non-pain-feeler who satisfies a description that is similar in various ways to the functional definition of 'pain', but not an entity who satisfies the actual functional definition of 'pain' without feeling pain.

But perhaps we have still not developed the example far enough. Suppose, then, that it was possible to set up some sort of monitoring device for the non-pain-feeler such that he would in appropriate circumstances, automatically be put into the state that gives rise to the beliefs and desires associated with pain (in us) and the appropriate avoidance behavior. This device, perhaps an implant into the brain (or the brain and the sensory receptors) would be activated by just the same environmental stimuli that activate pain in us; it could operate, at least in certain cases, at times at which the person was distracted, inattentive, or asleep. And, as in our case, it could even be set off sometimes by internal occurrences that do not herald impending injury at all. Would this count as a case in which a person satisfied the functional definition of 'pain' yet felt no pain?

It does seem as if such a situation is prima facie imaginable; at the very least, it meets our condition (a). I maintain, however, that it does not meet condition (b'):

It is not the case that there is no other natural, intuitively correct, description of what we have imagined.
Indeed, there is an alternative description of what we are imagining that, intuitively, seems to better capture the content of our imagination. What we are imagining is not that the person satisfies the functional definition of 'pain' without feeling pain, but rather that he has undergone an operation that enables him, for the first time, to feel pain. I believe that this description of what we are imagining has greater plausibility, as it is difficult to imagine what could act, for human beings, as such a monitoring device other than a sensation that is intrinsically disliked and which serves to alert one, at different times, to afflictions in different parts of the body. And what could such a phenomenon be but a sensation that is qualitatively like our last headache?

I believe that intuition favors the functionalist's account of what it is we are imagining, namely, that the person has undergone an operation which permits him to feel pain for the first time. However, I recognize that some may say that the situation is not so clear. True enough, it may be said, we can plausibly describe the situation we have imagined as a situation in which a person has just come to be able to feel pain. However, there is also at least some intuitive plausibility in maintaining that we have indeed imagined a situation in which an entity satisfies the functional definition of 'pain' without feeling anything qualitatively like our pains.
Now, strictly speaking, such a situation would not cut against my argument, since we still would not have a case which met our conditions (a) and (b'). My goal, as will be remembered, was to show that we could not imagine what seemed to be a world in which pain was distinct from a certain functional property, such that there was no other more intuitively correct way of describing the content of our imagination. If, as I have supposed, our intuitions are torn between two alternative descriptions, then the conceivability test cannot be used to argue for the falsity of functionalism; in this case, our so-called raw metaphysical intuitions provide no help in deciding this metaphysical question.

But though such a verdict would not block my argument, it nevertheless tends to render it somewhat suspicious. After all, consider this case in comparison with the case of Hesperus and Phosphorus. In that case, after careful reflection, there is no temptation at all to describe what we have imagined as a situation in which Hesperus is not Phosphorus. Why, then, does the threat of metaphysical inconsistency not rule out all descriptions of the content of our imagination other than the one that supports functionalism? And what are we imagining when we think that the case of the non-pain-feeler provides an example of a being that satisfies the functional definition without feeling pain?
Let me address the second question first. Here I must appeal to the fact that what we know of the functional definition of 'pain' is rather involved and complicated, and the fact that we may well not yet know all the details of the causal role of pain. These facts are important, in that they could provide an explanation of what is going on when a person alleges to imagine an entity that satisfies the functional definition of 'pain' without feeling pain. It could be said that what is going on is that we are not really imagining an entity with the relevant functional property, since we are not paying attention to, or perhaps do not even know, some of the relevant features of the causal role of pain; instead, we may be said to be imagining an entity that satisfies a somewhat different functional definition, which, even if it happens to pick out pain in normal persons, is not co-extensive with the actual functional definition of 'pain'. What we may be said to be imagining, that is, is a world in which an entity satisfies the approximation of the functional definition of 'pain', without having the appropriate quale. But this is not a world in which pain is distinct from the property expressed by the actual functional specification of that phenomenon.

Some may wonder, perhaps, about the plausibility of such an explanation. I believe, however, that the explanation becomes plausible when we reflect upon what has gone on in the successive stages of the example I devised of the
person who initially cannot feel pain. What motivated the temptation, at each intermediate stage, to say that here was a case of an entity that satisfies the functional definition of 'pain' without feeling pain, was our inattention to some aspect of the functional role of pain. In each case, what we had in mind was an entity that satisfied a functional description that merely approximated the actual functional definition of 'pain'. But if it was plausible to imagine that this is what was being imagined in the previous cases, it ought to be plausible to suppose that this is what is being imagined here as well. Our success in explaining away the previous alleged counterexamples to the functional definition of 'pain' provides inductive evidence that this sort of explanation will also be successful here.

So far, I have attempted to justify my diagnosis of what we are imagining when we claim to imagine that our non-pain-feeler satisfies the functional definition of 'pain' while retaining his affliction. But now I must say why it is that some people might, even upon reflection, evince even the slightest temptation to claim that what they have imagined was a person who satisfied the functional definition of 'pain', but did not feel pain. This would not occur, I imagine, in the case of Hesperus and Phosphorus, or the case of temperature and mean kinetic energy: no one (who agrees with Kripke's treatment of proper names and kind
terms) would claim to be imagining a world in which Hesperus was not Phosphorus, nor temperature, mean kinetic energy.

But the comparison of our case with the cases of temperature and Hesperus reveals other instructive differences as well. The functionalist case is the only one we have considered in which there is no independent reason for believing that the items to be conceived are in fact identical or distinct. We all believe that Hesperus is Phosphorus, and that temperature is mean kinetic energy, and we believe these things on the basis of well-established theories of astronomy and physics. On the other hand, the issue of whether or not pain is a functional property is still an open question. This suggests, then, that the stronger our antecedent beliefs about the identity of things, the easier it is to elicit firm judgments about the "intuitively" preferable description of the content of our attempts to imagine the distinctness of those things. But this suggests that the alternative descriptions of the contents of our imagination seem more or less intuitively plausible not on their own, but only in combination with the other beliefs about the world that we hold.

If this is so, however, then conceivability arguments, by themselves, have less utility than might be expected in resolving open questions about identity, since the very openness of the question makes it difficult to describe, that is, interpret, what it is that we are actually imagining.
This is not to say, however, that they have no use; on the contrary, considerations about what can be conceived, or what seems intuitively plausible are, and ought to be, used to determine the relative plausibility or naturalness of various philosophical theses. What I am saying is that the results of conceivability considerations may serve more as an indication of how well a certain statement fits together with the rest of our beliefs than as a definitive test of what is or is not possible. If this is so, then any split in intuitions about the identity of pain with a functional property may be due merely to the lack of a theory (or even a less formal body of other relevant beliefs) that would either support or overrule our intuitions in either direction.

If this is so, however, then I am committed to saying that the more we develop our functional theory, and show that it fits in with other of our views, the less our temptation will be to think that there is any other plausible description of what we have imagined other than an entity who has, for the first time, been made to feel pain. Indeed, I am committed to the claim that the success and entrenchment of functionalism would ultimately make any other descriptions of the content of our imagination strike us as intuitively false. I am willing to accept any such commitments, however, as I believe that our intuitions already favor the functionalist's description of our case.
An interesting case for comparison here is the case of Church's Thesis. According to Church's Thesis, the mathematically precise concept of **recursiveness** provides a property with which the less precise, intuitive concept of effective computability may be identified. However, this identity is certainly not **provable**. The identity is supported by showing that every effectively computable function so far examined has indeed been recursive, and that the apparent counterexamples have arisen from the failure to appreciate some or other of the details of the definition of 'recursive function'. Because recursiveness and effective computability have so far coincided, and because the apparent counterexamples have been successfully explained away, the denial of the thesis may be considered difficult if not impossible to conceive. If, in the case of Church's Thesis, the coincidence of concepts and the entrenchment of the identity statement makes its negation less and less conceivable, then so, possibly, in the case of pain and functional property P.

Suppose, however, that I am correct, and future conceivers indeed have no temptation to describe our case as anything other than a case of a person who has just become able to feel pain. Suppose they regard the suggestion that we have described a non-pain-feeler who satisfies the correct functional definition of 'pain' as **inconceivable**. Some might argue that I should not be so sanguine about
such a prospect, since it looks to be incompatible with some of my other views. To see why this is so, let us compare the way I hope to explain away any prima facie conceivable of the claim that pain is not a functional property with the way in which Kripke explains away the prima facie conceivable of the claim that Hesperus is not Phosphorus. According to Kripke, when we think we are imagining that Hesperus is not Phosphorus, we are really imagining that the heavenly body which in some other world appears first thing in the evening (as Hesperus does in this world) is not the same heavenly body as the one which in that world recedes last thing in the morning (as Phosphorus does in this world). We are imagining that two distinct things satisfy, in some world, two descriptions which happen, in this world, to be satisfied by the same thing, and not a world in which that thing is distinct from itself. Thus, by invoking alternative, contingent descriptions of Hesperus and Phosphorus, we have an explanation of what it is that we have imagined when we thought we had imagined that Hesperus was distinct from Phosphorus. Further, this explanation also permits a nice account of how this necessary truth could nevertheless be a product of empirical discovery: We have always been in possession of distinct descriptions of heavenly bodies that appear to us in characteristic positions, and have discovered that they are descriptions of the same thing.
In the case of the functionalist account of pain, however, I have used a different strategy. So far in explaining what it is we are imagining, I have invoked approximations of the functional definition of 'pain' which may well be alternative, contingent descriptions of the state-type that we suppose to realize the functional definition of 'pain', but I do not appeal to the existence of an alternative, contingent description of the functional property (or, for that matter, pain) itself. Rather, the prima facie conceivableability of pain as distinct from the relevant functional property is diagnosed as the product of inattention, ignorance, or haste. Further, it looks unlikely, that we have any such description in hand. But then, how could we empirically discover that pain is that functional property? I also suggested that if we were (at some future time) in possession of, and paying attention to, the complete functional definition of 'pain', we would no longer claim to be able to conceive that the property expressed by this definition is distinct from pain. But again, this makes the explanation of what is going on in the pain-functional property case closer to the explanations of the apparent conceivableability of the negations of mathematical and geometrical truths than to the cases of the empirically established identities of Hesperus and Phosphorus or temperature and mean kinetic energy. Indeed, I have explicitly compared this identity to Church's Thesis.
But then am I not saying that "Pain = Functional property P" is in some sense an a priori, as well as a necessary truth? And if so, how can this be reconciled with my earlier claim that the ultimate functional definition of 'pain' is not necessarily a product of our common sense theory of that phenomenon, but may depend upon the results of psychological investigation as well?

The answer, I think, is this: Yes, there is an a priori ring to the claim that pain is a certain functional property. It seems that, at the very least, we must say that "Pain is identical with functional property P" provides an explication of our intuitive concept of pain. However, it is not inconsistent to say that empirical considerations may in fact be relevant in helping us to come up with the proper explication. Such considerations are relevant, indeed in two different ways, in many cases of explicative identity.

First, such definitions are open to revision, on empirical grounds, in the following way: Take, for example, our functional definition of 'pain'. We might discover an entity that shares a good deal of our functional organization, but which, for independent reasons, we do not believe could feel pain. This could lead us to complicate the functional definition so as to rule in human pain feelers, but rule out these beings. Such a discovery, however, would merely provide data that could have been provided by
thought-experiment alone; data that, by sharpening our intuitions, force us to add some details to the definition of 'pain'. But there are many other examples of this phenomenon. Take, for example, a definition which used to count as a (crude) explication of free action—namely, 'actions performed without external compulsion'. Freud's investigations into the unconscious, however, revealed a source of compulsion other than that provided by external circumstances: the unconscious beliefs and desires of the agents themselves. If we accept the view, resulting from theory construction in psychology, that these mental processes are relevant sources of compulsion, then we are committed to refining the definition of 'free action' in order to rule out actions produced by the "compulsive" beliefs and desires. But here again, though the refinement would not have seemed necessary without the discovery of compulsive desires, the discovery merely calls our attention to a possible counterexample to the original definition. However, the counterexample in question could have been constructed by a (rather imaginative) thought experiment alone. And let us look again at Church's thesis that the intuitive concept of effective computability is correctly explicated by the mathematically precise concept of recursiveness. Even this identity could be disconfirmed by what look to be empirical considerations: If we discovered a machine that effectively computed functions
for which there was no recursive procedure, we would have found an empirical counterexample to Church's Thesis. However, once again, the discovery of a counterexample would merely have served to alert us to a possibility that had been there all along.

There is another source of empirical input into these explications as well. In the case of pain, we might need some rather sophisticated psychological experimentation in order to determine various features of the actual functional definition of 'pain'. And it may be wondered, how that could be compatible with my claim that it is, or will someday be, inconceivable to suppose that an entity had functional property P without being in pain. These claims, however, are compatible if we suppose that the role of empirical investigation is merely suggestive—that its role is to provide resources for the expression of a concept that will be, when understood, intuitively adequate as an explication of the concept pain. Again, there are similarities between our case and the cases of free action and Church's Thesis. In the former, it might be argued that the development of the theory of the unconscious was necessary for uncovering certain information about motivation that must be incorporated into any definition of 'free action'. And in the case of Church's Thesis, it took considerable mathematical research to come up with a concept
that could serve as an explication of effective computability. Thus the a priori ring to all these explicative identities is compatible with their seeming to be established by the research methods that are relevant to those respective fields.

Of course, the ultimate vindication of functionalism, if there is to be one, must await the development of functionalist theories that provide a comprehensive definition of 'pain'. What I have argued here is merely that considerations of what can be conceived do not militate against the identity of pain with some functional property; indeed, if anything, they support this identity. However, some may argue that, even so, I have not fully addressed the issue of whether absent qualia are possible. At the very best, I have shown that it is unlikely that we can conceive of a human being that satisfies the functional definition of 'pain' without feeling anything that is qualitatively like my last headache. However, there are other cases which may not be so easily explained away.

4. Non-human cases

In "Troubles with Functionalism", Ned Block argues for the possibility of absent qualia by having us test our intuitions on a case that involves a non-human being. We can imagine, he claims, that the population of China someday decides to duplicate my functional organization: each person agreeing to play, for a time, the role of some
particular neuron in my brain. Consider the resulting entity (hooked up, if we like, to an artificial body with the appropriate sensory transducers). If such an entity is functionally equivalent to me at time t, and I then have a headache, then it is in a token of a state with the functional role of my headache. But, Block argues, we can easily imagine that such an entity feels nothing like what we feel during headaches; indeed, we have strong intuitions that such a being would not be sentient at all. If so, however, we have just conceived of a being that satisfies the functional definition of 'pain', yet has no qualia. Thus, if Block is right, and we cannot explain away these intuitions, it is the non-human cases that provide the best argument against the adequacy of a functionalist account of sensation.

There are two different ways of responding to Block's case. One is to accept it as a counterexample, and attempt to modify our outline of a functionalist theory in a way that preserves its topic-neutrality and its capacity for multiple realizations, yet prevents the population of China from counting as a sentient being. The other is to deny that the content of what we have imagined, in Block's case, actually is best described as a case of an entity who satisfies the functional definition of 'pain' but feels nothing qualitatively like our last headache.
There are, on the face of it, modifications of functionalism that may well deny sentience to those entities that we believe to be incapable of being in states qualitatively like ours. One such may be viewed as an attempt to capture the intuitions behind the solution that Block himself offers to the problem raised by his case. Block, however, proposes not a modification, but a rejection of functionalism: He suggests that:

each pain is actually a composite state whose components are a quale and a functional state. Or what amounts to the same thing, each pain is a quale playing a certain functional role. If this view is right, it helps to explain how people can have believed such different theories of the nature of pain and other sensations; they have emphasized one component at the expense of the other. Proponents of behaviorism and functionalism have had one component in mind; proponents of private ostensive definition have had the other in mind. Both approaches err in trying to give an account of something which has two components of quite different natures. ... [p. 309]

He adds, in a footnote, that "the quale may be identified with a physico-chemical state." [p. 323] But this suggestion, although it avoids the threat of dualism implicit in the quoted passage, will encounter the very same problems that beset the original psycho-physical identity thesis. That is, it leads to an implausibly chauvinistic view of what counts as a sentient being: Not only must such a being share our functional organization, but it must share our physico-chemical properties as well.
Though this result is welcome in that it rules out computers and humunculi-headed beings as sentient, it is unwelcome in that it rules out certain beings such as Martians and silicon-based life forms who, intuitively, we would regard as sharing our sensations.

Yet the intuition behind Block's proposal is worth exploring. It does seem as if we want to withhold sentience from the entity composed of the population of China because the state that realizes pain in it is sufficiently different from the state that realizes pain in us. But, as in the case of the identity thesis, it is hard to pin down just what it is about our own states that is significant without succumbing to human chauvinism. I would thus like to suggest a modification of functionalism that attempts to satisfy the intuition that computers and homunculi-headed simulations are not enough like us to be sentient, without having to identify qualia with specific physico-chemical, or even biological, properties: indeed, without having to mention, in the functional theory, anything about the internal composition of its realizations at all. I believe that this view is ultimately unsatisfactory, but it is worth describing briefly in order to see just how it is to assuage our intuitions, and just where it fails.

I shall call this view "evolutionary functionalism", and as might be suspected, it takes the evolutionary
origins of entities into account in providing "functionalist" definitions of their sensations. In doing so, it retains the purely relational, topic-neutral characterization of sensation provided by a strictly functionalist view. Let us attempt a [crude and provisional] definition of 'pain' as follows:

[P] x is a pain if and only if x is a token of

some state of some entity that has the appropriate causal role and that evolved

as a link between injury and avoidance in response to environmental pressure.

So far, this definition is extremely vague. Some of the vagueness is deliberate, although some is, unfortunately, of the promissory sort that infects most explanatory appeals to evolutionary history. The deliberate vagueness occurs in the clause that refers to the environmental history of the being. I do not describe it as the unspecific "response to environmental pressure" merely out of ignorance of the details [although the ignorance is certainly there], but leave it open to insure that entities with different environmental histories that have nevertheless forged a bond between injury and avoidance can be said to feel pain the way we do. Thus, if the Martian or the silicone-based life form adapted to their environmental pressures as we have to ours, they may be considered as among the sentient. However, this definition does rule
out the computer and the population of China as entities capable of feeling pain.

But this definition may rule out too much: It seems as if a molecule by molecule reproduction of me, though it did not share my evolutionary history, would nonetheless feel pain. Can evo-functionals be modified to account for such a case? Perhaps we could count as sentient not only entities that actually had a particular environmental history, but also those who were enough like entities that had. This strategy, of course, bears the burden of having to state in what respects an entity must be like one with the relevant evolutionary history: I suspect that the spelling out of this condition might involve the same sort of chauvinism that I criticized above. If so, is our choice between one or another overly chauvinistic quasi-functionalist view?

I do not think so. I believe that the non-human challenges to functionalism are best met by pursuing the second strategy that I mentioned; that is, by denying the claim that what we have really imagined, in Block's case, is an entity that satisfies the functional definition of 'pain', but does not feel anything at all. 7 If my earlier remarks about the human case are correct, then it is easy to argue that, in Block's case as well, we have not in fact imagined the population of China as actually having produced a token of a state that satisfies the
functional definition of 'pain'. We have not done so because, just as in the human case, we are not paying attention to, and are probably not even aware of, many of the details of that definition. Even worse, because of its scale and its recognizably human components, it is hard to get an imaginative grip upon what it would look like for the population of China to duplicate even a crude approximation of my functional organization. Both these considerations support the claim that we have not in fact conceived of the situation that Block purports to have described.

However, as will be remembered, my diagnosis of the intuitions behind the putative human counterexample to functionalism had an important consequence, namely, that if we ever were able to fully and attentively imagine that case as satisfying the functional definition of 'pain', it would then seem intuitively unsatisfactory to describe it as a case of someone unable to feel pain. It would seem, then, that my analogous diagnosis of what we are (and are not) imagining in Block's case must have an analogous consequence: if we came to know the details of our functional organization, and were indeed able to imagine the population of China as duplicating it for some period of time, we should find it intuitively preferable to describe the content of our imagination as a case of an entity that does feel pain. But this consequence
might strike some right now as being even more counter-intuitive than the consequences for the human case: as Block argues, we have strong seemingly independent intuitions that no being composed in such a way could be sentient, no matter what its functional properties.

However, I believe it is worth asking, especially in light of our previous considerations about how our background beliefs influence our metaphysical intuitions, just what lies behind our strong intuitions that no such being could feel pain. I believe that such intuitions reflect a certain pessimism about our future willingness to treat such entities, because of their physical oddity, as capable of having a mental life sufficiently like ours. This pessimism, however, seems to be based less upon independent considerations about what sorts of things can or cannot be sentient than upon our relative inability to picture the population of China as producing tokens of states that have a certain functional role: it seems quite implausible that large entities, or entities in which processing takes a long time, or entities whose central nervous systems are run by homunculi are entities that for those reasons could not feel pain. Thus I believe that a functionalist's best strategy is to assert that if we are really able to conceive of a being that duplicates my functional organization, then we will no longer have doubts about its sentience. To assert this, to be
sure, is to make a claim about our future intuitive appraisals of cases that may well turn out to be false. But I believe that the assertion inherits much plausibility from our discussion of what was going on when we tried to imagine a human who satisfied the functional definition of 'pain' but had no qualia, and retains that plausibility once it is recognized that, in the China case, it is just much harder to imagine the scenario that must occur in order for some part of the population of China to have the functional property pain.

I acknowledge, however, that this view is not yet shared by many philosophers. Indeed, even among those most sympathetic to functionalism are those who regard absent and inverted qualia to be possible, and to provide inescapable difficulties for the doctrine.

What I would like to stress now is that most philosophers who have pointed out and taken seriously the possibility of absent or inverted qualia have seen no equivalent problem for the functional definitions of intentional terms such as 'belief' and 'desire'. For example, Shoemaker in "Functionalism and Qualia" says, "I think that where the view that mental states are 'logically' or 'conceptually' connected with behavior has its plausibility is in its application to such states as desires and belief..." And in "Troubles with Functionalism" Block says, "It is very hard to see how to make sense of the
analogue of spectrum inversion with respect to non-
qualitative states...qualia seem to be supervenient upon
functional organization in a way that beliefs are not..."[p. 305]

Why do these philosophers regard there to be a dif-
ference between the two cases? There seems to be a number
of different intuitions at work. First, we are less
committed to certain epistemic principles about our beliefs
and desires than we are to similar principles about sensa-
tion. For example, though it is sometimes claimed that
we, and only we, have indubitable access to our own sensa-
tions, we often claim not to know our beliefs and desires
with certainty, or even particularly better than anyone
else. Indeed, common sense includes a concept of uncon-
scious belief and desire, phenomena whose content is, by
definition, no more accessible to the person who has them
than to anybody else. Their content is ascertained, by
all parties, only by the effects of the belief or desire
upon other mental states and behavior. This could always
serve as an alternative explanation of an allegedly "absent"
belief. Thus it is more plausible to believe that a
functional definition of these phenomena, a definition
that appeals solely to their relations with other mental
states, stimulations, and behavior, captures all their
important properties.

Another consideration is that the important properties
of belief and desire, namely, content and structure, seem to be coherently ascribable only to a system of entities. Beliefs have implications; desires have derivative desires. Thus it is hard to imagine a change or "inversion" of beliefs without imagining a substantial number of other changes of belief and desire as well, and it becomes less tempting to think of belief and desire as states whose properties we can ascertain merely by "looking within" at some particular time. As a consequence, there seems to be nothing intrinsic, or irreducibly non-relational, about the content or representational properties of intentional phenomena which a functional definition must fail to capture.

Further still, the epistemic thought experiments that generated doubt about whether sensations could be identified with functional states do not create the same difficulties for belief and desire. On the contrary, it is rather difficult to imagine an analogue to an inverted or absent qualia problem for intentional phenomena that is in the least convincing.

Although these considerations add weight to the claim that functionalist theories have a greater chance of accounting for intentional rather than qualitative phenomena, I believe that the confidence that they can give an adequate account of belief and desires is misplaced. It is true, I believe, that there is no plausible
analogue of the *epistemic* arguments—the arguments that rely upon considerations about what can be *conceived*—for the possibility of inverted or absent qualia. However, in what follows I shall argue that, for other reasons, an explanatorily adequate account of belief and desire along the lines of Lewis' functionalist theory looks as if it must attribute certain intrinsic, non-relational properties to belief and desire states. Thus if I am correct, the prima facie difficulties for a functionalist account of qualitative phenomena will recur for a functionalist account of intentional phenomena. In the case of belief and desire, just as in the case of sensation, a functionalist will probably be able to resolve these difficulties. But in this case, as I shall argue, it will be at considerable cost to our common sense conceptions of intentional states.

Thus let us turn to the problem of giving a purely functional characterization of belief and desire.
V. Characterizing Belief and Desire

0. Opening remarks

I claimed in the last chapter that intentional phenomena such as belief and desire may, contrary to common belief, provide at least as many problems for functionalist theories as qualitative phenomena. More specifically, I want to claim that certain natural accounts of belief and desire that satisfy our common sense intuitions about their important properties appear *prima facie* to attribute certain intrinsic, non-relational properties to these intentional states. Such a consequence, however, would betray the spirit of functionalism, as it would lead to a doctrine as chauvinistic as the "componential" theories of qualia discussed in the last chapter. However, certain attempts to modify a functionalist characterization of belief and desire in order to avoid this problem appear to lead to a theory which simply fails to capture the desired properties of intentional states. As I shall suggest later, there may well be functionalist characterizations of belief and desire of approximately the right strength. However, this is not obvious, and most of the attempted characterizations of these phenomena that have appeared in the recent philosophical literature run into one of the two problems mentioned above. Thus it seems as if the problem of characterizing belief and desire ought to be taken at least as seriously as
the problem of qualia. I shall offer some comments along the way about why this has not happened.

The argument of this chapter will proceed as follows: First I shall discuss the claim (made by Fodor and Lewis, among others)\(^1\) that any adequate theory of belief and desire must construe these states as relational; that is, that beliefs and desires must be construed as relations of persons to items with certain structural and representational properties. In the last chapter, I used the term "relational" to denote any theory in which mental states were characterizable purely in terms of their causal and genus-species relations to other mental states, inputs, and outputs. Here I use "relational" to mean that each belief and desire state of a person is to be described, or further analyzed, as a relation between that person (at that time) and an item that is, or has, the content of that belief or desire. Indeed, one way of putting the problem I intend to discuss is this: Can and if so, under what conditions can, a characterization of belief and desire that is relational in the latter sense be relational in the former sense as well?

Next I shall argue that the most plausible relational theory of belief and desire construes these states as relations of persons to internal sentences, or sentences in a language of thought. I shall then argue that most of the conditions so far proposed for type-individuating internal
sentences and thus, beliefs and desires, fail in one of two ways: Either they lead to an implausibly chauvinistic psychological theory, or they fail to provide an account of some of the important properties of belief and desire required by any adequate theory of mind. The major challenge for such an account of belief and desire, I conclude, is to provide a method of type-individuating internal sentences that succumbs to neither of these difficulties. The results of such a method, however, may be so unlike the sentences of a natural language that the appeal to the language of thought in analyzing belief and desire may serve more to mislead than to edify.

1. Non-relational theories of belief and desire

The goal of a functional definition of qualitative states was to provide "topic-neutral" translations, or at least truth-conditions, for our ordinary statements about qualitative states and their properties, statements such as "The pain is beginning to throb", and "Dry ice feels hot even though its temperature is low". What features of belief and desire do we expect a functional definition of intentional states to capture? Let us look at a range of ways in which beliefs and desires are invoked in ordinary explanation.

First of all, we often explain why a person did a particular thing by appeal to beliefs and desires. For example,

(1) She opened the door because she believed that it
was the door to the classroom, she wanted to enter the classroom, and she believed that opening the door would let her in.

We also make **general** claims of this sort. That is,

(2) If someone believes that acting in a particular way brings about a particular result, and desires that result, then she'll act in that way.

Other sorts of statements for which we may expect a functionalist translation describe **relations** among beliefs that a person is likely to have. For example,

(3) Most people who hold two distinct beliefs will also believe their conjunction.

Finally, we expect a functionalist theory to provide us with the means for comparing the beliefs and desires of different individuals, as in:

(4) Everybody believes that the boy Flo loves is a Romeo.

(5) Whatever Lola wants, Bill wants too.

Our question is, can these ordinary explanatory statements that invoke belief, desire, and their structural and representational properties be expressed within a Lewis-style functionalist theory - a theory in which mental states are characterized purely in virtue of their being tokens of states that have certain typical causal roles?

What kinds of generalizations comprise such a theory? Let us begin by including information about beliefs and
desires along the lines of Lewis' functional definition of sensations. For each kind of sensation, say pain, or itching, or having the after-image of yellow, there is a set of common sense generalizations about its characteristic causal interactions with other states, stimulations, and behavior. (For example, "Being hit on the head usually causes headaches.") Also recorded are the genus and species relations among the states. (For example, "Headaches are a kind of pain.")

But it seems as if we commonly make generalizations about the typical causes and effects of particular kinds of beliefs and desires, just as we make them about the typical causes and effects of particular kinds of sensations. For example, a series of generalizations of this sort may begin: "Seeing steam rising from a kettle usually results in the belief that the liquid in it is boiling; the belief that the liquid is boiling and the desire not to get burned usually results in a person's not touching the kettle..."

Yet generalizations of this sort alone do not seem to permit the expression of many of the ordinary statements about belief and desire listed above. The difficulty is this: When a Ramsey sentence is constructed from the theory consisting of the conjunction of generalizations, the name of the just-mentioned belief would be replaced by a single variable that gets bound by an initial existential quantifier. Similarly with the names of the other particular
beliefs and desires. Thus no trace of any similarity in the contents of these beliefs and desires - no indication of what occurs within the 'that' clause - will be preserved by this method of defining beliefs and desires. For example, the names 'belief that snow is white' and 'belief that snow is white and grass is green' will be assigned different variables in the construction of the Ramsey sentences, since they are names of different mental states. The same is true of 'belief that snow is white' and 'desire that snow is white'.

But since this method of defining intentional states suppresses any orthographic record of similarities in their content and structure, it would seem to inhibit the expression of ordinary generalizations about these properties, even ones as simple as that people generally believe the conjunction of any two beliefs that they entertain. Each instance of this generalization could be expressed in the functionalist theory as it stands so far, but the instances would have nothing in common.

Unless, that is, these similarities in content and structure can be identified with or reduced to properties of the causal roles of the intentional states in question. For example, perhaps there are important similarities among the causal roles of the states identified as conjunctive beliefs, and those identified as components of the conjunctions in question. If this were the case, our desired
generalization may be expressible as follows: A person who has a belief with the causal role C and a belief with causal role C' will have a belief with causal role f(C,C') where f is some function defined over states that satisfy the functional definition of belief.

However, it is not easy to see how to define such a function. It is plausible to think that the belief that P, in some cases, will have at least some of the same causes and effects as the belief that P&Q. If I believe that it will rain, for example, then (given the appropriate desires) I typically will carry an umbrella. Further, if I believe that it will rain and that it will be cold, then (given those same desires) I typically will carry an umbrella. So far so good. But if I believe, for example, that it will rain and that there will be winds over 60 mph., then I typically won't carry an umbrella. In short, "conjunctive" beliefs may not share causes and effects with either of their components. Further, beliefs that share causal properties may nevertheless have no common content. If I believe that I must return your umbrella to you, I also will carry an umbrella. But the structure and content of this belief has no similarity to the structure and content of the belief that it will rain.

It may be suggested that we look only to certain specified activities involving beliefs in order to find the similarities in role that reflect their relevant similarities
in structure and content. An appeal to the role of beliefs in inference and in evidential considerations is suggested by Harman, in his book *Thought*.

Initially, this strategy might look quite bleak, since it is far from clear that all the logical or inferential relations among beliefs could be captured by a specification of the actual use we make of beliefs in inference. We may never in fact draw certain conclusions from many of the beliefs we hold. Such an objection, however, would be misguided; a theory of human inference need not record all the vicissitudes of our actual inferential procedures in order to count as a theory in which, as Harman desires, "the valid principles of (inductive) inference are those principles in accordance with which the mind works." Rather, as Harman suggests, an idealized set of principles of inference could be extracted from the inferences that a reflective person makes in simple cases, and used to evaluate the other inferences the person actually makes. That is, we may differentiate, say, the belief that snow is white from the belief that grass is green in simple evidential or inferential situations. Then we will be able to identify these beliefs, respectively, in more complex situations, and thus come to recognize inferential mistakes by noticing that these inferences do not conform to the principles extracted from the simpler cases. If enough do not conform, however, we may be forced to modify the principles, or to study
another sample. Further, we may extend these principles to beliefs not in the original sample by assuming, in these cases as well, that similarities in content and structure reflect similarities in (simple) inferential function. Again, any egregious divergences from the principles of inference adopted on the basis of the original sample may force a modification in those principles. But if an idealized theory of inference that nevertheless has psychological reality can be developed in this way, then Harman could argue that the relevant relations among beliefs could be captured by appeal to the way they function in a wide range of simple, but carefully considered, inferences.

But even if Harman's strategy could yield a specification of the relevant relations of content and structure among beliefs, there remain difficulties for giving an account of the common content of belief and desire (and wishing, hoping, and the rest of the propositional attitudes). The only way that I can see to do so would be to point to similarities among the functional roles of the various attitudes in question. However, since beliefs and desires function so differently as evidence, in inference, and in action, it is prima facie implausible that the relation between the belief that P&Q and the belief that P is relevantly similar to the relation between the desire that P&Q and the desire that P. For example, if I believe that P&Q, then (most probably) I will believe that P, but if I desire
that P&Q, it need not be so likely that I desire that P (alone). But since we required that a functionalist theory express the common content of belief and desire, as well as the structural relations among different beliefs, it looks as if such an attempt to define beliefs and desires singly, even using this strategy adapted from Harman, will not be adequate for expressing all the generalizations about belief and desire that are important to common sense psychological theory. Such prima facie problems have motivated many to turn to a different account of the content of intentional states.

2. Relational theories of belief and desire

So far we have been examining only generalizations about the typical causal roles of particular beliefs and desires. But Lewis himself, in a brief footnote, rejects the strategy of defining beliefs and desires singly, for reasons which I have until now ignored. He says:

It may be objected that the number of mental states is infinite, or at least enormous; for instance, there are as many beliefs as propositions to be believed. But it would be better to say that there is one state of belief, and it is a relational state, relating people to propositions. (Similarly, centigrade temperature is a relational state, relating objects to numbers.) The platitudes involving belief would, of course, contain universally quantified propositional variables. Lewis' suggestion that it is the potential infinity of beliefs that forces a relational view is misleading.
He seems to suggest that if there were an infinite number of possible beliefs, and if beliefs were singly defined, then there could be no single Ramsey sentence that expressed our commonly held principles about belief. (Of course, if the number of beliefs is merely enormous, there would be no problem, at least in principle.) This follows, however, only if, as I have argued, there are not enough similarities in the functional roles of all instances of individual beliefs and desires with the same, or similar, structure and content. If there were such similarities, we could translate Lewis' "common sense" psychological generalizations that appeared to quantify over propositions into statements which quantified over belief and desire states, and specified them further using predicates from the functional theory alone. For example, instead of:

'(P)(Q) (S believes that P&Q S believes that P)',

we could say:

'(x)(y) (x bears the role in inference to y that all and only conjunctions bear to their conjuncts, then S believes x S believes y.)'\(^5\)

Thus the real motivation behind the introduction of a relational view is the suspicion, or conviction, that the generalizations we can make about the typical roles of individual beliefs and desires to not suffice to capture the structural similarities that are important for a fullblown psychological theory, particularly the generalizations
that refer to the common content and structure of different intentional states. I stress this point, as it is important to my argument later on.

Lewis' suggestion, then, is to incorporate these very generalizations into the body of common sense theory itself. These generalizations would contain, as Lewis says, universally quantified variables that range over the objects of belief and desire, entities to which we may be said to be related when in the relevant state.

But what are the items in the range of the universal quantifiers? Propositions, suggests Lewis, are just the items that have all the properties we want an object of belief or desire to have. Thus the belief that the earth is round and the desire that the earth be round will have common content in that they will be treated as different relations to the same propositions. The belief that snow is white and the belief that snow is white and grass is green would be construed, respectively, as relations to propositions which themselves share content. In short, this method of defining belief and desire would seem to yield the desired relations among the contents and structures of particular beliefs and desires.

Why should it be obvious that propositions are the proper objects of belief and desire? So far, they have been better suited to the purposes of a theory of mind than were other candidates for the objects of belief and desire.
Among such candidates in earlier relational views were objects in the world, events and properties. Russell, in his "Lectures on Logical Atomism" documents and criticizes a number of these earlier views.

For example, theories in which beliefs were taken to be relations of people to facts or to objects and properties in the world could not provide an account of false belief, or of beliefs about non-existent objects. Theories of the second sort, moreover, were not adequate even to account for all the properties of true belief. Take, as its most sophisticated version, the view that beliefs are relations of people to ordered n-tuples of objects and properties. Thus, my belief that the earth is smaller than the sun will be analyzed as a relation of me to the ordered triple, the earth, the property smaller than, the sun. This analysis has a number of advantages; it permits the distinction between the belief that the earth is smaller than the sun, and the belief that the sun is smaller than the earth, and it distinguishes the belief that the earth is smaller than the sun from other beliefs about the relations between the earth and the sun that are co-extensive, but not co-intensive with being smaller than. However, this view, unmodified, would yield no distinction between the belief that the earth is smaller than the sun and the belief that the third planet from the sun is smaller than the sun, given that 'the earth' and 'the third planet from the sun'
denote the same object. We would want this distinction, as the beliefs in question may well have differential effects upon behavior, but it is hard to see how to get it on this approach. And although certain modifications of this view—for example, the claim that what one is related to is the object the earth, the property being the earth, and the property being smaller than the sun, may well provide an adequate account of belief, it is equivalent to certain versions of the view that belief is a relation of people to propositions. Indeed, it seems that any view that is too weak to be suspected of being equivalent to a propositional view would not make the proper distinctions among the objects of belief and desire.

Similar difficulties arise if beliefs are construed as relations to mental images standing alone, without, say accompanying descriptions. Any mental image that can be formed of an object or a situation will have a number of determinate properties: shape, size, arrangement, color. But the belief that, say, an object has the represented shape will be indistinguishable from the belief that the object has the represented color. Here again, certain intuitively important distinctions among beliefs that are necessary for a full account of the influence of intentional states upon behavior will be lost.

These considerations, to be sure, have accrued to the attractiveness of treating belief and desire as relations to objects with propositional content. But propositions
themselves, (whether or not they are given a possible worlds analysis) are not the only such objects. Other proposals such as Davidson's⁶ and Quine's have involved the treatment of belief and desire as relations of people to sentences in some specific natural language. On such a view, both I and a monolingual German or Chinese would be construed as having the same relation to the English sentence, 'The earth is round' when conditions were such that we could be said, by the English-speaking theorist, to believe that the earth is round. (These conditions may include the disposition to affirm that English sentence or its translation into German or Chinese, but there is no explicit reference to translation on this view.) Such a view, however, has been found to be implausible for the following reasons. First, it is often true that the belief that P is vaguer than its sentential analogue, the belief-true, 'P'. For example, "He found it two blocks from the doctor's office" and "It was discovered by him two blocks from the physician's" might be two equally good descriptions of the content of my belief. It also seems unsatisfying to talk of beliefs and desires as relations to sentences in a particular language for those who don't speak that language, and even worse to talk this way of those such as animals and children who do not speak at all. Quine rightly points out, in "Propositional Objects", that such talk is not contradictory, only infelicitous or odd. Nevertheless, this view, at least on the
surface, is much less satisfying than the view that belief and desire, as Lewis among others suggests, are relations to propositions.

On the other hand, another account of belief and desire as relations to sentences in natural language could be attempted along the following lines: The monolingual Chinese and I may be said to have the same belief if and only if he or she had the disposition to assent to a sentence in Chinese that was a translation of the sentence in English to which I had a disposition to assent (in certain specified circumstances). In this case, I would be related to a sentence in English, the Chinese to a sentence in Chinese; so far no implausibility. But this account does make explicit reference to translations in the condition that determines sameness of belief among those who speak different languages. And it is unclear whether the intuitive constraints upon sameness of belief can be satisfied without explaining "S's utterance is a translation of R's utterance" as "S and R bear a relation to the same proposition P". Thus this view looks to be a satisfactory account of sameness of belief only insofar as it is equivalent to the view that beliefs and desires are relations of persons to propositions. Thus the view that the objects of beliefs and desires are propositions seems to best fulfil the requirements of a theory of mind.
Many philosophers, nevertheless, express reluctance to admit generalizations that require the quantification over propositions into a theory of belief and desire. Even if we provisionally shelve the worry about whether these generalizations are part of common sense, or common sense slightly modified, other difficulties remain. These regard, first, a general worry about propositions: What sorts of objects are they; what are their identity conditions? Second is a worry about how and why it is that a person's relation to a proposition - an abstract entity - should be efficacious in the production of behavior.

It seems that Lewis would want to assuage these worries by developing the analogy between belief and centigrade temperature that he suggests in the passage I quoted above. Centigrade temperature is an always present, but often varying state or property of an object. It may be functionally defined as that state of an object which has a certain role in, for example, producing various effects when the object is brought into contact with a column of mercury. Objects, however, have specific temperatures; the centigrade temperature of an object at a time is to be the value of a function from objects, at given times, to the real numbers such that the number n is the temperature in degrees centigrade of the object at the time. The function itself must yield values that obey the constraints of the theory of temperature; for example, objects that have different effects
upon a column of mercury must receive different values.

Likewise, Lewis might suggest, belief may be functionally defined as that state or property of people that leads to distinctive sorts of inferences and actions. It, like centigrade temperature, is to be a relational state, definable as a function from persons at particular times to propositions, such that the proposition $P$ is the content of what the person believes at the time. Thus the value of the function for a person at a time reveals the specific belief that he or she is said to have. Further, just as objects may bear different kinds of relations to real numbers, so may people bear different kinds of relations to propositions. Having a centigrade temperature characterizes one sort of relation of objects to numbers; having weight and longitudinal location characterize other, distinct, relations. Similarly, we may characterize desiring and hoping as relations, distinct from belief, to the same set of objects, propositions.

On the strength of this analogy between the theory of belief and the theory of centigrade temperature, Lewis could respond to the above mentioned worries about propositions as follows: Propositions, to be sure, are abstract entities, but so are numbers. Yet we do not worry about construing the temperature of an object as a relation to a number, and we agree that the temperature of an object, though it is a relation of that object to an abstract entity, can account for various important properties and effects of
objects with that temperature. Thus we should disregard parallel worries about the theory of belief and desire. The strength of Lewis' argument, however, depends upon the strength of his analogy. I believe that a closer look at the cases will show the analogy to be lacking, and further, if apt at all, to support an alternative relational account of the content of belief and desire.

Lewis is correct in claiming that we appeal to the temperature of an object in explaining various phenomena in which it is involved, even though temperature may be construed as a relation of objects to real numbers, abstract entities. But it is not merely the abstractness of propositions that is at issue in the case of belief. We have a well-developed theory of numbers, their properties, and their interrelations, but we have as yet no analogously well-developed theory of propositions. Further, the need for numbers is motivated independently of the theory of centigrade temperature, while this is not clearly the case for propositions and the theory of belief and desire.

Yet there are difficulties more serious than this. If, as Lewis suggests, the potential infinity of beliefs were the prime motivation for the move to a relational account, then it would be easy to define a function that mapped people in belief and desire states (finitely specified as beliefs and desires by the Lewis-style generalizations that appear to quantify over propositions) onto propositions.
that would specify the content of those states. States with similar functional roles would be mapped onto propositions with similar content. However, in such a case one would not theoretically need such a mapping, since we could translate the generalizations that appeared to quantify over propositions into generalizations about the similarities in the functional roles of specific beliefs and desires, individually defined.

But recall that infinitude was not the only reason for a move to a relational view. Instead, the move was motivated by the suspicion that there were not enough similarities in the functional roles of beliefs, desires, wishes, and hopes that P to account for what we think to be the common content and structure of these mental states. The existence of a function that mapped all these states onto the proposition P would serve to explain something inexplicable by reference to the functional roles of these states alone - their common content.

At this point, the resistance to the claim that human behavior can be explained by appeal to a relation of persons to propositions is not to be diagnosed as a fear of abstract entities, or even as an aversion to propositions in particular. In the case of centigrade temperature, our belief that an object's relation to a number is explanatory depends heavily, I think, upon our conviction that similarities and differences in the relation of that object to
numbers at different times are underlain by structural similarities and differences in that object at those times. More strongly, we believe that distinct objects that bear relations to the same number at a time have something structural in common. But if there must be structural similarities among individuals who are related to the same proposition (or, more weakly, structural similarities within an individual each time he or she is related to the same proposition), then what need have we for propositions as objects of belief and desire? Why not construe the entities—brain states and events—that enjoy the structural similarities in question as the objects of intentional states themselves? This, if anything, is the consequence of Lewis' analogy of relational mental states with centigrade temperature: it provides the motivation for an alternative account of the objects of belief and desire.

Indeed, there is in the literature on this topic, such an alternative account of the objects of belief and desire, advanced for reasons independent of the ones adduced in the last paragraph. This view, proposed in different forms by Fodor in The Language of Thought, Harman, in Thought, and Field, in "Mental Representation", involves the claim that beliefs and desires are relations of people to neural structures interpreted as having sentential content; that is, to sentences in an internal language, or language of thought.
In the next section, however, I shall argue that the apparently promising hypothesis that intentional states are relations to sentences in an internal language will, at worst, lead to implausibly chauvinistic psychological theories, and, at best, serve as a misleading metaphor which distracts philosophers from the problem of characterizing belief and desire.

3. The language of thought: A neural individuation of internal sentences

All versions of the language of thought hypothesis suggest that the belief that P and the desire that P are best construed as relations of a person to something in his head. The objects in the head are to be neural structures or events, interpretable as sentences in a system of internal representation, or language of thought. According to this view, a person believes that P at t only if a token of that same sentence occurs in a different way, or place, in his brain. This view, like the propositional view of Lewis', is relational, and the objects of belief and desire have propositional content. Yet this view, unlike Lewis, does not require that a relation of persons to abstract objects whose properties are unclear play such a pivotal explanatory role.

This view, moreover, seems to have further advantages. The move to put the objects of belief and desire inside the head not only yields a more plausible framework for the
explanation of an agent's current behavior, but might provide an account of the mechanisms involved in certain diachronic features of belief and desire as well. An example would be the relation between a person's enduring beliefs (for example, the long-held conviction that Oswald did not act alone) and the occurrence to him of those beliefs at particular times (say, when passing a bookstore that displays copies of the Warren Report). If some neurological structure could be identified as the sentence "Oswald did not act alone", we could cite its formation and storage to explain that person's coming to have that belief, and retaining it. The occurrence of particular thoughts could be explained as the activation of that same structure by various stimuli. Further, the existence of such internal sentences might also provide a mechanism for the occurrence of such phenomena as the fading and strengthening of beliefs and desires. These claims, of course, are highly speculative. However, they add some weight to the claim that a language of thought hypothesis provides the most promising way of supplementing the body of common sense or common sense plus psychological generalizations in order to permit the expression of relations among the contents of intentional states.

I want to stress here that in the following discussion I shall evaluate the internal language hypothesis only with regard to its adequacy as a theory of the objects of belief and desire. I shall not hold it responsible for an
explanation of the meaning relations among sentences in natural languages; indeed, I want to leave open the possibility that a semantics for natural language may have to be given before, or in tandem with, a theory of meaning for the internal sentences. Thus the advantages of an internal language theory over a propositional account of the content of belief and desire regard its greater plausibility as a part of a materialistically acceptable psychological theory. As we have seen, this theory shows great promise as an account of the differential causes of behavior, and suggests an account of additional psychologically relevant features of belief and desire, such as the ones mentioned above.

In the literature on this topic, there is a lot of controversy about just what semantic and syntactic properties the internal sentences must have in order to explain regularities in the acquisition of beliefs and desires and their differential effects on other mental states and behavior. It is clear, however, that one thing that must be specified is the relation that holds among different tokens of the same internal sentence. What is wanted, according to our common sense desiderata, is a theory that insures that Smith's belief that P at t, his desire that P at t', and Jones' belief that P at t'' all account as relations to the same sentence. Also, the sentence to which Jones is related when she believes that P must be, somehow, part of her belief that P&Q.
What, then, must all tokens of the same internal sentence have in common? One natural guess (and one that suggests itself on a reading of Jerry Fodor's *The Language of Thought* or "Computation and Reduction"), is that there must be some neurophysiologically salient feature that all tokens of the same internal sentence share. This seems especially plausible, since an appeal to the intentions and decisions of language-users, though appropriate in determining when we have different orthographies for the same natural language, is unavailable as an explanation of what characterizes internal sentences of the same type. Only a constraint such as this, it might be thought, will permit the talk of a language of thought to be construed as more than metaphorical; only then will the internal language hypothesis have provided an explanation of the common content and structure of various beliefs and desires that P, and an account of how different individuals can have the same beliefs and desires. Only then will this be a legitimate alternative to the propositional view. Further, there must be systematic structural regularities among the tokens of different but related sentences, just as there are among the sentence tokens of a natural language. Otherwise, it will not make literal sense to talk of the simple components of complex intentional states.

These requirements, however, are taken by many who are hostile to the internal language hypothesis to completely
undermine its plausibility. The argument goes as follows: To say that exactly the same neural events (or events similar enough to determine sameness of sentential structure) must occur in the brains of all individuals who can be said to believe or desire that \( P \), and that such events must, in turn, occur as components of each belief that \( P \& Q \), is to tie the existence of beliefs and desires to a highly specific and extremely dubious hypothesis about human neural structure. Thus, any theory that entails such commitments should be rejected.

This argument, however, is much too quick, and assumes a calcification of our current ignorance of neurophysiological law. As Fodor often points out, we do not yet know what will count as a neurophysiological type; we do not know under what description entities will satisfy the laws of neurophysiology. Thus there may turn out to be quite a bit of variation among states that count, from the point of view of neurophysiology, as members of the same kind; perhaps enough to account for why individuals with neural differences may nevertheless be related to the same internal sentences.

Nonetheless, the demand for any neural description at all that is to be true of my belief that \( P \), my desire that \( P \) and your beliefs and desires that \( P \) may still be regarded as an implausibly strong empirical commitment for a theory of belief and desire. The question arises, then, as to
whether the neural constraints can be weakened without relegating the notion of internal sentence to the realm of mere metaphor. The answer, intuitively, seems to be an emphatic yes. Consider, for example, what might be said in the face of the following hypothetical neurological finding. Suppose it were discovered that every time Jones could be said to desire that P, there occurred in her head a neurological event of a certain type, and every time she affirmed a desire that P\&Q, a neural event which included the former occurred in her brain. However, every time she affirmed a belief that P there occurred an event of a different type, even though this type of event occurred as well as a component of the neural state which accompanied the affirmation of P\&Q. That is, although the neural events were different in the case of belief and desire, the structural relations among the objects of belief were isomorphic to the structural relations among the objects of desire.

We could say, in this case, that internal representation occurred in more than one language: a language of belief, and a language of desire. However, I doubt that we would. Instead, I believe that a finding such as this would provide grounds for giving a more abstract, or even perhaps a disjunctive characterization of the sentences of a single internal language, if no more abstract yet neurally salient feature of those beliefs and desires could be
found, rather than force the conclusion that the language of thought is composed of many structurally identical sub-languages. Similarly, the discovery that the inputs from different sensory modalities were neurally represented in different ways would also provide the same sort of grounds for considering these sets of representations to be merely notational variants of one another, and thus part of the same language.\textsuperscript{8}

If this hypothetical finding supports the intuitions that the type-identity conditions for neural tokens of the same sentence may be even \textit{looser} than the identity conditions for neural types themselves, then perhaps some of the initial implausibility is removed from the internal language hypothesis. Nonetheless, on this view the internal sentences are still to be type-individuated in virtue of certain specific, though perhaps abstract or even disjunctive, neural properties.

Thus, even on this \textit{liberalized} account of the type-individuation conditions for internal sentences, and thus for beliefs and desires, the question arises: On what grounds can this theory count as a \textit{functional} theory of intentional states? Prima facie, the theory does not meet the constraints of functionalism at all. So far we have, like Lewis, added to our common sense theory certain generalizations about belief and desire that make reference to commonalities among their structures and contents,
for example:

1. \((P)(Q)(S \text{ believes that } P \text{ and } S \text{ believes that } Q, \text{ then } S \text{ believes that } P&Q)\).

2. \((P)(Q)(S \text{ believes that } Q \text{ if } P \text{ and desires that } Q, \text{ then } S \text{ will do } P)\).

In the range of the quantifiers, of course, unlike Lewis' view, are not propositions, but sentences in the language of thought. It is hoped by proponents of the theory that these generalizations capture and exhaust the lawlike relations among beliefs and desires (and other mental states, input, and behavior) that involve their structure and content. Indeed, this is what the list of added generalizations was designed to do: The relational intentional states will be functionally specified as just those states that together satisfy all the relevant psychological generalizations of this sort. But also important to psychological explanation, as well as to the common sense rationalization of behavior, is the appeal to a particular belief or desire of a person in explanation of an action performed. How, then, do we say specifically what a person believes or desires at \(t\)?

According to a classical functionalist, S's belief that \(P\) at \(t\) ought to be characterizable as the state that has certain (typical) causes and effects (and genus-species relations) in \(S\) at \(t\). A person ought to count as believing
that the earth is round at t if and only if he is in a state that was acquired in any of a number of ways (perception, reading, etc.) and gives rise to certain other beliefs, desires, and behavior (desire to sail west from Spain to get to China, the utterance, given the appropriate desires, of "The earth is round."). But, according to the language of thought hypothesis, no state counts as a belief that the earth is round unless it involves the occurrence in S at t of a token of the internal sentence, "The earth is round." That S realizes a Lewis-style functional theory makes certain of his internal states beliefs and desires. But S's having the specific beliefs and desires that are thought to be casually efficacious in particular cases depends upon another feature of his internal states and processes, the existence of which is not obviously insured by a functionalist specification of the typical causes and effects of that state. That is, this account of belief and desire requires not merely the specification of their typical causal roles, but also reference to an intrinsically characterized feature of those states as well. The reference to the intrinsically characterized internal sentences is necessary, according to this view, in order to properly individuate intentional states and to account for the common content of belief and desire.

But this sort of account is one that functionalists were anxious to avoid in the definition of sensation, as
will be recalled from our discussion of qualitative states. The suggestion that "each pain is actually a composite state whose components are a quale and a functional state... or what amounts to the same thing, each pain is a quale playing a certain functional role," was considered by Ned Block to be an alternative to a purely functional characterization of sensation. The apparent need for such a componential account of sensation was regarded as a failure for functionalism. Yet many philosophers affirm, in the same breath, a functionalist account of belief and desire and the internal language hypothesis. The language of thought hypothesis is not regarded as an alternative to functionalism, but merely as an alternative to the thesis that beliefs and desires are relations to propositions.

Even worse, while the componential account of sensations suggested by Block had common sense intuition on its side, ordinary intuitions militate against such a componential account of belief and desire. Indeed, one could use, this time in the service of functionalism, an argument of the same form as that used against a functionalist account of sensation. After all, it seems quite conceivable that an entity could be in a state that has the typical causal role of my belief that the earth is round, and yet be in relation to a different internal sentence, or an "ungrammatical" neural string, or worse, have no neural structure at all. If so, we have a case in which
a state that functioned like a belief that the earth is round nevertheless failed to be one; that is, the above conditions for the type-individuation of internal sentences permit the generation of an analogue of the absent qualia problem for the case of belief and desire.

Indeed, not only is such a possibility conceivable, but it is empirically likely. It seems quite possible that the neural structures that functioned, respectively, as the belief that snow is white and the belief that the earth is round in my brain could function differently in yours. And in the brain or other "mental repository" of a Martian or a silicon-based life form, these functions could be fulfilled by items very different from any set of neural structures. Thus while the loosening of constraints upon the type-individuation or internal sentences suggested above go some way towards weakening the physiological commitments of an internal language view, enough remain to keep the view rather badly chauvinistic.

Yet while the possibility of absent or inverted qualia intended to cast doubt upon the adequacy of functionalism, and thus to justify, in some sense, the chauvinism of its alternative, the possibility of absent or inverted sentences casts doubt, rather, upon the hypothesis that beliefs and desires are relations to sentences in a language of thought. Common sense, I believe, would support the attribution of beliefs and desires to entities that shared my functional
organization, regardless of whether they had linguistically interpretable neural states like mine as well.

This is a predicament, however, of the functionalist's own making. Our intuitions that absent qualia were possible were the products, I argued, of certain unexamined, and not entirely reliable, metaphysical intuitions. In the case of belief and desire, however, the introduction of internal sentences has been proposed by functionalists who believe, for reasons untinged by such intuitions, that the structure and content of intentional states cannot be reduced to features of their typical causal roles.

Even if, contrary to my suggestion in the last chapter, absent qualia were possible, many theory-oriented functionalists would not be unduly worried, since all the distinctions among qualitative states that are important for stating the conditions under which they occur, and their effects on other states and behavior would be preserved by a purely functional characterization of sensation. The unaccounted for "raw feels", according to these philosophers, could simply be regarded as outside the domain of psychology. Though there would still be a mind-body problem (or a qualia/physico-chemical state problem, as on Block's view), no behavioral regularities that should be explicable by psychological law would remain unexplained. This appears not to be so, however, for generalizations about belief and desire. Here it looks as if no functional specification
of their typical causal roles will capture all their explanatorily relevant properties. It seems then that prima facie there can be no easy retreat back to a non-relational, purely functional view.

Why, then, is a componential account of qualitative states regarded as more damaging to the functionalist spirit than a componential account of belief and desire? There are, I believe, three reasons. The move to a componential account of intentional states goes unnoticed, or unremarked, because it does not signify; for functionalism, a forced retreat in the face of staunchly held common sense intuitions about the mind. Indeed, there may seem to be little discontinuity between the views to those with theoretical interests. The move to a componential theory was forced by the requirement that an adequate theory of belief and desire account for the explanatorily relevant structural properties of these states. But this requirement comes not from common sense but from theory, arising from the view that there must be a lawlike, physicalistically acceptable account of observable behavioral regularities. Internal sentences are theoretical entities hypothesized to account for these behavioral regularities in terms of differences and similarities in the content and structure of belief and desire. Qualia, on the other hand, are commonly believed to be essential properties of sensations; further, they are arguably there (or not there, as the case may be) before
the mind.

There is an **ontological** issue as well. Although neither a componential account of qualitative nor intentional states counts as a functionalist theory proper, it might seem as if the latter, but not the former, is at least physicalistically acceptable. That is, if one adopted a view in which beliefs and desires were defined in terms both of causal role and internal sentential component, one's only ontological commitments would be to functional states and neurophysiological entities. There is no controversial internal sentence/neural structure identity thesis, since internal sentences were hypothesized to be neural structures of certain kinds. However, one who accepted a componential view of qualitative states would still have to account for the ontological status of the **qualia**, a project that revives difficulties for the physicalist of the sort initially encountered by Place and Smart, and thus threatens to involve the commitment to dualism.

Even if we reject all conceptual arguments against the identity thesis, however, there remains for both kinds of componential view, the (quasi-) **empirical** argument against the identity thesis that originally motivated the turn to functionalism. This difficulty concerns the **chauvinism** (as described above) of a componential view. As in the case of qualitative states, if beliefs and desires
involve internal sentences which are to be indentified with neural states, then only entities with the relevant neural states will count as having beliefs and desires. This appears to force an overly tight restriction of the class of entities to which mental predicates may apply.

Along with chauvinism goes another empirically untoward consequence of a componential view, which I shall label the non-autonomy of psychology. The explanation of behavior by appeal to mental states is thought to be the proper subject matter of psychology. According to a componential view, however, we would have to consult a lower-level science — neurophysiology — in order to produce explanations of the effects of beliefs and desires upon behavior. This would mean that the contents and structures of beliefs and desires, like qualia (if the componential theory of qualia were correct) would be outside the domain of psychological explanation proper. This betrays the aims of those who hope to identify mental states with states for which psychological regularities hold.

But here too is the promise of a difference between the internal language hypothesis and the componential account of qualitative states. It seems as if there may be some leeway in the characterization of the internal sentences that is not available in the identification of qualia with physico-chemical states. Perhaps an acceptable
way can be found of individuating internal sentences that sufficiently reduces the prima facie chauvinism of the view, yet retains its explanatory power. And indeed, though we found a neural individuation of internal sentences to be chauvinistic, the considerations that led to this conclusion have prompted the question: Can we do for the sentential components of belief and desire what the functionalists attempted to do for the belief and desire states themselves, namely, give a functional account of what it is to be a token of the internal sentence P? That is, could we identify as tokens of the sentence P all and only those structures that functioned in certain ways in belief and (perhaps other) ways in desire within a particular individual? If so, then the intrinsically characterized objects of belief and desire could vary as to physical type among different individuals, as well as within an individual over time, and the internal language hypothesis might escape the charge of chauvinism.

In the recent literature on this subject, there have appeared other versions of the internal language hypothesis which attempt just such a functionalist individuation of internal sentences. Gilbert Harman, in Thought, and Hartry Field, in "Mental Representation", are two who offer such views. If either view suffices to alleviate the above-mentioned difficulties for a componential account of belief and desire, then there is good reason
for the lack of attention that its prima facie difficulties have received.

Before turning to an examination of these views, however, I would like to offer a caveat. I claimed that a satisfactory theory of belief and desire required an account of how different individuals could share the same beliefs and desires. This desideratum, however, is explicitly rejected by both Harman and Field. They suggest that speakers of different natural languages (and different ideolects within a single natural language) may well have different languages of thought. This is a consequence of the turn to a functional account of internal sentences: likeness of the internal sentences of distinct individuals can be no greater than the likeness of their functional organizations. Harman and Field could suggest that individuals with different internal languages nevertheless share beliefs and desires if they are related to sentences that are translations of one another, but this requires an account of translation that makes no reference to propositions. Otherwise, the view would collapse into a somewhat baroque version of Lewis'. The difficulties of getting such a theory of translation are well-known; this may explain why Harman and Field are content to give up this requirement. I too am willing to reject this requirement if an otherwise adequate theory can be formulated.
only without it. Let us see, then, whether these views are as promising as they seem.

4. The language of thought: A functional individuation of internal sentences

Harman's view in Thought is not only that functionalism is compatible with the internal language hypothesis, but that a functionalist theory of belief and desire is automatically an internal language hypothesis itself. For example, on page 58, he says, "We can simply take mental states to be instances or "tokens" of appropriate sentences", and on page 59, "A functional account of mental states and processes is an account of the language of thought, since mental states are instances of sentences in the language of thought."

It is not completely clear, from Thought, what Harman's proposal is, but one way of reading it is this: The theory of belief and desire includes Lewis-style generalizations about these states which contain quantifiers ('(P)', '(Q)') that range over the objects of belief and desire. Also included are laws relating beliefs and desires to environmental inputs and behavior. An entity realizes the theory if it has in it states and processes that interact in the ways prescribed. And, as for any other functionally characterized states, belief and desire may have different realizations in
different individuals.

But on Harman's view, like Lewis', belief and desire are relational states: entities that believe that P must be related to an item with the propositional content, P. Further, like Fodor, Harman asserts that an entity that believes that P is related not to a proposition, but to the internal sentence 'P'. How, then, is such a view no more nor less than functionalism itself?

According to Harman, S is related to the internal sentence 'P' at time t if and only if occurring in S at t is a token of 'P'. But a token of 'P' occurs in S just in case S is in a token of a state that has either the functional role, according to the theory, of the belief that P or the desire that P (or, the wish, hope, or intention that P). That is, S is in relation to the internal sentence 'P' just in case S is in a token of a state that bears the relations to inputs, other states, and outputs characteristic of either belief or desire, including the constraints upon the interaction of belief and desire that must be expressed with the aid of the Lewis-style generalizations that include propositional variables. Thus, for S to be in relation to the internal sentence 'P', it is sufficient for S to be in a token of one of a number of functionally characterized intentional states. Indeed, any of these state-tokens is to count, itself, as a token of 'P'.

198
By virtue of the fact that in believing or desiring that $P$, $S$ bears such a "relation" to the internal sentence '$P$', the belief that $P$ and the desire that $P$ may be said to have the same structure and content. Briefly, on this strategy, we are to use the generalizations that contain propositional variables in order to identify the states in an entity that have the roles appropriate to various beliefs and desires. Then we cleverly discharge any unwanted empirical commitments by taking the values of these variables to be tokens of these states themselves.

On Harman's view, absent or inverted sentences are logically impossible, since one could not be in the state characterized as the belief or desire that $P$ without being in relation to the mental sentence '$P$'. The reason, of course, is that each token of some realization of an intentional state involving the content $P$ counts as a token of '$P$'.

It may be wondered, however, whether the commitment to any interesting version of the internal language hypothesis follows as trivially from functionalism as Harman believes. For example, one who takes the internal language metaphor seriously might expect all state-tokens of an individual that count as tokens of '$P$' to share syntactic structure. But under what description can they be said to do so? The states that, in an individual, realize the belief that $P$ and desire that $P$ have different
functional roles; thus it looks as if tokens of them cannot be said to share syntactic structure under a functional description.

It might be argued, however, that I have not taken a sufficiently broad view of the respective functions of the belief that P and the desire that P. Though it is true that the belief that P leads to actions and figures in inference very differently from the desire that P, perhaps a look at the relations among the functional roles of different desires and those of different beliefs would show an interesting isomorphism. If there were such an isomorphism, then we could define the relation of type-identity for internal sentences as follows:

(T) x is a token of the same type as y iff. x interacts with other beliefs (and mental states, inputs, and outputs) in way R or x is a desire and interacts with other desires (and other states, etc.) in way R...(and mutatis mutandis for y).

As we have already discussed, however, there is no such isomorphism between the functional roles of beliefs and desires, since the common sense and decision-theoretic generalizations about these states describe great differences in their respective functions. For example, it is agreed that if I believe that P&Q, then (most likely)
I will believe that P, although if I desire that P&Q, I may well not desire that P (if P is to occur without Q). Also if I believe that P brings about Q, and believe that P, I will most likely believe that Q, but if I believe that P brings about Q and desire that P, I will quite often not desire that Q. (This is the basis for the deterrent theory of punishment.) Indeed, belief and desire are distinguished in the functional theory by means of the differences in the generalizations true of them. Thus the burden of proof is on Harman to show that there is some broader functional description under which the belief that P and the desire that P count as tokens of the same type. Given our observations, however, the prospect looks rather bleak.

The possibility of their sharing structural properties under a neural description is certainly not guaranteed by Harman's view, either. Indeed, given Harman's identification of tokens of internal sentences with tokens of the states that have certain causal roles, he has insured that there will be at least some relevant differences among the tokens of a sentence P in an individual: Since the different intentional states that involve the content 'P' have different functional roles, they must differ somehow in their physical realizations.

It may be, of course, that in the individuals that realize a Harman-style functional theory, the belief that
P and the desire that P (and the other intentional states involving P) have neural realizations which, though not type-identical, nevertheless have certain natural, neurally salient similarities. Or, they may be type-identical relative to some neural types but not others. This would account for their functional differences, yet make more plausible the claim that they were tokens of the same sentence. Such a circumstance, I believe, would give greater credence to the view that internal sentences could be given a functional individuation that preserves the structural properties that we expect sentences in a language to have. In fact, I argued, in discussing the prospects for a neural individuation of sentences, for flexibility in sentential identity conditions under circumstances such as these. What I want to stress, however, is that if such a circumstance obtains, it will be because the theory happens to be realized by states that have certain neural structures. Nothing in Harman's statement of the theory, however, demands that the realizations have these structures.

Further, it is not clear that we have reason to expect the realizations of Harman's theory to have such natural, neurally specifiable, "linguistic" properties, especially given that different intentional states involving P have such different, unisomophic functional
roles. (We might more plausibly expect neural similarities among simple and complex beliefs (or desires) involving the same content, but we would expect so because there are similarities in the ways in which they function in inference.) Yet Harman asserts that his theory is a theory of the language of thought merely by virtue of its being a functionalist theory of belief and desire.

Some, however, would claim that any proper realization of Harman's theory must have the structural similarities between instances of the belief that P and the desire that P that I described above. The argument is this: Harman's theory is a theory of belief and desire, and the effects of their interactions upon the formation of other mental states and behavior. The description of these states as having common content is necessary in order to state important regularities in the effects of their interactions upon behavior; the states conform to these regularities by virtue of their common content. Moreover, we need to make reference to these regularities in order to assign content to individual state - that is, in order to begin to identify the neural states that count as beliefs and desires that P. Shouldn't we then expect that, within an individual, the similarities in the content and structure of beliefs and desires ought to be reflected on a neural level as well?
As suggested before, I believe that the answer need not be yes; a comparison with a different case may help to indicate why. Pain, for example, is a state that is functionally characterized by reference to its causation by certain environmental conditions, and the other mental states and behavior to which it typically leads. Various generalizations about pain are important to psychological theory: these include generalizations about the typical causal role of the state pain, and laws that show regularities in the occurrence of particular kinds and degrees of pain given various stimulations and other mental states. Now suppose (a not implausible assumption) that our psychological theory contained the following generalization:

\[(n)\text{ (Beliefs of strength } n \text{ usually require application of pain of strength } n \text{ (or } 2n, \text{ or whatever) in order to induce change of belief)}\]

If this generalization were indeed part of our psychological theory, it would help to uniquely specify pain, or at least to serve as a constraint upon which neural states could be identified as states of pain. But it seems that the inclusion of such a generalization in psychological theory would not be taken to insure that beliefs of strength } n \text{ and pain of strength } n \text{ had any neurophysiological features in common.}
But this is just our situation in the case of intentional states as well. Here we specify which belief (desire) is the belief (desire) that the earth is round partially by use of generalizations that make reference to the syntactic structure of intentional states, and partially by laws that relate sensory impingements and motor outputs to particular internal states. Neither the structure nor the efficacy of our current psychological theories, then, determine which features mentioned in the laws must be reflected by neural similarities, and which are, in some sense, merely manners of speaking. Thus it is not insured that the belief that P and the desire that P have some underlying neural similarity, even if reference to the common content P seems to be needed to express generalizations about certain interactions among beliefs and desires.

There is, in general, a tension between Harman's advocacy of a functional individuation of mental states and his proposal that we regard the states themselves as tokens of internal sentences. At one place in Thought, Harman justifies the latter proposal by an analogy with speech act theory:

Just as various speech acts - promises, threats, warnings, and so forth - can involve instances of the same sentence of the outer language, various mental states - beliefs, desires, hopes, and so forth - can be instances of the same sentence in the inner language of thought. (My italics)
But Harman has pointed to the dissimilarity of the cases by his own words. The sentence "I'll see you this afternoon" may count at one time as a promise, at another time as a threat or a warning, depending upon why it was uttered and what effects it had. Nevertheless, the structural description of this sentence need not refer to any of its possible uses.

Fodor has pointed out that a state's having a certain functional role may insure its having a certain structure, because a state may have a certain role only by virtue of having that structure. Harman, however, is saying not that structure determines role, but that role determines structure. Further he has not argued that there are any interesting similarities, concrete or relatively abstract, between the functional roles of the belief and desire that P.

Harman's internal sentences, thus, are insured no structure except that bequeathed upon them by their functional roles plus the additional stipulation that certain states with different functional roles (and different neural realizations) are to count as tokens of the same internal sentence. It is not clear, however, in Thought, just why this stipulation is considered by Harman to be justified.

And indeed such a stipulation may be criticized, both by the propositionalist and by the philosopher who takes
the language of thought hypothesis more literally, as giving the belief that P and the desire that P common content by fiat alone. For purposes of comparison, let us look once again at the alleged (and by now well-worn) analogy between belief and temperature. I claimed that the strategy of construing temperatures as relations of objects to numbers was merely heuristic and in principle dispensible because there were both structural and functional similarities and differences that underlay the similarities and differences in the numbers to which objects with various temperatures were said to be related. We need not invoke the number n in order to determine what different objects with the same temperature have in common. Water and wood, for example, of 30 degrees centigrade, though quite different in molecular composition, both produce the same rise in a column of mercury (when other variables are relevantly controlled), and they both are composed of molecules that undergo equivalent molecular activity. (Either would have sufficed to discharge the apparent commitment to numbers by temperature theory.) In Harman's theory, however, we are insured neither commonality in functional role nor in underlying neural structure.

Does this mean, then, that belief and desire psychology is saddled either the commitment to propositions or with a chauvinistic psychological theory in order to be able
to state all the interesting regularities in the interactions among beliefs and desires? No; there are two different strategies yet open to the liberal anti-propositionalist.

One method is advocated by Hartry Field, in his article, "Mental Representation". According to Field, it is possible to regard the conditions under which neural states count as tokens of the same internal sentence as an empirical issue, as follows: We are to take the vocabulary of belief-desire psychology as including various one and two place predicates such as 'x is a token of a grammatical sentence' and 'x is a token of the same type as y'. Then, included in the theory of belief and desire itself will be existence claims about such properties and relations. Any individual that realizes the theory of belief and desire, then, must have in it physical realizations of these properties and relations as well. But the discovery of what physical relation, in any individual, realizes the functional 'same token as' relation (call it 'R') must await significant theoretical advance. The question I would raise, of course (and it is one that is equivalent to the question raised against Harman), is this: Is it likely that humans - individuals to whom we routinely apply the decision-theoretic and common sense generalizations of the theory of belief and desire - do in fact
realize the supplemented theory suggested by Field?

Relevant to this issue is a certain ambiguity in my statement, so far, of Field's view. The view can be seen either as a clearer, more explicit version of a Harman-style theory - one that explicitly displays the theoretical commitments of a "liberal anti-propositionalist" - or as a more substantive alternative to Harman's view. The difference lies in whether Field believes that the neural states that realize various of S's attitudes towards P are to be identified with, or rather, are merely to have as components, tokens of the internal sentence 'P'. The consequence of this distinction for the issue at hand is this: If Field, like Harman, identifies realizations of intentional states with tokens of internal sentences, then he, like Harman, has insured that there will be some physical differences between belief-tokens and desire-tokens of the same internal sentence. Though, as in Harman's case, there could be more abstractly specified neural types instantiated by P-tokens of both beliefs and desires, this need not be the case. If not, R will be realized by at least n+1 disjuncts, where n is the number of distinct kinds of intentional states invoked by the psychological theory. The number will be larger, of course, if the relation that holds between, say, belief-tokens of P and desire-tokens of P is different from the relation that holds
between belief-tokens that $P$ and hope or intention-tokens that $P$. The number will be larger yet if there is no one relation that holds between the belief that $P$ and the desire that $P$ for all $P$.

Such a situation, then, would involve a disjunctive (and perhaps badly disjunctive) realization of those very properties and relations of mental states that were supposed to insure that beliefs and desires could be said (in appropriate circumstances) to have common content. The internal language hypothesis claims that $S$'s belief that $P$ and desire that $P$ have common content by virtue of $S$'s relation, in each case, to the same sentence of his internal language. According to the view under discussion, however, the language of thought would be a system in which "sentences" could possibly have identity conditions so baroque as to raise the question of whether it is a language at all. If the type-identity conditions for sentences in the internal language differ too radically from the type-identity conditions for sentences in languages with which we are familiar, then the internal language hypothesis becomes a metaphor that tends to mislead rather than explain.

Field's project of building into our psychological theory existence claims about properties and relations that can be trivially realized may be described as allowing
the identity conditions for internal sentences to be empirically determined. But if the theory can be true even if the predicates 'is a sentence' and 'is of the same type as' are (badly) disjunctively realized, then the language of thought hypothesis seems much less of a substantive claim.

As I have said before, it may be the case that some of the individuals that realize a Harman-style theory - perhaps humans - do have neural states that bear the desired "nice" relations to one another. In this case, one could argue that these realizations did, while the "trivial" realizations did not, have a language of thought. In such a case, the internal language hypothesis might retain its explanatory power. But one could not argue in this way and remain a classical functionalist, since one would have to make reference not just to the functional theory of belief and desire, but to certain preferred realizations as well. And such a view would leave unexplained the apparent common content of the beliefs and desires of the trivial realizations of the theory.

So far, I have attempted to examine the consequences of what amounts to a clearer statement of a Harman-style view. Earlier, however, I claimed that Field could be taken either as advocating a Harman-style view in which belief and desire states were identified with tokens of
internal sentences, or as advocating a view in which internal sentences were proper parts of the intentional states in which they were involved. In reflecting upon the theory that results from the latter interpretation of Field, we might form greater expectations for "nice" realizations of his relation R, since there is no reason why the tokens of P that occur as components of beliefs that P should not be type-identical, under all descriptions, with the tokens of P that occur as components of desires that P. This is so because, on this interpretation, it is not required that belief-tokens of P have at least some important physical differences from desire-tokens of P. Yet if Field accepts the possibility of trivial disjunctive realizations, and if I am right to claim that we would withhold the description 'language' from a system that trivially realized R, then there will be, possibly, entities that realize belief and desire theory which cannot be said to have an internal language. Thus the hypothesis that beliefs and desires are relations to internal sentences will serve to explain the content and structure of the intentional states of only some of those to whom we apply our commonly held generalizations about the interactions among the states.

I believe that my argument that there can be trivial realizations of a Harman-style theory applies to this case
as well. It may be objected, however, that my argument in Harman's case boiled down, ultimately, to the claim that our belief and desire psychology may not be literally true. And this, it might be said, is a possibility that any theory must face. My point, however, is that according to the Field-Harman method of individuating internal sentences, there will be theories that appear to quantify over internal sentences which turn out to be literally true, even though the items in the range of the quantifiers cannot be considered as having linguistic properties in any reasonable, intuitive sense. That is, that the Field-Harman method of individuating internal sentences trivializes the hypothesis that beliefs and desires are relations to sentences in a language of thought.

5. Conclusions

What, then, is the alternative for the philosopher who wishes to retain (the bulk of) the theory of belief and desire, yet is opposed both to the quantification over propositions and the chauvinism that results from a strictly neural individuation of internal sentences? One could say, I suppose, either of the following things: First, we could drop the requirement that internal sentences must be type-individuated in a way that reflects
the type-identity conditions we impose upon sentences of a natural language. This proposal seems unsatisfactory, however, since the power of the internal language hypothesis derives from the alleged similarity of the language of thought with the natural languages with which we are familiar.\textsuperscript{10} The second alternative is to drop the pretense that the sort of theory proposed by Harman or Field involves, in any intuitive sense, the relation of persons to internal sentences. Indeed, one could drop the pretense that these are relational views; that what looks to be the common content of various intentional states is to be explained by citing various relations of the relevant persons to the same thing.

It may be asked, however, whether this is even a remotely plausible alternative. After all, I have devoted a great deal of discussion to the reasons that motivated the turn to a relational theory of intentional states; it may not seem that these motivations can now be abruptly ignored. The reasons people have given include the potential infinity of beliefs and desires, the structural relations of complex beliefs and desires to their simpler components, the shared beliefs and desires of different individuals, and the common content of different kinds of intentional states. But I have already argued that potential infinity of beliefs and desires gives no reason to prefer a
relational view. Further, I have suggested, following Harman, that one might be able to translate the generalizations about the relations of complex beliefs (desires) to their simpler components which appear to require propositional variables into generalizations about the common inferential relations that hold among all instances of beliefs (desires) with those structures.

Thus we are left with the problems of accounting for shared beliefs among individuals, and the common content of belief and desire. However, Harman and Field have explicitly given up the claim that people, especially those in different language communities, can routinely be said to share beliefs and desires, and if my arguments are correct, they have implicitly given up the claim that beliefs and desires that P are different relations of an individual to the same object. That is, the remaining motivations for the turn to a relational theory are not in fact honored by these weakened internal language views.

But here I might be charged by Harman and Field as being theoretically short-sighted. On a view which claimed that the belief that P and the desire that P were relations to the same internal sentence, but that the type-identity conditions for the internal sentences were intuitively adequate, one could nevertheless state the important psychological laws that referred to common
content in describing the typical behavior-producing interactions among belief and desire. My response is this: If there are regularities that involve the common content of belief and desire, but there is not enough neural or functional commonality between the belief that P and the desire that P to justify their counting as relations to some one thing, then there ought to be another way of stating the regularities in question.

Until such a way is discovered, it would be permissible, in my view, to keep the propositional variables in the generalizations about belief and desire, while keeping quite open the option of remaining instrumentalists about some or all of the relations among intentional states that they purport to describe.

Such a strategy, I believe, would have the following advantages: As a model of our decision-theoretic and common sense psychology, the language of thought hypothesis has limited usefulness. It appears to generate more (not clearly relevant) questions than provide explanatory answers to why people believe, desire, and behave as they do. The dropping of the model would put a stop to questions about which of its features must be taken literally, and which merely metaphorically. It would also mute complaints from those hostile to all theories of internal representation that the language of thought hypothesis merely pushes explanation one
step back. And, most important for a functionalistic psychologist, it redirects attention to the distinctions that can be drawn among the functional roles of beliefs and desires, instead of suggesting that the explanations of the interactions among these paradigmatically psychological states must await any sort of verdict about the relation of psychology to neurophysiology.

What are the prospects, then, of finding a non-relational, but intuitively satisfying way of stating the relevant generalizations about the interactions of beliefs and desires? I agree that we do not have such a way now, and that the prospects may look somewhat bleak, but perhaps the following observation will prompt some optimism. At the beginning of this century, sense-data theories were seen as providing the most plausible answers to the question of what my seeming to see red and my actually seeing red had in common. The answer, of course, was that in each case, I was related to a red sense-datum. The appeal to this allegedly common component of the states in question was thought to be essential for explaining the similarities in the behavior prompted by those states. The postulation of sense-data, however, raised knottier ontological and epistemological problems than the ones it was designed to solve, and the sense-data approach to such problems was eventually given up.
I do not want to place too much weight upon this analogy. One difficulty, for example, is that the most commonly believed alternative to a sense-data view (and one that I have argued for in the last chapter) is the view that we can give a purely non-relational functionalist account of what these states have in common. But, as we have seen, it is not clear that we can do so in accounting for the common content of belief and desire. However, the strength of the analogy, in my view, is in its suggestion that it may not be so clear that we can't.

In closing, I would like to contrast the problems I have raised for a functional characterization of belief and desire with another set of difficulties that has recently received attention. I am interested here in the objection (suggested by Putnam, developed by Stich) that purely functionalist theories cannot account for a very important property of specific beliefs and desires - their references, or what they are about. To adapt Putnam's example from "The Meaning of 'Meaning'", a person with my functional organization who is now in the same state and is receiving the same sensory impingements as I, will be in the same belief state, functionally defined, as I am. However, suppose further that we are both being impinged upon by the colorless, odorless, tasteless liquid that fills the surrounding lakes and streams,
but that she and I are a great distance apart (on different planets?), and that while the stuff responsible for my current sensory impingements is indeed water, the stuff responsible for her sensory impingements is not water, but has molecular structure \text{XYZ}. It would be natural to conclude, according to this objection, that though I have beliefs about water, her beliefs are not about water but about XYZ. Therefore a purely functional definition will not distinguish these beliefs. Further, an adequate account of belief and desire must make reference to circumstances independent not merely of the functionally characterized states of an individual, but independent of any states and properties, however characterized, of that individual as well.

Most philosophers agree with the intuition that my counterpart and I have different beliefs. The issue most debated is whether such distinctions among beliefs and desires with respect to what they are about has any differential effect upon the production of other mental states and behavior. If not, it is countered (by Fodor and Field, for example) that properties such as aboutness are not necessary for the statement of psychological laws, and thus no responsibility of functionalism. Indeed, it is at issue among these philosophers whether it is necessary to invoke, in the
laws of psychology, any semantic properties of belief and desire at all.

The interest of this issue is evident, and its resolution is important, but it is an issue whose discussion depends upon the possibility of stating psychological generalizations that invoke the relevant syntactic properties of belief and desire. And this, I have claimed, might be much more difficult than it seems.
FOOTNOTES (Introduction)

1 See Ned Block, "Troubles with Functionalism," for a comprehensive account of the different varieties of functionalism.

2 See especially Chapter III.

3 James Cornman, for one, had such a view. Also, I shall claim in Chapter IV that some philosophers are needlessly driven to such a view.

4 Are functional properties of state-tokens second-order properties? It depends upon how the term 'second-order' is understood. If a first-order property is one had by particulars, a second-order property one had by first-order properties, and, in general, an n+1th-order property one that is had by nth-order items, then this functional property is not second-order. If, on the other hand, an n+1th-order property is one in the definition of which occurs quantification over nth-order properties, then the functional property described above is second-order. Because of the ambiguity of the term 'second-order', I shall attempt to avoid it in the chapters to come.

5 This terminology is not entirely standard. For example, David Lewis would object to the identification of the abstract universal pain with the property of being in pain, but I shall compare my usage to his in Chapter III.

6 Also, as Nagel points out in "Physicalism," it is not clear that dualism can account for the epistemological assumptions, either. He argues that even if mental states are taken to be states or properties of an immaterial substance, they will be states and properties that can be objectively characterized, and will thus not be, in principle, knowable in special ways by those who have them.

7 This was Cornman's point in "The Identity of Mind and Body."

8 Again, Cornman suggests this in the above-mentioned article.
Some of Dennett's arguments in *Content and Consciousness* against the identification of thoughts, etc., with central states of the nervous system that have been assigned content can be seen as arising from such considerations.

Putnam, for example, in "The Nature of Mental States" and other early articles.


Such a strategy is suggested by Lewis, and perhaps by Shoemaker.
As will be clear when the view is presented, the rubric "radical behaviorism" need not be associated specifically with a Skinnerian theory of behavior, but rather with a certain philosophic treatment of any empirically sound behavioristic view.

2 From "Have Some Madeira, M'Dear" by Flanders and Swann. Example thanks to Hartry Field.

3 As in Sommers, "Types and Ontology", and Cornman, Materialism and Sensations.

4 Putnam, "Brains and Behavior."

5 These "radical" examples are less plausible in the case of belief and desire than in the case of sensation. Perhaps it would help, in the former cases, to imagine a situation in which it is possible to have two conflicting beliefs or desires, and yet exhibit, perhaps for a lifetime, the same set of behaviors. See Dennett, "Brain Writing and Mind Reading" for a fairly convincing example.

6 Geach, Mental Acts; Chisolm, Perceiving and "Sentences About Believing"; Fodor, Psychological Explanation.

7 "brains and Behavior"

8 David Lewis, in "An Argument for the Identity Theory" argues in this way.

9 In Chapter III, I give a more detailed description of these functional definitions.

10 I mean by this a method of providing functional definitions, like David Lewis', which purports to preserve the meanings of our mentalistic terms.

11 Again, this has been pointed out by Chomsky, Fodor, Dennett, Taylor.

12 Thanks to Dan Osherson for help in developing this example.
13 I'm sure Dennett says this somewhere.

14 The above example was designed to illustrate certain points about the relation of "higher" to "lower" level theories which have been made by, among others, Fodor in The Language of Thought, and Putnam in "Explanation and Reduction". Putnam and Fodor draw certain conclusions from these points which I have not yet discussed, namely, that either the traditional constraints on theoretical reduction have to be liberalized, or psychology is not reducible to lower level physical sciences. Thus, Skinner's suspicion that mentalism is equivalent to dualism, though based upon positivistic views other than those traditional views about the nature of theoretical reduction, may have been, if premature, quite prescient.
FOOTNOTES (Chapter III)

1 Lewis's views are to be found primarily in his "Psychophysical and Theoretical Identifications" (Australasian Journal of Philosophy, Dec. 1972) and "How to Define Theoretical Terms" (Journal of Philosophy, July 9, 1970). They are foreshadowed in his "Argument for the Identity Theory" (reprinted in Rosenthal), and are alluded to in "Radical Interpretation" (Synthese).

2 Block, "Troubles with Functionalism"; Warner, "Mental States and Neural Codes" (forthcoming).

3 Fechner, Elements of Psychophysics; Stevens, (for example), "Hearing".

4 See "PPTI", and "RI", and Convention.

5 Generalizations of type (3) and their legitimacy will be discussed further in Chapter V. Here I am concerned only with the conditions under which such statements can be regarded as part of common sense. Also, it is an issue whether our commonly believed generalizations about the effects of belief and desire upon behavior and other mental states are best expressed in the form of (2) or (3). This issue as well will be discussed in Chapter V.

Also, all these examples of platitudes are to be taken as approximate, not exceptionless, generalizations.

6 Further, these particulars are not the right sorts of entities to enter into the causal relations in terms of which mental states are to be defined.

7 Convention, pp. 64-68.


9 In "How to Define Theoretical Terms" Lewis considers two proposals for what happens to the meanings of terms introduced by a theory that subsequently gets revised. On one proposal, the terms change meaning, as their meanings are to be determined by the Ramsey-sentence of the theory currently in use. On the second proposal, terms retain the meanings given to them by the theory in which they
initially appeared. Lewis claims not to like either alternative, and to believe that there must be something in between (p. 431-32). The difficulty with judging the plausibility of the second proposal has to do with an ambiguity in its presentation. Lewis says that terms in a revised theory can keep their old meanings only if the realization of the new theory is a near-realization of the old. But he implies that the theories need not have such a close relation. Thus the plausibility of this view is difficult to judge, and I shall ignore it for now. However, see note 13 for the discussion of a consequence of this proposal.

10 That is, when questions of the determination of reference are not at issue.

11 "HTD'TT," p. 432.

12 In my terminology (here and in the subsequent definitions): The property pain =df. the property of being a token of a state-type that (in some entity) has the appropriate role in T.

13 One might want to give (P') the form of (P*) instead. I do not think that it matters here.

14 On Lewis' second proposal about the meanings of theoretical terms in times of theoretical revision, (outlined in note 9) we could say something stronger, namely, that the meaning of 'pain' has remained the same. (We have assumed, of course, that the unique realization of T' is the nearest near-realization of T.)

15 The myth of Jones itself involves a two-stage introduction of theoretical terms--first, the postulation of thoughts or beliefs, and second, the postulation of sensory impressions.
In this chapter I discuss the problem of absent qualia to the exclusion of the question of whether qualia inversion is possible. Most of my arguments, however, can be modified to deal with the latter problem as well.

But see Feldman and Davis.

How firmly may we believe of a proposition that passes the test that it is possible? It seems that if $P$ passes the conceivability test, then, at the very least, a rather heavy burden of proof is on the person who says that $P$ is not possible.

Actually, I do not accept the principle that anything qualitatively like pain at moment $m$ is qualitatively like pain. It seems to me that in order for a state to count as qualitatively like pain, it must endure through an extended, if quite short, period of time. Nothing in this chapter (or others) rests upon this view, but I think that it is plausible, and can be used to clear up various puzzles about qualitative states.

Or, in any case, to be a functionalist sympathetic with physicalism of the sort I described in Chapter I.

There are a number of philosophers who have pursued this strategy as well, e.g., Lycan, in "Form, Function and Feel." I believe, however that the view about conceivability arguments that I developed in the early sections of this chapter lays the groundwork for making such a strategy much more plausible.

Lycan makes this point in the above-mentioned article.

Putnam raises considerations such as these in a number of his early articles on mental states, e.g., "Machines or Artificially Created Life?"

Lycan, in "Form, Function, and Feel" makes the first point. Dennett has, in conversation, made the second. Third, the fact that the central nervous system of the entity composed of the population is made up of (large)
homunculi should not, in itself, encourage intuitions that the being is not sentient. After all, homunculus-invoking psychological theories are advocated both by dualistic and physicalistic philosophers to explain a wide variety of mental phenomena.
FOOTNOTES (Chapter V)

1 In, respectively, *The Language of Thought* and "Psychophysical and Theoretical Identification".

2 For ease of exposition, I shall adopt Lewis' manner of speaking of mental states as the state-types that have a certain role, rather than as the properties of being tokens of states that have that role. But his locutions can be easily translated into mine: for example, where Lewis would say "Pain is (the state) caused by bodily injury that leads to the desire to avoid such circumstances..." I would say, "Pain is the property of being a token of a state that is caused by bodily injury and leads to a state-type that has the role characteristic of desire..."

3 *Thought*, p. 18.


5 Nevertheless, in my "translation" of Lewis' generalization, I permit universal quantification over beliefs; thus there are a perhaps infinite number of beliefs in the range of the quantifiers of the functionalist theory. Some have taken this to be a problem, I think, for the following reason: To say that each belief is a distinct functional state, and that people have a potentially infinite number of beliefs, and that functional differences (in an individual) must be underlain by physical differences seems to imply that people must have an infinite number of different neurological, or at least physical, states in order to realize the functional definition. But this does not follow. A person will in fact form and retain only a finite number of beliefs, or at least stored beliefs, in his lifetime. To say that a person has a potential infinity of beliefs is merely to say that if a person did have unlimited storage space, an unlimited lifetime, and sufficient motivation, there would be an unlimited number of beliefs he could form and retain. A person need not meet these conditions, however, in order to realize a functional theory in which there are an infinite number of beliefs in the range of its quantifiers. The functional definition
couched in terms of universal generalizations about belief and desire is designed merely to insure that whatever beliefs a person does in fact have function in certain ways, and bear certain relations to each other and to other psychological states. It leave the question of how many beliefs a person has, or could have, open.

6 In Davidson's "On Saying That", and Quine's Word and Object, for example.

7 Fodor himself has said that internal tokens ought to be individuated functionally, in terms primarily of their combinatorial properties, and their role in producing motor responses. However, some (myself included) have understood him to be requiring stronger neural similarities among tokens of the same type. Thus, even though this "neural individuation" view was not actually intended by Fodor, I believe it is worth discussing, as it has come to be seen as a response to the problem of how to characterize internal sentential tokens of the same type.

8 There is an issue, among psychologists, about whether there is a common representational system, or, rather, distinct systems employed by the different sensory modalities. As I do not want to legislate that issue, I will assume that, in my example, there are no relevant processing differences that would support one way of construing the internal language over the other.

9 In his comments on this chapter.

10 As the old joke goes: "How does the wireless work?" "Well, it's like you have a big dog: you pull its tail in Minsk, it barks in Pinsk. The wireless? It's just like that - but no dog!" (Thanks to Michael Friedman for the joke, and for suggesting its relevance here.)
Bibliography


Dennett, D., "Intentional Systems," in *Brainstorms*.

Dennett, D., "Skinner Skinned," in *Brainstorms*.


Lycan, W.G., "Form, Function, and Feel" (forthcoming).


Warner, R., "Neural Codes and Mental States," unpublished.