INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame. If copyrighted materials were deleted you will find a target note listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in “sectioning” the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.
Wood, George Darlington

EXPLANATION AND MENTAL ENTITIES

The University of Arizona

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106

Ph.D. 1982
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark \(\checkmark\).

1. Glossy photographs or pages____
2. Colored illustrations, paper or print____
3. Photographs with dark background____
4. Illustrations are poor copy____
5. Pages with black marks, not original copy____
6. Print shows through as there is text on both sides of page____
7. Indistinct, broken or small print on several pages____
8. Print exceeds margin requirements____
9. Tightly bound copy with print lost in spine____
10. Computer printout pages with indistinct print____
11. Page(s)_________ lacking when material received, and not available from school or author.
12. Page(s)_________ seem to be missing in numbering only as text follows.
13. Two pages numbered _______. Text follows.
14. Curling and wrinkled pages____
15. Other__________________________________________

University Microfilms International
EXPLANATION AND
MENTAL ENTITIES

by
George Darlington Wood

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF PHILOSOPHY
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA

1982
As members of the Final Examination Committee, we certify that we have read the dissertation prepared by George Darlington Wood entitled Explanation and Mental Entities and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

April 15, 1982

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director

April 15, 1982
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

George Darlington Wood
# TABLE OF CONTENTS

I. Introduction 1

Part I: Scientific Explanation and the Mental

Chapter 2: Covering Law Explanation and Mental Entities 8
   Hempel and Salmon 10
   Hempel 13

Chapter 3: Theoretical Explanation 16
   Reductionism: a first glance 19
   Hempel's two levels of explanation 24
   Ramsey & Craig 27
   Quine against the Dogma of reduction 32
   Summing up 35

Chapter 4: Sellars on the "Levels" view 37
   Relating Observation to theory 40
   Correspondence rules 40
   Explanatory Models 47
   Summing up 54

Chapter 5: Some Features of Language about the Mental Items: A Digression 59

Chapter 6: Psychological Metatheory 69
   Psychology? 69
   Two Domains of Observation 71
   Behaviorism and Linguistics 73
   Physical Reduction 76
   and Parsimonious Behaviorism 76
   and Language 80
   Mentalism and Rich Reductive Behaviorism 87
   Levels in Psychology 89
   Appendix: Fodor's View 94

Part II: Physical Reduction and Identity 97

Chapter 8: Reduction and Identity 101
   Identity and discernibility 101
<table>
<thead>
<tr>
<th>Chapter 9: Possible Worlds and Conceptual Systems</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Arguments for Lewis' view</td>
<td>118</td>
</tr>
<tr>
<td>A. For Possible-Worlds realism</td>
<td>120</td>
</tr>
<tr>
<td>B. Against Trans-world Identity</td>
<td>122</td>
</tr>
<tr>
<td>II. Criticism of Lewis: Identity, Counterparts, and Composite Individuals</td>
<td>136</td>
</tr>
<tr>
<td>Gibbard Revisited</td>
<td>139</td>
</tr>
<tr>
<td>Sortal Identity and Nomologicals Theories and Their Entities/ Systemic Relations</td>
<td>141</td>
</tr>
<tr>
<td>Systemic Relations</td>
<td>144</td>
</tr>
</tbody>
</table>

Chapter 10: Nomologicals and Reduction | 149 |
| Quick Recap | 149 |
| Davidson's attack on Psychophysical Laws | 150 |
| Sellars: Verbal Behaviorism to Scientific Mentalism | 157 |

Chapter 11: Tokens and their Functions | 167 |
| Review | 169 |
| Token Identity: Objections and Rejoinders | 169 |

Chapter 12: Aspects and Stances | 179 |
| Dennett's Stances | 179 |
| Stances and reality | 183 |

Conclusions to Part II | 188 |

Part III: Philosophical Applications | 192 |

Chapter 13: Immediate Knowledge and Self-Presenting States | 193 |
| Chisholm's Theory | 195 |
| Sellars attack on the Given | 203 |
| Fodor and Innateness | 212 |
Chapter 14: Two Aspects of Personal Identity 224
   I. Introduction 224
   II. Lewis and Parfit on Survival and Identity 228
   III. Persons and Personalities 235
This dissertation centers on issues central to a scientific account of mental entities.

In part I, I first consider the general question of whether, and when, it is legitimate to postulate theoretical entities in science. I argue that interpreting theories as if their theoretical terms do not refer to real items lowers their explanatory value by eliminating the opportunity to provide connections with other theories and provide a cohesive account of the world. I then question whether psychological theory needs to postulate mental entities in order to provide adequate explanations of observable behavior, concluding that, behaviorist claims notwithstanding, talk about the mental cannot be reduced to talk about behavior. Finally, I argue that it is incumbent on such a theory to investigate the operation and constitution of its entities.

In part II, I address attacks on the view that mental items might be made of physical stuff, but not analyzable in purely physical terms. I argue that while revised dualist arguments can show that mental entities cannot be defined in the vocabulary of physics, this is nevertheless consistent with their having a physical constitution. I conclude that in addressing the issue
of mind-body identity, the solution lies as much in understanding "identity" as in understanding "mind."

In part III I apply the scientific account of mental items developed in the preceding parts to two traditional philosophical issues: the problem of explaining what it is to know, and the problem of personal identity. I sketch solutions to these problems and conclude that many remaining problems require scientific investigations rather than philosophical analysis.
CHAPTER I
INTRODUCTION

My purpose in writing this dissertation is to investigate some of the reasoning surrounding what has come to be a leading theory of mental entities, one which postulates the existence of items having 'mental' properties in order to explain human behavior and many 'obvious' ordinary truths about belief, thought intention, volition, desire and the like. This theory characterizes those the postulated items according to their functions, as will be brought out in following chapters. Thus, a mental item of type $m$ will be that item, whatever it is or whatever it is made of, that performs the function $m'$. I am not attempting to formulate a complete theory of the mental, and must admit that the features I consider are consistent with several versions of a more complete theory.

This work is divided into three parts, each centering around an aspect of the mental. My intent is to show how one type of theory of the mental can satisfy important requirements of science, ontology, and epistemology in a way which provides a coherent solution to puzzles raised by all three. I am motivated, at least
in part, by a view I argue for: that a philosophical theory of the mental can be satisfactorily justified only by showing it to be part of a coherent system of scientific and philosophical views.

Before describing what follows in greater detail, I must say what I am not doing. First, I am not attempting to analyze the ordinary meaning of mental terms or solve ordinary language antinomies regarding the reference of such terms. However, since many problems in the philosophy of mind are rooted in the uncritical acceptance of some 'obvious' ordinary language 'truths,' it will be incumbent on me to investigate and clarify some of them. I have included a short section on the types and goals of philosophical analysis, and a comparison of analysis and theory, in order to make my approach to these ordinary 'truths' clearer.

Second, I will have little to say about the nature and ultimate ontological status of sensations, sensa, and the like. In particular, I will not attempt to resolve the issues of the recent Sellars-Cornman debate, though I consider that a worthy undertaking in its own right. In the section on epistemology I will discuss the epistemic and functional status of sensations, and contrast them with thoughts, about I hope to do so in a manner which is neutral with respect to their ontology.
With those initial disclaimers out of the way I will describe each of the three major parts of the dissertation.

**Scientific Explanation and the Mental**

In the first part I investigate several accounts of scientific explanation with respect to their applicability to the explanation of the mental. In particular, I consider the views of Hempel, Salmon, and Sellars, and compare behaviorism in psychology to instrumentalism in science generally. I find several arguments to the conclusion that scientific realism is required for adequate theoretical explanation, and, following Sellars, Fodor and a host of more recent authors, I am led to the view that an adequate psychological theory requires the postulation of 'inner goings on' to explain a wide range of behavior, most notably intelligent linguistic behavior.

I aduce some support for this view by arguing that reductive behaviorism fails to show that mental entities, qua items postulated by psychological theory, are in principle, less real than the entities postulated by physical theories.

Second, I argue that the postulational account of the mental considered is consistent with other scientific theories and can help satisfy the requirements of
linguistics and the theory of meaning. It also poses a useful research strategy for neurophysiology. Finally, I begin considering the relation of this postulational theory to physics and the doctrine of the unification of science.

In the second part I consider issues of logic and ontology, focusing on the relationship of (postulated) mental items to the theoretical entities of physics. Versions of psychophysical reductionism are considered and found wanting. It is argued that the position that only physical stuff really exists is consistent with the postulational theory of the mental, even if it is not possible to reduce mental to physical items by providing translations of statements about mental phenomena to statements in physical theory without loss of meaning.

Part II also contains a brief consideration of the logic of propositional attitude statements and an investigation of questions of personal identity and survival.

Part III takes up some epistemological issues relevant to the philosophy of mind and argues that the postulational theory allows an escape from the traditional foundationalist/coherentist dilemma. In particular I focus on recent versions of R.M. Chisholms 'theory'
of knowledge, and its intuitive cartesian support, and of recent developments in the coherence theory. I argue that features which apparently support one or the other horn of the traditional dilemma can be shown to support an epistemology which rests on the postulational theory of mind, and within which one can slip between the horns.

Although this dissertation is largely inspired by the systematic philosophy of Wilfrid Sellars, it parts company from his views in each section. In particular I attempt to motivate the postulation theory of mind and an associated epistemology independent of questions about the ontological status of abstract objects, not so much because I object to his view as abstracta (I don't) but because I hope to abstract from such issues. A fully developed account would not be able to refrain from questions surrounding the problem of universals, and I shall have to make some assumptions with only the briefest sketches of arguments for the sake of brevity, I hope my readers will bear with me on those occasions.

Questions

It is uncontroversial that the philosophy of mind is a central area of philosophy. The core question in the philosophy of mind is 'what is the nature of the
mental?" Subsidiary questions clarify the relation of the philosophy of mind to ontology, epistemology, the philosophy of language and of psychology, and to ethics; 'Do mental entities exist?' 'Are they physical,' 'What is the nature of belief', 'What is it to understand a sentence or proposition), 'what are meanings,' 'How are mental states related to behavior,' 'What is a reason,' 'What is a person', 'What are the criteria of identity for persons?' Each of these questions leads to a host of others; for instance, in determining whether mental entities exist, one must first consider a broader question: 'What is it for something to exist?'

It does not seem possible to completely answer all of these questions in a single essay, and I shall not attempt to do so. Rather, I intend to examine the kinds of reasoning involved in philosophising about the mental. To this end I shall discuss 1) the role of scientific explanation in questions of ontology, 2) questions concerning identity, and 3) considerations stemming from psychology, the theory of language, and epistemology.
PLEASE NOTE:

This page not included with original material. Filmed as received.

University Microfilms International
CHAPTER II
COVERING LAW EXPLANATION
AND MENTAL ENTITIES

What are mental items? Do they exist? What kinds of properties do they have? What criteria of identity are relevant to them? Are they in some sense physical? What role can they play in explaining the phenomena of meaning? What is their role in knowledge? These are traditional problems in the philosophy of mind. My question is "To what degree can science, or scientific reasoning, or philosophical reasoning about science, clarify and answer these questions about the mental?"

The first issue I will investigate is that of the existence of mental items. Consider the following claims

1. The best explanations of human behavior use terms referring to such mental items as hopes, beliefs, desires, etc.

2. The best explanations are scientific explanations.

3. Scientific explanation of human behavior implies that mental items exist.

Do scientific explanations imply that mental items exist? This raises a more general question: What is the relation between scientific explanation and the existence of objects which are not publicly observable?
In this section I will show that covering-law explanations as described by Hempel and Salmon do not require the existence of theoretical entities and argue that 1) they are open to counterexamples and 2) are not explanations in any but a stipulated sense. In the following section, I will argue that only under a scientific realist interpretation of the referring expressions of explanatory theories can those theories yield genuinely satisfactory explanations. My argument will proceed by examining competing accounts of scientific explanation.

A corollary of the argument against instrumentalist attempts to avoid ontological commitment to theoretical entities is an argument against some behaviorist attempts to deny the reality of mental entities.

Scientific Explanation: Covering-Law Accounts

What is scientific explanation? Hempel, Salmon and Sellars have provided accounts which differ in content and emphasis. While all three seem to agree that realistic interpretations of theoretical terms are to be preferring to non-realistic interpretations. I will argue that neither Hempel nor Salmon provide strong arguments to the conclusion that a realist interpretation is required in order for a theory to provide a satisfactory explanation. Further, adding pragmatic
suggestions by Carnap and others to the accounts of Salmon and Hempel does not yield a much more satisfactory explanation, and while recent moves by Salmon require explanations to have causal components, it is not yet clear that this requirement is strong enough.

Our examination of covering law explanations will show that they are ontologically neutral in the sense that they require no ontological commitments and are hence consistent with instrumentalist interpretations of theoretical explanations which thus avoid ontological commitment to theoretical entities.

My strategy in this part will be to argue that because covering-law explanations are ontologically neutral in this sense, they are not as satisfactory as accounts similar to that of Wilfrid Sellars, according to which theories explain by providing models whose explanatory power depends on the referential force of theoretical terms.

Covering Law Explanation: Hempel and Salmon

In this section I briefly discuss some of the major features of contemporary accounts of scientific explanation, focusing on the neutrality of their ontological implications.
Hempel's DN and IS Accounts

Carl Hempel\(^1\) provides two accounts of scientific explanation: Deductive Nomological (DN) and Inductive Statistical (IS). Both analyze explanations as arguments, where the conclusion is a statement describing the event being explained (explanandum) and the premises include both a) statements describing circumstances antecedent to the explanandum event and b) Nomologicals, or law-like statements. Since in both cases the connection between antecedent circumstances and the explanandum event is a nomological, they are called 'covering law' explanations, in which the explanandum and antecedent events are subsumed under (i.e. are instances of) the covering law. [Technically speaking, the explanandum is the event (or perhaps object or state of affairs being explained, while the explanans are the events, states of affairs, laws or other factors in terms of which the explanation is made: and the statements describing them are explicandum-statements and explicans-statements, respectively. I will, however, use the unhyphenated expressions where context disambiguates them.]

The difference between DN and IS explanations is that in DN explanations the nomological statements are universal and the argument is deductive, whereas

---

IS explanations include some statistical nomologicals, and the argument is inductive, depending on a high degree of inductive support for the explanandum statements. One major problem for IS is that of specifying the degree of support required of a good explanation.

In both accounts, Hempel requires that all the statements making up an explanation be true, that the explanandum not be derivable from the descriptive or nomological premises alone and that the law or laws used play a non-trivial role in the derivation.

Hempel's IS account has the unfortunate consequence that low-probability events cannot be explained, since the inductive support for a statement describing them will be correspondingly low. Take for instance the case of radioactive decay of a single atom. The probability that any given atom will decay at some particular time is vanishingly small: hence no statistical argument will strongly conclude that the atom will decay at that time. This leaves the IS theorist in the position of either denying that the decay is explainable at all -- even if the mechanism of decay and reasons for its low probability were known -- or of admitting that they could only provide a very weak explanation.
Although Hempel's DN and IS models are widely accepted, it is also widely recognized that they are open to counterexamples, since not all arguments meeting Hempel's DN criteria are explanations: for instance from the premises that a flagpole's shadow is \( n \) feet long and that the sun is at a certain angle, and laws of light propagation and geometry, one can deduce that the flagpole is \( r \) feet high. But surely those premises don't explain the height of the flagpole. Thus, meeting DN criteria is not a sufficient condition for explanation. Nor can it be necessary that an explanation meet either DN or IS model: such a requirement would rule out explanation of low-probability events.

Salmon's SR Account

Wesley Salmon's Statistical Relevance (SR) model avoids these problems by characterizing explanations not as arguments but as sets of statements whose truth is statistically relevant to (descriptions of) the explanandum event. The goal of scientific research is to discover all conditions which are statistically relevant to the explanandum, arriving at last at a reference class specified in terms of those conditions which cannot be further partitioned (divided) in statistically relevant

---

ways: such a reference class is said to be objectively homogeneous, and even if the probability of the explanandum given the occurrence of conditions in the reference class is not high, the event is explained. Of course, if the probability is high enough, one could construct an IS (or DN) explanation from the relevant factors, but this is not required by Salmon.

It should be pointed out that merely adding the requirement of statistical relevance to the DN model would not eliminate the flagpole counterexample, since the angle of the sun is statistically relevant to the height of the flagpole. After noting Reichenbach's suggestion that causally antecedent events are 'screened off' from explaining their causal precedents, Salmon suggests that in such cases causal priority may play a role in establishing an explanatory assymetry, but fails to show that SR could add a screening off rule more readily than DN or IS. Thus, the principle virtue of SR over Hempel's accounts is that it allows the explanation of low-probability events. 4

One possible objection to the SR model is that the notion of objective homogeneity is vacuous, since

3. Ibid. p. 184.
one could never know, in principle, when (or whether) all statistically relevant conditions had been discovered. This complaint seems, however, to rest on an unacceptably simplistic variety of verificationism: one could similarly complain that the geometrical concept of a circle is vacuous, since one could never know whether an observable object was geometrically (perfectly) circular.

As a matter of common sense it is clear that neither DN nor IS nor SR provide explanations in the ordinary sense unless the statistically relevant statements making up an SR explanation include a description of the casual etiology of the explanandum event. In his more recent work, Salmon has added a requirement like this.

Having set the stage for it, we can now move on to a more thorough discussion of the role of theoretical terms in explanation.

---

5. Ibid.
CHAPTER III

THEORETICAL EXPLANATION

Theoretical Explanation: Hempel

Hempel seems to accept the account of theoretical explanation as explanation in terms of unobservable postulated entities: theoretical entities. He views such entities as having an important role in science:

Scientific systematization is ultimately aimed at establishing explanatory and predictive order among the bewilderingly complex "data" of our experience, the phenomena that can be "directly observed" by us. It is a remarkable fact that the greatest advances in scientific systematization have not been accomplished by means of laws referring directly to observables, i.e., to things and events that are ascertainable by direct observation, but rather by means of laws that speak of various hypothetical, or theoretical entities, i.e., presumptive objects, events and attributes which cannot be perceived or otherwise directly observed by us.¹

According to the characterization of theories provided by Hempel² a theory is a set of sentences expressed in terms of a specific vocabulary consisting of primitive terms and defined terms and whose sentences are exclusively primitive (axioms or postulates) or derivative (theorems).

1. "The Theoretician's Dilemma", p. 177
2. Ibid., Sec. 4

-16-
This characterization admits almost all deductive systems as theories. Hempel distinguishes empirical from non-empirical theories by requiring that the former have an interpretation given by reference to empirical phenomena.\(^3\)

As previously noted, Hempel distinguishes two levels of scientific systematization: empirical generalization and theory formulation. On his view, theories, involving the postulation of theoretical entities, are intended to explain lawlikeness of the empirical generalizations which are based on observed regularities. He also distinguishes observational from theoretical terms: the former are those whose applicability to a given situation is decidable on the basis of direct observations which are, at least in principle, intersubjective.

Reichenbach's Casual Relevance Argument

Theoretical entities are often postulated as casual links connecting distinct observable events. Reichenbach, evidently assuming that statistical relevance must be explained by casual relevance, postulates a common cause of simultaneous events exhibiting a high statistical correlation. Salmon reports that Reichenbach argues for this view by means of analogy:

---

Reichenbach invites us to consider an observer who is confined to the interior of a cube in which a bunch of shadows appear on the various walls. Careful observation reveals a close correspondence between the shadows on the ceiling and those on one of the walls: there is a high statistical correlation between the shadow events on the ceiling and those on the wall. For example, when the observer notices what appears to be the shadow of one bird pecking at another on the ceiling, he finds the same sort of shadow patterns on the wall. Reichenback argues that these correlations should be explained as shadows of the same birds cast on the ceiling and the wall; that is, birds should be postulated.... [Even though it is] physically impossible for the inhabitant to observe the birds directly, he should infer their existence. Reichenback is doing precisely what he advocated explicitly in his later work: he is explaining a relation of statistical relevance in terms of relations of causal relevance ....

Of course, theoretical posits don't always serve as common causes of observable regularities: when one explains that a light goes on when a button is pushed by saying that pushing the button allows electrons to flow through the lamp, causing it to emit light, the unobserved electrons mediate a casual process, but are not construed as a common cause of the two connected events.

While Hempel considers theoretical explanation to be a kind of deductive-nomological (DN) explanation, Salmon has argued that the essential features of theoretical explanation are preserved within his own statistical

---

4. Salmon, W., "Theoretical Explanation", p. 136
5. Ibid.
relevance (SR) model. Construed as a species of SR explanation, theoretical explanations are freed from the DN requirement that they must be arguments. Salmon's model of theoretical explanation is rooted in Reichenbach's principle that statistical relevance is to be explained by casual relevance and the view that where observable spatio-temporally contiguous casual processes are not available, it is in general preferable to postulate unobservable contiguous processes than to resort to a belief in remote causation. Thus, according to Salmon and Reichenbach, statistical relevance implies causal connection. Which, if unobserved, implies theoretical entities.

A major basis for objections to the practice of postulating unobservables is Uckham's Razor, which can be paraphrased as "posit not entities without necessity." Uckham probably did not use this formulation, but it is close to one he did: "We must not affirm that something is necessarily required for the explanation of an effect. If we are not led to this by a reason proceeding from a truth known by itself or from an experience that is certain." Objections generally continue by arguing

---

6. Ockham, William, "Reportation" question 150, in P. Boehner, ed. Philosophical Writings of William of Ockham
that, for various reasons, it is not necessary to postulate the entities: one can refrain from doing so, it is said, and still have good explanations for observed regularities. Let us consider some alleged means of eliminating posited entities without sacrificing explanatory power.

Reductionism: A First Glance

will construe reductionism as the broad thesis that statements containing a certain kind of term can be translated or analyzed without loss of meaning, into statements not using these terms. The idea is that the reduced statement will say everything you want to say without using ontologically suspicious locutions. Thus, if one were suspicious of the ontological status of smiles, one might propose a reduction of smiles to faces:

John has a nice smile
would be a paraphrased as

John smiles nicely
which does not make it seem as though smiles are things that could exist without smilers. Like Lewis Carroll's Cheshire cat.

Reductionism is an ontological thesis having several forms: taking the example of the reduction of
smiles to faces, a reductionist may admit that reduced items exist, but insist that they are really items of their preferred type. Thus, a reductionist might say,

Sure smiles exist, but they are really only states of faces.

On the other hand, a reductionist might deny that the reduced items exist (or 'really exist'), saying that terms allegedly referring to them are merely a convenient shorthand for more convoluted, but ontologically correct, expressions referring only to items in the reduction base; thus:

Smiles don't (really) exist. Smile talk is really talk about the ways faces can be.

Finally, a reductionist might be able to have his cake and eat it too, by emphasizing the importance of the 'really' in the second approach. It serves to point out the difference between an ordinary sense of 'exists' and a metaphysical sense according to which only ontologically basic items, those forming the basis of one's ontology, really exist. This sort of reductionist might say

Of course smiles exist: but they don't really exist, since they are really only configurations of things which really exist, faces. The fact that people sometimes speak of smiles as if they were things in their own right is evidence of conceptual confusion imbedded in language, and should be dealt with by thorough philosophical analysis.
This broad construal of reductionism allows of many special cases, depending on the mode of reduction, the types of items being 'reduced,' the reduction base (i.e. the types of items they are 'reduced' to), and an account of what is to count as an acceptable paraphrase or translation. It must be noted that some reductionists are willing to give up intuitions about the meanings of words, proposing their analyses as correct accounts of these expressions, prescriptions as to what these terms should be taken as meaning.

Historically there have been forms of reductionism which have reduced things of various sorts to perceptions, ideas, observables, or operations performed at the level of observation. I will be concerned for the most part with forms of reductionism stemming from logical empiricism. In an early paper, Hempel argues for what I take to be a paradigm of this form or reductionism with respect to mental entities:

All psychological statements which are meaningful, that is to say, which are in principle verifiable, are translatable into statements which do not involve psychological concepts, but only the concepts of physics.7

Hempel points out that his physicalistic conception of psychology, which he calls "logical Behaviorism," is derived from the principle of verification:

The meaning of a statement is established by the conditions of its verification. . . . Furthermore, a statement for which one can indicate absolutely no conditions which would verify it. . . is wholly devoid of content and without meaning. 8

He also says that Logical Behaviorism, as he describes it, does not claim that mental states do not exist: talk about their existence is a pseudoproblem, since terms apparently referring to them are merely abbreviations for physicalistic expressions. And while Logical Behaviorism may focus on the responses of organisms to stimuli, it is not in principle restricted to this since the conditions of verification may also include "the most minute details of the phenomena of the central nervous system." 9

This reduction is a special case of those forms of reductionism according to which all theoretical terms can be eliminated in favor of observational expressions. These broader forms include instrumentalism, and Operationalism. The former holds that theoretical terms are merely a convenient shorthand (verbal instruments) which serve to abbreviate complex conditional

compounds of observable conditions which, unabbreviated, would render explanations containing them opaque. Operationalism is similar but includes not just observable conditions but the result of prescribed operations. The distinction between them will not play a crucial role in what follows. It is important to note that for both, the complex expressions being abbreviated are conditional in nature. Thus, solubility is not a thing literally had by a chemical compound: rather it is explained in terms of an analysis of such open sentences as "x is soluble in y" which is in turn analyzed into a complex expression starting with "if x were placed in y, x would dissolve -- if conditions are . . . ." I will not attempt to fill these ellipses here.

Although Hempel embraced reductionism in his early papers, his view had somewhat weakened by the time he argued in "The Theoretician's Dilemma" ¹⁰ that since no conclusive argument has been made for -- or against -- the view that reduction is, in principle, always possible, and that since it is at present impossible to formulate observational definitions for all the theoretical terms in current use, theoretical

¹⁰. Hempel, Carl, 1958/1965, Sect. 4
terms are, at least for the present, necessary parts of explanations, employing them. We will now examine the wider role that Hempel attributes to theoretical terms in this later paper.

Hempel's Two Levels of Explanation

Part of Hempel's reason for dropping his earlier view that theoretical terms are abbreviations for complex observation expressions is that they serve a higher-order function than the expressions of observation language: they explain observed regularities in a way they could not if they were equivalent to statements describing observations, observed regularities, or generalization based on them. Recognizing that theoretical terms

... usually purport to refer to not directly observable entities and their characteristics: they function. ... in scientific theories intended to explain empirical generalizations. 11

Hempel endorses

... the familiar rough distinction between two levels of scientific systematization: the level of empirical generalization and the level of theory formation ... where research is aimed at comprehensive laws, in terms of hypothetical entities, which will account for the uniformities established at the first level. 12

11. Ibid.
12. Ibid.
Hempel seems, at this stage, to think that theoretical entities are, in principle, unobservable. This would imply that there could be no observational discovery of a previously postulated theoretical entity. Later, when discussing specific gravity, he allows for theoretical properties which are 'less directly observable' than, for instance, the fact that a piece of wood floats.

Oddly, Hempel does argue for a realist interpretation of theoretical terms. His argument begins with his characterization of the theoretical/observational distinction, which leads him to ask "whether the systematization achieved by general principles containing theoretical terms can always be duplicated by means of general statements couched exclusively in observational terms." This question, together with the assumption that the sole purpose of theoretical terms is to state general principles ultimately relating observables to observables, leads to what he calls the theoretician's dilemma:

If the terms and principles of a theory serve their purpose they are unnecessary . . . and if they don't serve their purpose they are surely unnecessary. But given any theory, its terms and principles either serve their purpose or they do not. Hence, the terms and principles of any theory are unnecessary.13

13. Ibid., p. 186
Hempel's general solution to this dilemma is to deny that the assumed purpose that of being part of a formal, though perhaps uninterpreted theory from which law statements linking observables may be deduced, is the only purpose of theoretical terms. Running the gauntlet of verificationism, reductionism, and both Ramsey's and Craig's methods of eliminating reference to theoretical entities, he argues that simplistic reductionism fails to account for the varying roles of theoretical terms and that while the more complex forms do account for the role of mediating observations, they can do so only at the cost of tremendous complexity and a loss of explanatory and neuristic utility. (E.g. the Craigian expansions of some theories are so complex that it would be impossible for humans to comprehend them.)

After a brief look at Ramsey and Craig, we will review some criticisms of their approaches before turning in the next chapter to arguments in favor of a realist interpretation of theoretical entities and/or their roles in explanation. In later chapters we will examine the explanatory role of mental entities qua entities postulated by psychological theories.
Ramsey and Ramsification

Frank Ramsey has suggested a way of avoiding ontological commitment to theoretical entities which, while not entirely satisfactory, provides what might be called a functional analysis of theoretical terms. Ramsey's suggestion came in two forms, one involving existential quantifiers, the other involving infinite disjunctions. On the first form, the theory is taken as the conjunction of its sentences and variables are uniformly substituted for its theoretical terms: then the conjunction is existentially quantified. Using a different quantifier for each replaced term-type. This formulation will yield all the observational consequences of the original theory, yet use no term not included in the observational vocabulary: the Ramsified version is 'functionally equivalent' to the original theory.

The second form avoids the use of existential quantifiers, at the cost of making the functional variant of the theory into an infinite disjunction of conjuncts. (On the substitutional interpretation an existential quantification is equivalent to a string of disjuncts that exhaust the domain of discourse: i.e.,

14. Hempel discusses this approach in sections 8–9, "The Theoretician's Dilemma".)
exhaust the domain.) But taking this route robs the theory of its simplicity -- makes it unstatable --, to say nothing of what it does to its systematic explanatory power.

On the other hand, using the existentially quantified version may avoid theoretical terms, but it does not avoid the postulation of entities, at least on a standard (objectual) interpretation of quantifiers for the quantifier says, "there exists an x such that...."

Finally, the quantified theory, while more perspicuous than the disjunctive one, is nonetheless far less simple and perspicuous than one naming theoretical entities. Pragmatic considerations, including simplicity, perspicuity and heuristic value would seem to weigh in favor of a theory naming theoretical entities even if it has no more observational consequences than its quantified counterpart.

Ramsey's analysis is, however, useful in understanding one of the things theoretical explanations do. They provide functions from one kind of observable state to another. The ordinary mathematical concept of another, with the restriction that no argument of the function can have more than one value. For inductive theoretical explanations of indeterministic but statistically relevant relations between observables, the
values of functions must be allowed to be disjunctive, with probabilities assignable to the disjuncts if functional explanations (in our present sense) are to work. Thus, for instance, we could describe a function for the roll of a die: \[ f(x) = \{1, 2, 3, 4, 5, 6 \} \] (where "1" reads "the one-face of the die is up") with equal (1/6) probabilities for each face on a fair die. Construing these disjunction-probability assignment pairings as unique values of functions makes the mathematical concept of a function directly applicable to theoretical explanation without requiring that theoretical explanations be deterministic.

As an example of indeterministic theoretical explanation consider mendelian gene-theoretic explanation. In Mendel's theory, genes are postulated entities performing certain trait-transfer roles. If parents \( \mathcal{P} \) and \( M \) have certain traits, then their child will have various probabilities of having each of those traits or their mixtures: genes causally mediate statistical regularities. Even low-probability traits which occur are explainable. Thus, this account of theoretical explanation is consistent with Salmon's SR model of scientific explanation.
A Brief Digression: Intension and Extension

It is useful at this point to make a neo-\textsuperscript{1}regian distinction. It is relatively non-controversial that a term's reference does not exhaust its meaning, for other aspects of the term's function in a language makes an important contribution to meaning. This is most clear when the term is syncategormatic or otherwise non-referential: e.g. 'and' has no reference, but gets its meaning from the role it plays relative to other expressions. A term's extension, the object (or objects) to which it refers can be distinguished from its intension, which is a function defined over use-conditions, reference, and other functions. (Lewis, Kaplan and Montague have, following Carnap, defined intensions as functions whose domains are states of affairs -- possible worlds -- and whose ranges are usually extensions, but may be intensions.)

Graig's Account

William Craig went beyond Ramsey by proving that any theory could be replaced by an observationally equivalent (though perhaps infinite) set of sentences having neither theoretical terms nor variables of quantification whose values are unobservable: the replacement proceeds by using bridge laws to replace every occurrence of a theoretical term directly with its
observational consequences. Since the observational consequences are recursively enumerable, the observationally equivalent set of sentences is at most denumerably infinite. It will also be functionally equivalent to the theory in the same sense that the Ramsified version is, so the Craig set will be logically equivalent to and in that sense have the same intension as the theory taken as a whole. And even though it contains no theoretical terms, the bridge laws which 'translate' them provide them with intensions in this sense.

Using the distinction between intension and extension, we can say that **reductionists** believe that theoretical terms have intensions but no extensions in the real order (i.e., the actual world). Their intensions are functions whose values are extensions, that is, **observable** extensions.

By contrast, the scientific realist claims that theoretical terms belonging to good theories have both intensions and extensions. If things had been other than they are, if, for instance, a term which is actually a theoretical term referred to an observable, then, the instrumentalist would agree that the term had an extension, but would not call it a theoretical term. As it is, however, the instrumentalist says, the term serves merely as a function relating observables. Thus its intension is a complex function from the actual
world to functions relating observables and from other possible worlds to either observables or such functions, depending on whether or not it has an observable reference in those worlds. (It is interesting to note that the instrumentalist is precluded from discovering theoretical entities which had not been previously observed: for in his parsimonious account, there is no term even allegedly referring to an entity. We will give greater consideration to this feature in the next chapter.)

But the instrumentalist is not concerned so much with language as with ontology. Thus, he may focus on the theory as a whole rather than just on its theoretical terms: the point he wishes to insist upon is that theoretical explanation does not entail ontological commitment to entities, either in virtue of the use of apparently referential terms or for other reasons. Thus he might argue that the theory, taken as a whole, does not entail the existence of supposedly postulated theoretical entities, whether or not it uses theoretical terms. Thus, the theory would be taken to be equivalent to a version of it which makes no reference to theoretical entities.

Quine and the Dogma of Reduction-in-Principle

In his important essay "Two Dogmas of Empiricism," Quine argues that the reductionist claim that
theoretical terms can be reduced to the terms of an
observation language is always possible in principle
is a dogma which cannot be given empirical support.
His argument is that there can be no empirical test of
a proposed reduction, since the inference rule or
principle invoked in the reduction cannot be interpreted
by itself. This follows from the Duhemian thesis of
wholism, according to which individual terms have no
meaning outside a language or theory. Thus a reduction
rule (or bridge law) replacing an individual theoretical
term can only be tested in conjunction with the theory
and the observational language which the rule supposedly
links, as a whole: for if a reduction rule yields some
prediction which is contrary to the observational reports
of competent individuals, there are several possible
explanations: the reduction rule could be incorrect,
the theory could be in error, or the observation
language itself could be in error. Examples of the
latter can be drawn from history: The ancients observed
the sun 'going down' and took it as obvious that the
earth did not move: they observed the acts of gods:
Lightning bolts were the spears of Zeus, etc. Our
'observations' that tables are solid are questionable,
at least to the degree that we have modified our under-
standing of 'solidity' so as to understand that macro
physical solids may be micro-physical lattices, full of empty space.

Thus, since the reduction rule (or bridge law) cannot be tested independently of a whole system without assuming both the adequacy of the language involved in reporting experience and the correctness of the theory, there is not, and cannot be, empirical support for the view that any particular reduction is correct, and hence for the view that reduction is always, in principle, possible.

Quine's claim can be restated in terms of the intension/extension distinction: The Duhem thesis would entail that individual terms do not have intensions outside of whole languages. (This may seem to do violence to Quine, for he would insist that intensions are in no way respectable entities because they are indeterminate and he holds that there can be "no entity without [determinate] identity.") Thus, no amount of evidence could show that a translation of it into another language was correct. (This is a version of Quine's indeterminacy of translation thesis.) If two theories are extensionally equivalent with respect to all observable evidence, Quine would say that there is no legitimate empirical question as to which would be right or wrong, and suggest that we choose between them on pragmatic grounds.
It should be noted that there is no reason to believe that a Craigian version of a theory that postulates as yet unobserved entities will be (in the future) equivalent to that theory, since the bridge laws are based only on past observation. But, as we shall see, there may be reason to postulate theoretical consequences in the sense that the entities of one theory play roles in the theories of other disciplines even though these consequences are not be observationally verified or verifiable.

Summing Up

Hempel's account of explanation seems neutral in the sense that the role of theoretical terms in DN and IS explanations can be played by equivalent Ramsifications or Craig sets. This is because theories explain, according to Hempel, in virtue of their ability to support the deductive or inductive inference of the explanandum, and functionally equivalent alternatives will support the same inferences. If Hempel had shown that science would be in some sense impossible without a realist interpretation of theoretical terms then his account would be committed to scientific realism: but the closest he has come to an argument of this sort is to suggest that the use of theoretical terms has significant heuristic utility in virtue of
simplicity of expression and to suggest that they may play other important roles.

We have been cautiously approaching a discussion of competing interpretations of theories and theoretical terms. Since Hempel's DN and IS accounts are neutral with respect to the status of theoretical terms, and since he has not shown their heuristic utility to be necessary to science, his account of explanation does not require reference to theoretical entities. The time has come to deal with other issues surrounding theories and especially to consider other roles of the realistic interpretation of theories in explanation.
Another account of theoretical explanation is argued for by Wilfrid Sellars. He contends that views which equate theoretical explanation with deduction, from a theory, of empirical laws, which in turn explain observable events, is a mistake: theories explain observable events directly and explain laws only by explaining why observable exhibit the regularities they capture.  

The levels view gets much of its plausibility, as Sellars suggests, from the intuition that we seem to explain particular observable events in terms of laws and laws in terms of theories; but this intuition is based on confusing explanation with derivation, and on taking the level of observation as a basis which is immune to revision on the basis of any theory. This, in turn, would be tantamount to a rejection of even the law of noncontradiction in the face of conflicting observations. Sellars puts it thus:

---

1. Sellars, W., "The Language of Theories," Sec. III, in SPR.
to conceive of the explananda of theories as, simply, empirical laws and to equate theoretical explanation with the derivation of empirical laws from theoretical postulates. . . . is to sever the vital tie between theoretical principles and particular matters of fact in the framework of observation. Indeed the idea that the aim of theories is to explain not particular matters of fact but rather inductive generalizations is nothing more nor less than the idea that theories are in principle dispensable. For to suppose that particular matters of fact are the proper explananda of inductive generalizations in the observation framework and of these only, is to suppose that . . . the conceptual framework of the observation level is autonomous and immune from theoretical criticism.2

The levels view, then, seems to come to the idea that particulars are explained by abstract generalizations, still within the conceptual framework of observation, and that these abstract generalizations are explained by theories, conceived of as higher level abstractions, at a conceptual rather than individual level, from the observation framework. It should be no surprise that theories, so construed, are dispensable, and that entities postulated within theories, so construed, should be viewed as convenient fictions and not as particulars. IF this account of theories were correct, theories could not, in principle, add real content to explanations. On the other hand, if theories can explain particular observable events without the mediation of

2. Ibid., pp. 120-121
laws stated in terms of observables, or if theories can be construed in a manner in which they add content and do not serve merely as higher level abstractions, then the levels theory fails to provide a satisfactory account of genuine theories.

How then do theories explain events, according to Sellars? They do so by providing more adequate descriptions of the antecedent circumstances and events to be explained. Thus a theory would explain the increase temperature of a volume of air which was compressed to half its original volume by redescribing the air as a system of randomly moving molecules and its temperature as mean kinetic energy, and showing that the kinetic energy of a volume of molecules increases when those molecules, each having a given antecedent kinetic energy, are placed in a smaller volume. For the redescription to explain, of course, sense must be made of the claims that the air is a system of molecules and that temperature is mean kinetic energy. Surprisingly, Sellars does not take the simple Leibniz-Russell route of identifying air with the molecular system it 'consists' of, at least not in any simple sense of identity. The reason for this is

that on his view there is a substantial difference between identity within a conceptual system and 'identity' across conceptual systems. But to say more about this here would be to get ahead of ourselves.

It should be pointed out that this account of explanation is not, strictly speaking, inconsistent with the Hempel and Salmon accounts except to the degree that they are committed to the view that theories explain individual events by explaining laws: for, as shown in the last chapter, they claim to allow either realist or instrumentalist interpretations of the terms involved in explanations. Insofar as DN, IS, and SR explanations are characterized more by their syntax than by semantic constraints, and insofar as they are semantically characterized, they are neutral with regard to the question of whether or not any of the terms used in them refer to real items. This, from my viewpoint, is their weakness; they are too liberal. I need not point out that insofar as Hempel's account is committed to the level's view just discussed, it is inconsistent with his tolerance of realism.

Relating Observation to Theory: Correspondences

Sellars' notes that there must be some way of relating an observation-language description of an event
to its theoretical counterpart. This cannot be appropriately achieved by redefining observational terms in terms of theoretical terms. Most simply put, this is because 'air' does not mean 'system of molecules,' because those expressions do not play the same role. Sellars calls the rules relating observational language to the language of theory 'correspondence rules,' and, after rejecting the view that correspondence rules are merely rules for deriving laws from statements of theories, Sellars argues as follows

But what is a correspondence rule if it is not simply a device for deriving laws from theoretical postulates? We have seen that a correspondence rule is not a partial definition of theoretical terms by observation terms. Nor, obviously, is it a definition of observation terms as currently used by means of theoretical terms. But might it not be construed as a redefinition of observation terms? Such a redefinition would, of course, be a dead letter unless it were actually carried out in linguistic practice. And it is clear that to be fully carried out in any interesting sense, it would not be enough that sign designs which play the role of observation terms be borrowed for use in the theoretical language as the defined equivalents of theoretical expressions. For this would simply amount to making these sign designs ambiguous. In their new use they would no longer be observation terms. The force of the 'redefinition' must be such as to demand not only that the observation-sign design correlated with a given theoretical expression be syntactically interchangeable with the latter, but that the latter be given the perceptual or observational role of the former so that the two expressions become synonymous by mutual readjustment. And to this there is an obvious objection; the meaningful use of theories simply does not require this usurpation of the
observational role by theoretical expressions. Correspondence rules thus understood would remain dead letters. 4

As we shall see, the presupposition that identifying the objects represented within an observation language with those of a theory requires a definitional link -- one which preserves meaning -- is mistaken. Sellars suggests that correspondence rules allow the language of identity to be used to link theory to observation as a sort of promissory note for a more adequate language containing definitions of observation terms in terms of theoretical terms (and not vice-versa, as traditional reductionists would hold). As Sellars puts it,

Does it make sense to speak of turning empirical predicates -- and in particular observation predicates -- into definitional abbreviations of complex theoretical locutions? Could observation predicates be so treated while continuing to play their perceptual role as conditioned responses to the environment? I see no reason in principle why this should not be the case.

From the standpoint of the methodology of developing science, it might seem foolish to build physical theory into the language of observation and experiment. A tentative correlation of theoretical and empirical terms would seem more appropriate than redefinition. But this is a truism which simply explains what we mean by developing science. But the perspective of the philosopher cannot be limited to that which is methodologically wise for developing science. He must also attempt to envisage the world as pictures from that point of view -- one hesitates

to call it Completed Science - which is the regulative ideal of the scientific enterprise. As I see it, then, substantive correspondence rules are anticipations of definitions which it would be inappropriate to implement in developing science, but the implementation of which in an ideal state of scientific knowledge would be the achieving of a unified vision of the world in which the methodologically important dualism of observation and theoretical frameworks would be transcended, and the world of theory and the world of observation would be one.5

Sellars' point seems to be this: ordinary language objections to identifying theoretical and observational items and properties can only be 'met' by changing from contemporary language (to the more adequate language of unified science) and that current acceptance of such 'identification.' In the form of correspondence rules, expresses the belief that such a system is an acceptable regulative ideal.

The key to understanding Sellars' account of correspondence rules is to recognize that they express correspondences between items which belong to different conceptual frameworks, not correspondences between items in a single framework. They are then 'translations' which must fail to preserve exact equivalence of meaning unless the frameworks are functionally identical; they are 'translations' that relate the concepts of one system to their closest analogues in another system.

Thus, for instance, "mean (molecular) kinetic energy" in a contemporary conceptual framework corresponds to "phlogiston" in earlier theories of heat. Neither means the same as the other, for the latter has been succeeded by the former. But they are analogous in that each played the role, within its framework, of explaining phenomena associated with heat, fire, compression, friction, and the like. But it must also be noted that "heat" also has a different conceptual role in the two systems, and thus, even though they use the same sign design, the word "heat" in the system incorporating phlogiston has a different meaning from the word "heat" in the more recent system. They do, of course, correspond to each other (perhaps more closely than do "phlogiston" and "mean kinetic energy," since they have nearly identical observation roles); but the former, unlike the latter, entails statements about phlogiston, including some whose contemporary counterparts are false, e.g., "Heat is transported by an etherial fluid."

Thus, correspondence rules express relations between items in different conceptual frameworks, which are conceived of as whole linguistic and scientific systems, since different languages can contain the same theories). Sellars seems to agree with Duhem and Quine in thinking that such systems "race the tribunal of
experience as a whole" and that the elements of one are incommensurable with those of another. But he goes on to recognize that from within the standpoint of a new conceptual system, we can recognize and discuss the role of elements (terms) of earlier systems.⁶

One further note must be made here. According to Sellars it turns out that observational statements like "labels exist" are literally false -- if made within an Aristotelian framework within which observed substances must be essentially as they appear (the framework Sellars calls 'the manifest image'). The terms of a strict observation language will, in the end, fail to denote real things since the 'observational meaning' of terms like "table" neither denote the objects of ideal science nor are defined in terms that do:

According to the view I am proposing, correspondence rules would appear in the material mode as statements to the effect that the objects of the observational framework do not really exist -- there really are no such things.⁷

Thus construed, correspondence rules will identify observables as theoretical entities (or structures of them), but not identify them with each other, thus:
"The items we believe to be labels are really structures of quarks."

⁶. Sellars, W., 1967, Science and Metaphysics Chapter 5, Sections VII-XII
⁷. Ibid.
Does Sellars believe that tables don't exist? Of course not; for to correctly interpret the above passage one must construe the 'observation language' as the language of what Sellars has called the manifest image. This idealization specifically avoids questions of the ultimate makeup of observed items, and, to put it very roughly, construes them as really being the way they appear (the way they manifest themselves to us). Since tables appear to be solid. In the manifest image they are really (essentially) solid, just as they appear to be. What the manifest image specifically excludes is the sort of scientific view of tables as appearing to be solid but really consisting of unseen postulated theoretical entities. The manifest table thus cannot be the table of contemporary common sense, which is, as 'everyone' knows made of atoms and consists of a lattice spread through largely empty space; it is thus more 'spongy' than 'solid.' Sellars believes that something like this common sense table exists. That something is the ideal scientific counterpart of the 'vague' common sense table. This counterpart is the item that plays a role in the scientific framework which is similar to the role played by tables in the manifest framework. Note that the essential solidity -- among other manifest properties -- of the table precludes
it from having an ideal scientific counterpart, at least if ideal science develops further along the lines that have produced current science. 8

I once thought that Sellars' was both too willing to allow the inadequacies of present-day usage, and too radical in thinking that we are mistaken in thinking that observable macro-objects, e.g. tables, exist. Surely tables exist; and if they are really structures of theoretical entities our mistake is not acknowledging that fact in our usage. It now seems that what separated my former view from that of Sellars' was more a matter of attitude than a matter of substance, so long as Sellars allows that observable tables do exist and are structures of theoretical entities. The question of the identification of observable objects with theoretical ones will resurface later.

Explanatory Models

According to Sellars, a theory explains an event by describing it, and its causes, as they 'really' are (or as they are according to the theory, assuming it to be the best scientific account available). Science is the measure of things, and an ideal science would provide ideally adequate representations, or pictures, 

8. Sellars, W., 1963, Science, Perception and Reality, pp. 4-18
of the world. One way to put this, using the language of contemporary logic, is that a theory, viewed as an uninterpreted calculus, does not explain, but that the theory, together with a realistic interpretation, provides models with genuine explanatory power.

By a 'realistic interpretation' I mean one in which the domain is taken to be existing entities, those postulated by the theory, having the properties the theory ascribes to them. But Sellars seems to reject 'models' as being genuine explanations:

Now the deductive system of a theory is often formulated with reference to a model. A domain of objects is pointed to, which either behaves in ways which satisfy the postulates of the deductive system, or can be imagined to do so without an absurdity which would deprive the reference to them of any value or point. The model serves a number of purposes. The most obvious is to make the theory intuitive, and aid the imagination in working with it. But more than this it fills an important need in that whereas the basic magnitudes of the empirical framework are operationally defined and are therefore rooted in a background of qualitative content, the basic magnitudes of the theoretical framework, in the absence of a model, would in no way point to a foundation in nonmetrical, qualitative distinctions which might stand to them as the qualitative dimensions of observable things stand to the metrical properties which the operationally defined with respect to them. The theory would leave them 'abstract' in a sense which reminds us of Whitehead's charge of 'vacuous actuality' against scientific materialism. The basic magnitudes of the theory would simply point forward to the more complex theoretical properties which can be defined in terms of them, and would find their be all and end all in the theorems which save the appearances (empirical
generalizations). Now be virtue of their visualizable character, models provide a surrogate for the "qualitative" predicates which must, in the last analysis, be the underpinning of theoretical that could "really exist," if this phrase can be given a stronger interpretation than that of the irenic instrumentalist. Needless to say, the qualitative dimensions which provide the content for the metrical form or theoretical entities need not be the perceptual qualities of the model, it would be odd if the only qualitative dimensions of the world were those which are, in the last analysis, tied to the sensory centers of the human brain.

Thus, however important the heuristic function of the model, and however important the layer of analogical meaning which the theoretical predicates acquire by being explained in terms of the model, the theory is not about the objects of the model, nor do theoretical predicates stand for properties of the objects in the model. The reference of the theory, if it can be said to have reference, and the meanings of the predicates of the theory, insofar as these are more than an adumbration of things to come, are to be understood in terms of the deductive system and the coordination of the theory with the empirical generalizations it is designed to explain. 10

We will return shortly to the issue of the interpretation of the predicates in a theory. Sellars wanted to insure that his view was not equated with that of Mary Hesse, who argued that models provide the interpretation of theoretical predicates, i.e., that the properties used by the model, as clarified by its glosses and disclaimers, are identical to the properties

picked out by predicate constants in the modeled theory.11

Several senses of 'model' need to be distinguished; models can be simulations, replications or descriptions. A simulation model is one which produces similar observational effects which are similar to those of the item being modeled, even though the means by which these simulated effects are achieved are radically different than they are in the original case. A chess-playing computer, of current vintage, simulates the chess-playing of a person even though the causal mechanisms and reasoning are radically different than those of human players. Note that the 'means' which bring about the effect need be neither the same kind of stuff, nor need they operate in the same way as the original for a simulation model to be an appropriate simulation. Hence a simulation model of human behavior need not provide an explanation of human psychology.

A replication model provides a replica of the item or process being modeled. The important difference

11. In "Scientific Realism or Irenic Instrumentalism (SRII, in pp. 196) Sellars charts a course between Nagel's view, according to which all theoretical predicates are mere predicate variables and Hesse's, according to which they are constants whose content is specified by heuristic models.
between a replication and a simulation is that the replica must operate in the same manner as the item being replicated. Now, this does not require that the model be constructed of the same stuff as the original, but that the elements of the replica reproduce the causal patterns being modeled. Thus, it would be possible to create a replica of a person from transistorized circuits rather than neurons, so long as those circuits function in the manner that the original's counterparts do. The distinction between replicas and simulations is actually one of degree; in a replica, the functional components, and not only the overall functions, of the original are simulated. The extreme case of replication is duplication; this is where the components of the components are functionally simulated *ad infinitum* (actually, not to infinity, but to the limit of science's ability to reveal levels of composition and determine their functional roles). An exact duplicate would have exactly similar construction and operation. A science of persons capable of constructing an exact duplicate from scratch would presumably be able to explain all there was provided, of course that it was capable of describing the functional relations of the constituents as the model was assembled.
The apparent conflict between Sellars' view and the description in the terminology of logic I have sketched for it is due to a difference in the use of 'model.' In the above passage, Sellars views a model as an interpretation which gives theoretical entities properties which are either qualitative, or analogous to qualitative properties with which we are familiar; its purpose is to enable us to visualize the theory's descriptive content, "to make the theory intuitive, aid the imagination in working with it." But this visualizable, or Heuristic model is not the same as the realistic model I described. One can consistently speak of two kinds of models for a theory, one designed to be formally correct, even if it involves properties which persons cannot sense (like valence), and one which serves the heuristic purpose of making the theory intuitive (like saying that electrons are like little balls). According to Sellars, then, only the former, mathematically pristine, models would provide genuine scientific explanations; but the latter, based on analogies to everyday objects, are what serve the ordinary explanatory function of 'making the phenomena understandable' or 'reducing the unfamiliar to the
familiar.' Sellars felt that it was a mistake to identify the content of theoretical predicates with their model's analogues, as did Mary Hesse. As long as the distinction between these two ways of modeling is retained, and the important fact that theoretical entities do not literally have the intuitive qualities attributed to them by the heuristic model, but only scientific properties somehow 'analogous' to them, then we can account for the requirements of ordinary understanding within a single account of explanation.13

This suggestion makes it clear that the disanalogies between postulated properties and their ordinary counterparts must be kept clearly in mind in order to distinguish the real commitments of theory from what it might seem

13. See SRII for Sellars' account of the content of theoretical terms and his discussion of Hesse and Nagel. He distinguishes three possible ways of interpreting theoretical predicates: the interpretation given by a heuristic model, a refined version of this with disanalogies glossed, and one according to which their interpretations are analogous to, but not identical with those provided by the glossed model. Sellars thinks that Hesse's mistake was in not recognizing the third and hence identifying theoretical properties with those of the heuristic model. I incorporate this distinction into my distinction between heuristic and realistic models. I believe that Sellars would find my realist models acceptable in the long run, but unavailable now.
to be committed to if we were to interpret it according to everyday usage. This feature is brought out in contemporary physics, by the attribution to quarks of such features as charm, strangeness, and beauty. Few people would be tempted to interpret these in an ordinary sense, or to criticize quark-theory on the grounds that "subatomic particles couldn't be beautiful." (It is interesting to note that these terms, unlike 'spin' and 'charge' from earlier atomic theory, do not seem to invite being taken literally.)

The suggestion that explanations are models of the circumstances culminating in an event is not, of course, completely novel. It can be found in the semantics of Montague's paper "Physical Theories," and it seems to me to be central to Sellars' discussions of explanation; for he holds that theories directly explain observable phenomena by picturing it in terms of theoretical entities and their behavior. This view requires a realist interpretation of theoretical terms, a point to which we will return below.

Summing Up So Far

Thus, Sellars' account of theories and theoretical explanation rejects the usual distinction between observational and theoretical terms and the 'levels' picture of explanation according to which
theories explain laws and laws explain observations. A good theory explains an observed phenomena by describing it as it really is. When a person observes an object, say a table, he is observing a structure of theoretical entities as a table.

Sellars seems to say that it is literally false that tables exist; this flies in the face of common sense. We saw that this reading of Sellars is only plausible if 'literally false' is taken as meaning 'literally false within the manifest framework'; and this is trivially true given Sellars' account of the manifest framework.

Must one reject the common-sense belief that tables exist in accepting a largely Sellarsian account of explanation? Sellars' main reason for denying that tables, as conceived of within the manifest image, exist is that, as a scientific realist, he is committed to all and only those entities postulated by good scientific theory, and that manifest tables are neither theoretical entities nor identical to them or collections of them. Scientific common sense would tell us that tables are, in some sense, collections (or structures) of theoretical entities; and since scientific realism embraces these entities (and presumably collections or structures of them) we are led to conclude
that there are tables. It would indeed follow that tables exist if it were true that tables were identical to (existing) structures or collections of theoretical entities. But are they identical? We have seen that Sellars denies their identity, on the basis of the indiscernability of identicals; for tables are solid and colored but theoretical structures are mostly empty space, colorless (since, on Sellars' view, it is sensa which are genuinely colored) structures of theoretical entities.

This raises a question central to ontology (and, in my view, to reference); what is identity? Quine makes the connection between identity and existence explicit in his motto "no entity without identity"\(^{14}\) in recent literature questions of identity have been approached within discussions of modality and reference. The next chapter will begin with a discussion of possible worlds realism and lead back to the question of identity. For the remainder of this chapter I will focus on other aspects of theoretical explanation and scientific realism.

Explanation

My positive account of scientific realism draws largely on the Sellarsian account just sketched. A

\(^{14}\) "Speaking of Objects", p. 23
theory explains macro-object \( \phi \)'s being \( \psi \) be describing \( \psi \) as theoretical item \( \phi \) and describing \( \omega \) as \( \psi \), where \( \omega \) is a theoretical property corresponding to \( \psi \).

Note that this is precisely the manner in which scientists actually do explain observed phenomena. For instance, layman Jones compresses a volume of air in a cylinder and asks why its temperature increases \( (\psi) \); the appropriate explanation is that the air \( (\phi) \) is a collection of molecules \( (\omega) \) and that heat is molecular motion and that compressing the air concentrates the molecular motion in a smaller volume, resulting in more motion in the area. The theory explains a phenomenon by redescribing it in a way which makes the phenomena understandable.

It must be noted that this manner of explaining observed events requires the reality of the postulated entities, at least if the truth of sentences required in explanation requires the reality of the theoretical entities those sentences refer to. Note also that this view is open to the possibility that the correct reference of terms is not a matter of a correspondence between referring expressions and a reality which is totally independent of all the possible ways of representing it. The theory explains relevant laws as well, by explaining the particular events which the laws
generalize. Finally, a theory is good to the degree that it explains phenomena (and hence laws) in its domain, and conforms to other theories which are similarly successful explainers. Thus, molecular and atomic theories explain observations and conform to one another.

With respect to reasoning about the mental, this account of explanation puts the following constraint on Psychological theory; the theory should conform to physical theory as well as explain observed phenomena. This constraint would seem to imply some form of materialism or identity theory; I will begin the examination of Psychological explanation in the next section, reserving questions about the relation between Psychological theory and Physical theory to Part II.
It will be convenient to distinguish between several temporal and/or dispositional senses of terms about thoughts or beliefs. If someone says

Jones thinks that p,

he could mean that Jones is, at this very moment, thinking that p; alternatively, he could mean that Jones, although occurrently thinking about q, is disposed to think p (as opposed to thinking not-p, or thinking that p may be the case or may not be the case) if some occasion or circumstance should bring the issue of p or not-p to mind. Ordinary usage allows the contrast between the occurrent and dispositional senses to be made by using the present progressive

... is thinking ...

for the occurrent sense, and the present

... thinks ...

for the dispositional. While the latter usage is almost always followed, the present progressive allows not only a pure occurrent (right now) sense, but also a sort of 'local' dispositional sense, illustrated by the case
where Black has recently been considering the p issue and has decided, at least for the time being, to accept p, but is at the moment (occasionally) thinking about (or that) q. In this case "Black is thinking that p" can imply the impossible contrast "but he will probably change his mind soon."

I should point out that on most classical views, mental states are considered to be occurrent states, not having any intrinsically dispositional features. It may, however, be desirable to analyze even occurrent thinkings in terms of dispositional features. If this is done, dispositions to have these occurrent thoughts would be second-order dispositions, i.e., dispositions to be in dispositional states. Sellars uses an electromagnet as an example of something having second order dispositions of this sort. When energized, the electromagnet is occurrently magnetic; but occurrent magnetism has the intrinsically dispositional feature of being such that if iron filings were brought near, they will be attracted, if a compass were brought near, it would point toward the magnet, etc. Now consider the electromagnet in its de-energized state; it is such that if energized, it becomes magnetic. That is, it has a disposition to acquire a disposition to attract iron, etc. This would be a second-order disposition.
The main point here is that states which appear to be 
occurrent may have intrinsically dispositional properties. 
This possibility will play an important role later.

The Alleged Infinity of Belief

"Believes" has these temporal senses and may add 
a further complication as well. It has sometimes been 
claimed that people have many more beliefs than they are 
aware of, perhaps even an infinite number of beliefs. 
According to this view, people who have studied both 
the moon and Mount Rushmore but who have never considered 
whether or not they are identical may believe that the 
moon is not identical to Mount Rushmore, and also, even 
though they have never considered the issue, believe 
that mount Rushmore is identical with Mount Rushmore. 
If this were correct, there would seem to be no limit 
to the number of propositions a person might uncon-
sciously believe, even though these propositions have 
ever crossed their minds. A person could, on this 
interpretation of 'belief,' believe an indenumerable 
number of mathematical truths; all the consequences 
of truths he consciously believes.

At first glance this would seem to pose a 
problem for a postulational account of thoughts which 
held that the postulated entities are (or consist of)
neurophysiological constituents, no one could possibly believe all these propositions.

This point is sometimes used not only as an objection to any form of materialism or identity theory, but as an argument for a relational theory of belief; i.e. one which holds that belief is a two-place (at least) relation connecting persons and propositions or states of affairs. This is because even though a person can have at most finitely many distinct neurophysiological states, he could be related to infinitely many propositions, etc. Surely, it is argued, everyone is related to each proposition by the relation of non-identity, among countless others. It is sometimes argued that if thought is a relation between a thinker and a proposition, then thinking must be non-physical. Of course, not all relational theories of belief are dualistic; some would hold, for instance, that when Jones believes P it is because his brain is in a linguistic state which expresses P. On this view, the relationship between a brain state and a proposition is the expression relation, and is the same as the relation between printed sentences and the propositions they express. Although that relation still requires explanation, this move reduces two mysteries to one, a net advantage.
Unfortunately this move is, of course, still open to the objection based on the infinity of a person's beliefs, since it would require distinct brain states to express each of infinitely many beliefs.

As we will see later, Sellars invokes a non-relational account of belief. According to this account, the connection between a person and the content of his belief is analogous to the connection between words and their content, i.e. what they express; but this 'connection' is not a genuine relation between two objects. As a nominalist, Sellars denies the real (or independent) existence of propositions and other so-called abstract entities; and what does not exist cannot be a genuine relatum. Put this bluntly, Sellars' account may seem implausible; its cogency rests on his account of the nature of abstracts, which we will consider in due course. Intuitively, though, the claim is that thinking is no more a relation between two independently existing objects, a thinker and a thought, than smiling is a relation between two objects, a smiler and a smile; few would find plausible a claim that smiles are objects which exist apart from smilers.

Returning to the problem of the infinity of belief, we will consider two related suggestions. They both share the claim that in the case of beliefs which
a person can be said to have even though he has never explicitly entertained them, the reason is that although he represents, in some form, all of the information or content of that belief, he has not assembled that information into a single mental item.

As will be seen, both suggestions are developments on the theme that an occurrent state may explain a second-order dispositional state. It seems clear that the sense in which a person may be correctly said to believe all of the consequences of what he believes in an ordinary sense is that the person is such that if he considers the consequence in question -- and has the power to draw the previously undrawn inference -- he would believe it in an ordinary sense. Thus, taking beliefs as having a first-order dispositional nature explained by the occurrent state of a person, these undrawn consequences are dispositions to believe; second-order dispositions.

The first suggestion holds that this is possible because most of a persons information is stored as 'dictionary' or 'encyclopedia' items in the persons 'library,' memory, and not as histories or the like. On this view it is appropriate to say a person believes that, say, Paris is not on the moon, even when he has never explicitly considered this possibility, because
he has the information required to reach that conclusion should the need arise. Similarly, he can be said to understand an infinity of sentences he has never spoken or encountered, and to judge many of them true or false, because he is in command of the lexical, semantic, and often factual information, stored piecemeal, together with the means of manipulating these pieces of information required to analyze and judge them. Note that the dictionary, of itself, is not sufficient — there must also exist a means of sorting through and arranging entries taken from it.

The second suggestion differs from this one primarily in claiming that some of the information required is stored in non-ropositional, nonlexical, form; either as images, or in what recent theorists have called 'frames.'

Now, in distinguishing various senses of terms referring to mental items, one must keep ones purposes in mind. On the one hand, such distinctions may be intended to clarify ordinary usage; here, one is not creating distinctions, but discovering those implicit in ordinary usage. On another approach, however, one may be creating a new set of meanings for mental terms, intended primarily to spell out a theory by attributing appropriate properties to the items postulated in the new theory, even though one uses existing terminology
to name the new entities. It could be argued that this approach is dangerously confusing, for people are likely to interpret such uses in traditional ways even though much of the traditional meaning of the terms invoked may not be consistent with the theory. Thus, for instance, when physicists like Bohr used the traditional term "atom" in the development of 20th century atomic theory, and said that atoms were divisible into even smaller parts, scoffers objected that that was absurd, since "atom" entailed "indivisible." In this case, theorists could have replied, "You are right; we were mistaken in thinking that the objects we called atoms were, in fact, atoms," instead of the reply "no, even though these items are divisible, we will call them atoms; our theory, not your Greek lexicon, determines what properties our atoms have, and stipulates a new definition for the word 'atom.'" That the world has accepted this stipulation may, in retrospect, be an accident of history, and arbitrary in the sense that there is no completely satisfactory justification for the new usage; but it stuck even though it violated traditional usage.

Similarly, in discussing thoughts, a theorist may adopt the term "thought" because it is the existing term which most closely approximates that which his
theory refers to, without being committed to all of the ordinary language entailments that traditional usage carries with it; and philosophical objections based on such entailments are strictly inappropriate.

As in the case of atomic theory, the degree to which the new theory and its associated definitions are adopted may depend to a large part on historical accident; or on a more positive side one might claim that the success of the new theory in explanation and prediction gives it such pragmatic value that its adoption is as inevitable as the survival of the fittest species. It is not my intent to quibble over these issues; my point is that it may be allowable and appropriate for a theorist to usurp an existing term, for special use in his theory, especially when his theory is designed to provide an improved, or scientific, account or explanation of phenomena to which traditional usage was relevantly tied. The long-run justification of such stipulative redefinition may be a matter of the success of the new theory.

In what follows I will generally speak of thoughts and beliefs in the purely occurrent sense if I do not specify otherwise; and since the issues I raise are of a very general nature, I will not distinguish between various species of thought but will
generally include such propositional attitudes and mental acts as desiring, hoping, and fearing -- in their occurrent sense -- as kinds of thought. Thus, the issue directly addressed involves primarily the nature of occurrent thought; and although I think that the following discussions will apply *mutatis mutandis* to other varieties of mental actualities (with the noted exception of sensory states), I will not argue for this assumption here.
Psychological metatheory is theory about psychological theory; it involves such questions as, "What is the domain of psychology," "How should psychological theories be constructed," and "What are criteria of adequacy for psychological theories." The major dispute in 20th century psychological metatheory has been over the postulation of theoretical entities; three major contenders in this dispute are the philosophical behaviorists, the physical reductionists, and what I shall call mentalists. Each of these general positions has the questions of whether or not internal states must be postulated in order to provide adequate psychological explanations and whether or not the language of physics is sufficient for explanations.

Philosophical behaviorists either deny that internal states need to be postulated or claim that explanations referring to them can be reduced to explanations in terms of observed behavior; physical reductionists claim that the language of physics is sufficient for psychology, i.e., that all statements referring to theoretical terms in psychology can be
reduced to statements couched in the primitives of theoretical physics. Mentalists assert that internal ('mental') states need to be postulated for explanations in psychology, but that there can be no adequate reduction to purely physical (or purely behavioral) terms. In this usage, 'Mentalist' is not synonymous with 'Dualist'; mentalism is consistent with physicalism, the thesis that all entities are physical entities.

Since Psychological metatheory is theory about Psychology, it would be convenient to know what psychology is. Unfortunately, I'm afraid that I cannot give an uncontroversial definition of psychology. The problem is that different psychological theories tend to describe the domain of psychology in different ways; some definitions come close to the etymological source of the expression, 'the science of the mind or soul (psyche),' while other psychologists would deny that there is a psyche to be studied. Of course, if one were to claim that the psyche is whatever it is that psychologists study, there would be little controversy (except, perhaps, over this misuse of language), but this answer yields little information as well. It does suggest a sort of Feyerabendian approach though; psychology is what psychologists do. (But this could go on a while: "What is a psychologist?" -- "a person who studies the psyche? . . . ," etc.)
What do psychologists study? People? For the most part I imagine this may be true; but then what distinguishes psychology from anthropology? And who are those guys running rats through mazes, and studying the behaviors of chimps and other animals, if psychologists study only people. Psychologists don't just study everything about people, nor do they focus only on people; they seem to focus on the behavior of people and other organisms which are, at least in some weak sense, like people. This seems to agree with intuition and authority; Jerry Fodor claims,

It is a point of definition that a science of psychology seeks to provide a systematic explanation of the behavior and of the behavioral capacities of organisms.¹

Two Domains of Observation: Rich Behavior, Poor Behavior

Thus, Psychological theories are at least theories about the behavior of animate organisms, especially people. But what, then, is behavior? While this question may seem oblique, a little investigation reveals that it can be especially troublesome for the behaviorist, for different answers carry different presuppositions. For instance, if behavior is merely physical movement or physiological reaction and is described in a purely

¹ Psychological Explanation, p. 3
physical, physiological and anatomical vocabulary, then, it may be claimed, behaviorism will not explain activities couched in our ordinary vocabulary of action, since different movements can be used in performing the same actions, and vice-versa. But if actions are to be explained, and the richer vocabulary of action is used in describing behavior, then it may be argued that behaviorism is committed to mentalist concepts, since action descriptions are typically intentional, being couched in the language of thought and belief.

The behaviorist thus seems to be faced with a catch-22: Either he uses an impoverished vocabulary to describe and explain phenomena, in which case the domain of his science will exclude much of what is interesting, or he uses a richer vocabulary and seemingly violates the anti-mentalist scruples of philosophical behaviorism. The situation is made worse for the philosophical behaviorist by the fact that while the restricted vocabulary may seem adequate in highly structured and controlled experimental circumstances, it does not seem readily applicable to cases outside such constraints. It is important to recognize that there is a broad range of richness and poverty of behavioral description. As we will see, Sellars shows how a form of rich methodological behaviorism couched in terms of speech
dispositions can be rich enough to provide an account of behavior in an interesting sense, but leads naturally into (and seems to require) an explanation of those speech dispositions in terms of inner states.²

Why should one thing that theories of behavior should postulate unobserved entities or use theoretical terms in their explanations? Well, if, as has been argued, behaviorist theories taking an impoverished domain of behavior (in the sense just discussed) fail to provide adequate explanations of human activity, and if richer forms of behaviorism require further explanation of their added richness, and if this further explanation cannot itself be reduced to a purely physical vocabulary, then the postulation of theoretical entities would be, if not forced, at least reasonable.

Behaviorism and linguistics

Can impoverished behaviorism provide adequate explanations of human activity? Consider explanations of language use and acquisition. It has been argued by Skinner and others that language use can be explained by means of complex functions relating observed stimulus conditions with observed verbal responses, and that

language learning can be explained by the theory of operant conditioning, making no reference to unobservables. However, as Chomsky points out in his review of Skinner, the data of language use and learning makes these claims highly improbable for two reasons. First, language users are capable of understanding and creating expressions which they have never before encountered; but behavioristic explanations, including those offered by Skinner as consistent with his theory of operant conditioning, depend on the relations between behavior and stimuli (captured in schedules of reinforcement), and explains only to behavior which is brought about through repeated conditioning. Even if the use of individual words or phrases was acquired through (and hence explained by) operant conditioning, the understanding of novel sentences requires that users be endowed with the ability to combine and permute combinations of words. The explanation of these abilities would seem to require some talk about the user's abilities over and above the facts of behavior alone.

Second, in language learning, the input (stimulus and reinforcement conditions) underdetermines the output (verbal behavior). That is, the total data available to the learner is insufficient to explain his
behavior unless his abilities, presumably resulting from his constitution, are considered. (These two arguments have been used by Fodor, who generally attributes them to Chomsky.)

Note that if the stimulus conditions and responses included in the domain of linguistic observables are couched in an impoverished vocabulary, in terms of patterns of ink on paper and of vibrations in air, not even in terms of words (much less their perceived meanings), these arguments have even more force, while they are weaker against behaviorist theories which include not only particular words but also sentence functions in their domain of observables.

To the degree that these arguments are plausible, they show that one must postulate unobserved abilities or attribute dispositional characteristics to language speakers in order to explain linguistic phenomena. But these abilities and dispositions seem to come down to the same thing from the viewpoint of explanation, since neither are explained in terms of behavior, and both seem to point toward internal states.

I do not intend these arguments to be taken as proofs of a mentalist approach to psychological theory.³

³ Fodor's The Language of Thought and Psychological Explanation are filled with alleged proofs of the necessity of mentalism.
My intent in this section is to show the initial plausibility of psychological theories incorporating terms supposedly referring to unobservables. Behaviorists might admit the terms but claim that they refer only to functions performed by observable persons in virtue of their acquired or innate dispositions, and a physical reductionist could admit the existence of entities causally mediating these functions but claim that a specification of these entities in purely physical terms is adequate for the psychological explanations of the observed regularities.

Arguments for these competing views often consist in arguments against their competitors. As physical reductionism argues against both accounts based on postulated mental entities and the richer versions of behaviorism, we shall begin with it.

**Physical Reductionism and Parsimonious Behaviorism**

Physical reductionism can be construed as a claim about the nature of the observables to be explained by a psychological theory. Mental terms can be reduced to physical terms if and only if the functions associated with mental terms are defined over the same domain and range (or subsets of the domain and range) as the functions associated with physical terms. (This is a logical feature of functions.) As Fodor has
pointed out it has been argued (by impoverished behaviorists) that the observables in psychology are motions and antecedent physical conditions. Motions are purely physical features of beings in the world, describable in terms of space-time location. The physical reductionist can embrace impoverished behaviorism (as a special case of itself), but must reject the view that any richer sense of behavior, construed as motion under some description involving non-physical terms, is observable. Two arguments for this view deserve close attention.

First, it is claimed that answers to why-questions about behavior construed richly must be in terms of reasons, not causes (We will examine Davidson's argument against the possibility that psychological explanations are causal explanations in chapter 8). Hence if the domain of psychology were behavior in a rich sense, it could not be a causal science; but this is scientifically unacceptable. As it turns out that this provides a better argument for the irreducibility of psychology to physics than for reductionism, I will delay a deeper look at this issue until considering arguments for the positive account developed in Part II.

Second, it is held that taking rich behavior, as distinct from mere motion, as a primitive leads to circularity; for this assumption presupposes that the non-physical aspect of behavior must be something which has to be interpreted by the theory which is supposed to be based on observation. Hence such intentional features as a person's seeing a newspaper as a newspaper rather than merely as a bunch of white stuff with black marks, or seeing a block of ice as something with a dispositional character (if warmed, it would melt), could not be properly included in the domain of explanation. Thus it is not, for instance, Jones' picking up the newspaper which is to be explained, but the motions of Jones' body as the physical collection of stuff we interpret as a newspaper is moved up.

Is this claim any good? Let us restate it for a case where there is an unobserved connection between two observed physical events, a button-pressing and light-glowing; here it is assumed that the physical connection between these two events, i.e. electrons flowing through a wire, is something which has to be interpreted by the theory which is supposed to be based on observation. This in turn would seem to imply that

5. Fodor, p. 167-68.
the button-pressing is observed as the establishment of an electrical connection. But if the physicist's observation of the button-pushing interpreted it as the establishment of an electrical connection, his explanation would, in a sense, be circular.

Is this restatement a proof that taking 'rich observations as primitive leads to circularity? If so, is the circularity objectionable? Perhaps so, but only if the choice of primitive observables is taken as an attempt to justify the science using those observables. Thus, attempting to justify the physical theory of electricity by saying that electrons are required to explain how establishing an electrical connection causes photons to be emitted from a resistive element in the electrical path would seem circular.

The alleged circularity comes from an attempt to use a definition of a kind of science as a justification for it; an appropriate justification is not derived from a definition alone, but from the need to explain the kind of observable regularities in question. Some of the major advances in science have come about by expanding the range of acceptable observation. The progress from mechanics to thermodynamics to molecular to atomic physics can be viewed as expansions of the domain of observables from motions to volume, compression
and temperature phenomena to electromagnetic phenomena (including light emission). In mentioning temperatures and electrical charges (as opposed to heat or shock) we might seem to be speaking of unobservables. But heat differences and sparks are observable, and it is hard to see how these can be construed as macroscopic motions. Furthermore, Quine and others have argued that the distinction between observable and theoretical entities is vague at best (is one observing a bacterium through a 2000X microscope?), and most observation terms are 'theory laden' in the sense that the theory puts constraints on what counts as an observation. But we do not, in general, reject a theory as circular because it defines its domain. Thus, accepting observations of behavior not specified in terms of motion, as the domain of psychology should not lead us to reject it, especially if less rich accounts fail to provide adequate explanations.

Physical Reductionism and Language

Another version of physical reductionism requires that all of the types spoken of in psychology be definable in terms of the types of physics, not just that each individual named in a psychological statement be namable using solely physical terms, for otherwise
physical language could provide physical names for Jones' belief that \( P \) and for Smith's belief that \( P \) but not able to show that both Jones and Smith had the same belief. Thus, to say that mental language can be reduced to physical language is to say that each class-term in the mental vocabulary is true of an individual if and only if some class-expression (which may be very complex) in the physical language is true of it.

But such reduction is typically not possible in the behavioral sciences. Linguistics provides an example of a case where the observation base of a science is not specifiable in terms of purely physical parameters, because linguistic functions are functions of, among other things, the uses to which expressions can be put and the behavior involving them. Thus, the token "Gift" is arguably, identical with a certain amount of ink distributed in a spatial array on this page, but there is no purely physical description (or 'impoverished' behavioral description which provides type-identity conditions for the English expression "Gift" as opposed to the German expression "Gift", for the two have wildly different linguistic functions which a 'rich' linguistics may include as 'observable' insofar as it allows the recognition of physical sign-designs as words in a language. (In English, "Gift"
functions as "object which is given, i.e., present" does: in German, it functions as "poison" does in English). The required type-identity conditions make irreplacable reference to the linguistic roles played by the term. It follows that physical language without information about linguistic functions, is too weak to make distinctions required in linguistics. This is borne out on the practical level by work in machine recognition, in which the most promising methods of classifying characters rely on heuristics based on the linguistic context in which the characters occur, and other information outside the scope of the purely orthographic pattern of the character. It may, perhaps, be objected that linguistics cannot take complete linguistic functions as observables; but it seems to me that the domain of a rich linguistics must include such obvious facts as 1) "Gift" is a word of English and 2) "Gift" has a different use in German than it does in English. These are included in the 'observables' of a rich linguistics in the sense that they are among the phenomena that such a science is to explain.

6. John Pollock objected that the linguistic functions are not observables; my response to this is that a rich linguistic science includes among its observables the fact that the expression "gift" has different functions in English and German.
On the other hand, it might be argued that any necessary condition, based on purely physical criteria, of tokens of a given linguistic type will be too strong: for any purely physical criterion of what counts as a token of the "Gift" type which includes no other expressions will rule out some genuine tokens of "Gift" because there are an infinite number of ways of instantiating tokens of that type, even without a given language. In the first place, there would seem to be an infinite number of possible media for such instantiations: carbon-ink on paper, pencil on cloth, tomato juice on silk ties, water vapor in air, morse code in electro-magnetic radiation, pebbles in sand, and so on. Worse yet, there are an infinite number of distinct typographical styles: Roman, Italic, Gothic, futura bold, architectural, and the initials in illuminated texts are only a beginning there will be permutations of every combination of styles, and further permutations created by varying, say the height-to-width ratio, the relative line thicknesses, and so on, not to mention pointillist typographic effects or typographical irregularities like letters with unintended gaps, e.g., "T". The problem is not merely the number of possible formations of a token, but that there is no reason to think that they can all be captured by a finite number
of mathematical rules of transformation not based on the extrinsic linguistic functions of the characters in question.

Now this should not be taken to imply that no purely physical device could be capable of recognizing physical patterns as words, at least in most cases; what it implies is that such devices will have to rely on heuristic procedures using linguistic information at a higher level than just the purely physical properties of the tokens: for instance, that the token in question occurred in a specified linguistic context (speaker, sentence, conversation or paragraph, facts about the speaker's language and environment, etc.). Recent work in language recognition presupposes such information and has been largely devoted to developing heuristic procedures based on such knowledge. It should be pointed out that the problem of recognizing linguistic expressions is a special case of the problem of object recognition, for which the use of such 'knowledge based' heuristics was first proposed. The more specific field cryptanalysis has long used such linguistic information as the language (probably) encrypted, the frequency of distribution of letters in that language, and hypotheses about the topic of the encrypted message in approaching their problems.
While this may fail as an absolute killer, it seems convincing as a burden of proof argument; for until there is at least some degree of success in explaining such a linguistically basic functions as reliable character or word recognition based solely on physical features, there can be no reason to think that they physical reductionist has a prayer of explaining more complex linguistic or psychological phenomena.

Worse yet, the study of translation between languages and codes, languages and languages, and within a language (synonymy) introduces a radically new twist; for in this study there is a sense in which the English expression "Poison" is of the same type as the German expression "Gift", and in which the coded expression "eatcsmnis" is of the same type as the ordinary English expression "semantics," and in which the English expressions "bachelor" and "unmarried marriagable man" are of the same type. We will return to a discussion of function and meaning in chapters 8 and 9.

Assuming for the moment that these arguments provide a prima facia demonstration of the impossibility

7. "Eatcsmnis" results from "semantics" by taking every other letter, beginning with the first, out of the word, then appending these to the end of the result, generating a rule-based anagram).
of reducing linguistics to physics, it follows that it is likewise impossible to reduce psychology to physics, for included in the domain psychology is linguistic behavior.

And even if the argument fails to prove the impossibility of such reductions in principle, it demonstrates their "impossibility" in practice. Thus it constitutes at least a prima facie objection to physical reductionism. Until a method of obtaining such reductions is proposed, tested, and acquires a high degree of confirmation, there is no reason to accept the reductionist view.

Of course, the reductionist may respond by saying that so long as reductionism is not disproved in principle, it is reasonable to use the reductionist thesis as the basis of a research program. We will consider several arguments against the possibility -- in principle -- of reductionism in sections to come, but at this stage it should be clear that the mere claim that reductionism is possible cannot be used to preclude other research strategies, especially those with less dubious presuppositions.

It should also be obvious that the failure of physical reductionism does not entail the falsity of the

---

8. Cf. chapters 7 and 8.
physicalist hypothesis that only physical objects and events exist; for even without a suitable translation of psychological theory into the physical language, individual mental states may be identifiable as neurological states, just as individual expressions in a language, e.g. this token: "Gift," can be described in the physical language ('a two-dimensional array of ink on paper with the dimensions . . .') without there being acceptable physical criteria for tokens of that type. This example is especially appropriate for cognitive psychological theories such as those proposed by Sellars, Fodor, Aune, and a multitude of Johnny-come-latelysts in which an internal language, whose tokens are brain states which intermediate observed relations, is postulated. We will return later to the relationship between functionally classified tokens and their physical constituents in considering varieties of identity.

Mentalism and Rich Reductive Behaviorism

Having decided for the moment physical reductionism is practically untenable, let us consider the dispute between two competitors which agree as to the realm of observables to be explained. Both mentalism and rich behaviorism hold that psychological theories ought to explain regularities in observed
behavior, i.e., regularities between stimulus conditions and responses, where observation is not necessarily limited to the observables of physical theory.

As I have previously argued, philosophical behaviorism, as opposed to more tolerant (and often richer) methodological behaviorism, can be viewed as a form of instrumentalism according to which mental states and events don't actually exist: talk about them is merely a convenient instrument in calculating explanations and predictions. Although it is contrary to my usual characterization of his position, the Philosophical behaviorist may try to have his cake and eat it too, by adopting a rich domain of observation and claiming that it provides a sufficient explanation of all interesting behavior, i.e. that all psychological phenomena can be reduced to it. Thus construed, the dispute between philosophical behaviorism and mentalism is of a piece with the dispute between instrumentalism and scientific realism. What these two theories share is the view that psychological explanations involve functions between observables, relating environmental conditions (including stimuli) to behavior.

As we shall see, Dennett's view in "Intensional Systems" emphasizes the functions characterized in terms of intensional states but considers the question of
whether these states exist to be a pseudoquestion. Thus his position seems clearly in this instrumentalist tradition.

Levels in Psychology: Inductive Generalization and Theories.

In the second half of "Explanations in Psychology," Fodor argues that psychological explanation is a two stage process and discusses the stages. He assumes that psychology is a systematic attempt to explain and predict the behavior of organisms and that at any instant behavior is the product of an organism's current stimuli and internal state (which is in part a product of previous stimuli). In the first stage, functions relating stimuli to behavior are postulated on the basis of observed relationships. Functional terms are introduced and it is assumed that internal states and processes are responsible for mediating the functionally related input (stimuli) and output (behavior). A stage one explanation is adequate to the degree that it is capable of predicting and explaining behavior given the current stimulus situation. Explanations at this state are not causal explanations, as no particular mechanisms are proposed as causally mediating observed and predicted relations. This sort of functional

explanation is compatible with many theories about the physiological -- and causal -- structure of the organism. The second stage is the stage at which a theory specifying the causal processes involved takes place. This state must take into account macro- and microneurophysiology and aims at discovering causal mechanisms which underlie the functionally described processes of the first stage.

Although Fodor does not explicitly reject this view of levels in his later book *The Language of Thought*, it seems to offer the reductionist a reply to one of his arguments against reductionism. I have placed a discussion of his views from *Language of Thought* in an appendix to this chapter.

Like Hempel's two-level view discussed earlier, this two-stage analysis has both instrumentalist and scientific realist aspects. In the first stage, functional terms (apparently describing dispositions) serve as devices for calculating in making predictions and formulating explanations. Historically behaviorists have denied the existence of entities postulated by mentalists to causally mediate the observed functional relations. This denial is often justified on the basis of parsimony. When behaviorism is directed against dualism, the doctrine that there are two kinds of existing substances, mental and material, the argument
asserts that since we don't postulate invisible gremlins to explain the observed behavior of watches (we wind them, and then they tick and their hands rotate), we shouldn't postulate minds (mental substance, what Ryle called the 'ghost in the machine') to explain human behavior.

This argument cannot be used against the sort of mentalism we have been discussing, for it does not assert that there are things which are not physical; it asserts that there are mental items and leaves open the question of what they consist of. Using the watch analogy, our mentalist does not postulate an invisible genie, but rather, unobserved clockwork, including perhaps unobserved handturners and energy-storers, which turn out to be physical stuff fulfilling functional descriptions. Cognitive psychology, a form of mentalism, makes explicit use of a machine analogy. Animate organisms are viewed as indeterministic computing machines, some of whose internal states are so organized as to be a machine language. Their behavior resists a completely physical analysis for roughly the same reasons that computations in general do; the regularities observed are more logical than physical in nature.

If there is a second sort of existence, it is the sort of 'existence' had by abstract 'entities' such
as numbers and functions, which can be interpreted such as to avoid any ontological commitment, not the sort of existence that Cartesian mental substance is supposed to have.

Thus, Fodor's first stage is common to both behaviorism and mentalism, while the second stage is rejected by the rich philosophical behaviroist but embraced by the mentalist. In the second stage it is assumed that there are unobserved entities and mechanisms which causally mediate the functions discovered in the first stage. Fodor supposes that we will find that neurological mechanisms mediate physiological functions. Thus, the second stage allows a means of exploring and establishing relations between psychology and the physical sciences; a second-stage explanation is adequate to the degree that it permits scientific systematization by rendering explanations produced by different scientific theories compatible and intelligible.

This provides at last an argument, alluded to in my initial characterization of the philosophical behaviorist as committed to an impoverished basis of observation, against the rich reductive behaviorist. It is not that he cannot provide some degree of explanation in terms of his basis, but that he disregards the possibility of explaining the functions (including
functionally or dispositionally classified behaviors) he countenances as observable in terms of the underlying structure of the organism exhibiting them. It is as if he would be satisfied with an explanation of what a clock does in terms of its propensity to move its hands when wound, without considering what it is about the clock that accounts for this 'rich behavioral description.'

As long as Fodor's two stages of explanation are not related as levels, with the theoretical level explaining laws which in turn explain observed regularities, a la Hempel, but are related as (temporal) stages in the development of science, where the discovery of regularities which seem law-like suggests the need for a theory which explains individual observations and why they exhibit the regularities in question, his view is consistent with the Sellarsian account of explanation arrived at in the last section. Thus the argument there which concluded that a realistic form of theoretical explanation is to be preferred to an instrumentalist or simple covering-law theory can be called upon in the dispute between behaviorism and mentalism to favor mentalistic psychological explanations and hence a realistic interpretation of psychological theories which postulate mental entities, states and processes.
In Part II the relation of these mental items to physical ones will be investigated.

Appendix: Fodor's View

Jerrold Fodor's arguments for mental items are based on the requirements of psychological theory. Fodor is initially concerned to refute two forms of reductionism which would "vitiate the psychologist's claim to study mental phenomena." Both would do so by showing that there really aren't any mental phenomena distinct from behavior on the one hand, or from physiology on the other. The first form of reductionism, logical behaviorism à la Ryle and Skinner, would 'reduce' mental phenomena to behavioral phenomena by providing logical equivalences for mental terms in terms of behavior. Fodor responds to this supposed attack on psychology by first trotting out standard skeptical objections to the logical behaviorist's view that all mental talk can be eliminated, or translated without loss of cognitive significance, into purely behavioral talk. He then points out that even if the logical behaviorists equivalence claim was true, it would not eliminate the need to provide causal explanations of behavior; and that the only plausible candidates for theories of behavioral etiology locate the immediate
causes of behavior in the internal constitution and
operation of the organism which is behaving. Fodor is
aware that this argument cannot support a dualistic
theory of mind; but if mental processes are, in some
sense, physiological inner processes then a psychologist
could explain the etiology of behavior in mental terms
referring to physiological states. Fodor's attack on
logical behaviorism is an attack on the view that there
is a conceptual tie, or logical equivalence, between
mental and behavioral terms; and his argument is
designed to show that the logical behaviorists have not
provided \textit{a priori} reasons for thinking that the sort of
psychological explanation of behavioral etiology just
sketched cannot be given nor reasons for thinking that
a program of research designed to discover such explana-
tions must fail. Fodor seems to suppose here that without
an \textit{a priori} argument against his form of mentalism, the
logical behaviorist cannot defend his claim that there
is a conceptual tie between mental and behavioral terms.

Fodor then turns to a second form of reductionism,
suggested in part by the rejoinder to the logical
behaviorist. If mental states are physiological states,
in any sense, cannot all mental talk be reduced to
physiological talk? And if a physiological reduction is
available, can psychology retain its status as a
special science? Fodor apparently thinks not, for his
arguments against reductionism are intended to save psychology. He apparently fails to realize that even where a reduction of one science to another is available, as there is between chemistry and atomic physics, the 'reduced' science is not thereby eliminated, for it still retains its own methodology and domain of phenomena to be explained, even if it resorts to the reducing science at times to explain laws governing its special domain. Thus the reductionist has a reply to Fodor's claim that psychology is a special science based on an account of levels of science, according to which psychology remains a special methodological level.
PART II: PHYSICAL REDUCTION AND IDENTITY

CHAPTER VII

INTRODUCTION

In this part I examine in greater detail recent arguments concerning the relation of mental to physical items, especially those relevant to various forms of physicalism. One variety of physicalism says that the basic theoretical entities of physics are ontologically basic, that only these entities and structures of them exist, and that everything is ultimately explainable in terms of these basic items and their occurrent properties. This view holds that physics is the ultimate science, and that a unified science based on physics is 'around the corner,' waiting for a mode of reduction which will define the primitives of other sciences. It allows the other sciences their separate existences on methodological, practical, and heuristic grounds but claims they are reducible, in principle, to physics. I call this view Reductive Physicalism: its two aspects, the reduction thesis and the unification thesis are logically separable, but linked historically, especially in the logical positivist tradition.
Chapters 8 and 10 focus on two approaches to reductive physicalism, one based on a variety of the identity theory, the other incorporating bridge laws. Arguments against the former approach lead to the latter; arguments against it lead to chapters 11 and 12, on two forms of non-reductive physicalism. The non-reductive physicalist holds that the theoretical entities of physics are ontologically basic, and that only these basic items and structures composed of them exist, but does not claim that all properties of such structures can be ultimately explained in terms of the occurrent properties of basic items. Chapter 12 ends with a defense of a version of non-reductive physicalism and leads to a refrain on the threat of instrumentalism, and on the promises of reductionism and the utility -- or futility -- of pursuing them as a research strategy.

Before embarking on the upcoming examination of physicalism, I will mention some factors motivating physicalists and their opponents. Ockham's razor, the principle which is often stated as 'posit not without necessity' provides a major argument for the reductionist; he feels that it has led modern science away from such fantasies as witchcraft, demon possession, magic, aether and phlogiston and is willing to follow its lead toward dismissing other mysterious metaphysical entities such as souls, spirits, and mental substance.
A major thrust of the logical empiricists of the Vienna circle was the development of a scientific philosophy incorporating an empirical methodology which dismissed much of traditional metaphysics as meaningless pseudo-problems. They were convinced that empirical science alone could provide a rational measure of the world and reflected it as it really was.

Many factors have lead away from early logical empiricism. The verificationist principle of meaning, according to which a sentence was meaningful only if it could be empirically verified or disproved, failed its own test. As shown in Part I, recent explorations of scientific explanation have led toward explanation in terms of postulated, usually unobserved entities, as opposed to instrumental explanations. Recent work in psychology suggests that mental items must be postulated in adequate psychological theories. Nonetheless, the spirit of the logical positivist moves many to seek out a parsimonious ontology, reducing the entities postulated by the special sciences to those of physics in order to achieve a single, united, coherent ultimate world view.

On the other side of the coin, religious, ethical and epistemological factors continue to motivate

dualism, as they have at least since Descartes. For if only physical objects exist, can there be immortal souls, or god? Or if humans are purely physical and human actions are either ultimately determined by physical law, or ultimately based on the random occurrence of indeterministic events, how can it be reasonable to hold persons responsible for their actions, to praise or blame them for what they have no control over? And if there is no mind, how can we introspect as we do and find it in an act of thinking, how could we be directly aware of our mental states, our sensings and believings? Finally, without this awareness, how could we have a firm foundation to build our knowledge on? Of course, a lot is required to turn these motivations into arguments. Part II is devoted to an examination of the possibility of physicalism in the face of these factors; part III is devoted to questions of personal identity and knowledge in the light of the kind of theory discussed in part II.
CHAPTER VIII
REDUCTIONISM AND IDENTITY

The Identity Theory and Its Problems

I will begin this section by briefly recounting the history of a version of Identity Theory to provide both historical background and philosophical motivation for the following discussion of identity and of the relation between the physical and the mental.

In the 1950's a number of philosophers - including U.T. Place, J.J.C. Smart and D.M. Armstrong - advanced versions of the theses that mental items are identical to certain physical items. Sensations, thoughts, beliefs, and consciousness itself were said to be identical to brain states, C-fiber firings and/or other physical items in the body.

Identity and Discernibility

One traditional argument against the identity theory is based on the principle of indiscernibility of identicals (I=), which states that if any items x and y are identical, then they have all properties in common:

(I=): (x)(y) [x=y \rightarrow (\phi) (\phi x \leftrightarrow \phi y)]
Two premises and this principle yield the result that sensations are not brain states:

1. There are blue pentagonal sensations.
2. There are no blue pentagonal brain states.

Since the range of physical objects and events which anyone would be tempted to identify as sensations excludes things which are blue and pentagonal, this result is generalizable to

No blue pentagonal sensation is identical with any physical item (within the range...)

To be sound, this argument requires, of course, that 'blue' is univocal in 1 and 2, and that each of them be true. Each seem, at first glance, true, since people ordinarily believe that nothing in the brain is blue and pentagonal, and that since after-images are sensations, the result of staring at a bright yellow pentagon is a blue pentagonal sensation. Dualists like Descartes would not allow that sensations have shape and color in the same sense, for insofar as mental states are unextended, they do not have shape; but they would be satisfied with color alone, for material things do not have color.

I= generates many alleged counterexamples to versions of the identity theory:
Indubitability:

a) I cannot doubt that I think;
b) I can doubt that I have a body (or any other physical parts or properties)

Thus, c) Thinking is not a bodily process

Incorrigibility:

a) I cannot be wrong when it seems to me that I have a red sensation.
b) I can be wrong about my brain states.

Thus, c) Red sensations are not brain states

Color:

a) Some sensations have vivid colors that never occur in the body.

Thus, b) Sensations are not in the body

Truth:

My thoughts can be true or false; it makes no sense to say of a brain state, etc., that it is true or false.

Intentionality ('aboutness')

my thoughts are intentional, i.e. they are 'about' objects in a way which it makes no sense to say that merely physical things are about objects.

Now it must be noted that the identity theorists were quite suspicious of Indubitability, Incorrigibility and other epistemic or doxastic properties of mental states, arguing that they introduced violations of I= in even the most clear cases of identity, insofar as they seem to rely more on what an object is called
than on what it is. They pointed out correctly that
the fact that a person believes that Cicero was an
orator but does not believe that Tully was an orator
could not establish the nonidentity of Cicero and Tully,
for the appearance that this shows that Cicero and Tully
have different properties is merely an appearance,
depending on not recognizing the difference between
having a belief about an object under a description or
name and having a belief about the object itself,
independent of what it is called. In the latter sense,
anyone who believes of the person (object) Cicero
himself that he is an orator, believes it of Tully,
the very same person, even though the believer may never
have heard the name "Tully."

The dualist has a response to this with respect
to incorrigible belief, since in incorrigible belief,
the object of belief is said to be directly apprehended,
not known via name or description, and known in such
a way that all of the objects properties are directly
an immediately perceived. These ways (or modes) of
knowing an object are thus thought to be immune to the
possible confusion of belief about an object itself and
belief about it under a description.

This defense rests on doxastic modalities which
seem to presuppose epistemological theories requiring
the antecedent truth of dualism; I will take up the issue of direct apprehension in chapter 14, where it will be argued that a plausible epistemology consistent with physicalism is available. For the moment, however, we must note that I= raises prima facie problems for identity even if these doxastic modalities are called into question, for there remain two problems; the intentionality of belief, and the 'ordinary' properties of sensations.

On the other hand, the principle of the identity of indiscernibles (=I), which states that indiscernible items are identical

\[(=I): (x)(y)\{(\psi)[\psi x \leftrightarrow \psi y]+ x=y}\]

can also pose problems for the identity theorist, at least if \(\psi\) is restricted to non-relational properties, for it has been claimed that it is possible for two people to be in the same mental state (the states are indiscernible) without being in the same physical state. Thus, for instance, it can be argued that no physical state of mine is identical to any physical state of Keith Lehrer's, but we both believe that Tucson is in Arizona. That is, we have the same belief, but are in distinct physical states. Similarly, it is claimed that a person can be in indistinguishable mental states on different occasions, yet be in distinct
physical states on those occasions. These puzzles seem to evaporate when one draws the distinction between thoughts as types and thoughts as tokens (Lehrer and I have distinct thought-tokens of the same type; if his thought-token is one of his physical-tokens, and likewise for mine, there is no problem; likewise, if our (shared) thought types are identical with our (shared) physical types, there is no problem. Of course, this response requires a theory making an acceptable type-token distinction. In a generous mood, one might say that the point of the alleged counterexample is to bring this out.

Thus both of Leibniz's principles relating identity to properties can be said to be violated by versions of the identity theory, and it is the burden of the identity theorist to provide satisfactory solutions to them. Some lines of response to these objections either deny that the mental states in question literally have properties not had by neurological states

1. I do not wish to engage in the dispute as to which of these principles is more properly attributed to Leibniz; I= is often called 'Leibniz' Law,' though sometimes this phrase is used to identify the (much stronger) combination of I= and =I, the claim that two things are identical if and only if they share all properties. See Hide Ishiguro's Leibniz's Philosophy of Logic and Language, especially ch. 2, for more on Leibniz' views.
or assert that, contrary to the supposition, the physical states do have mental properties. As an example of the former strategy, it is argued that sensations are not literally blue but are 'of' blue, or are blue in the sense that they represent blue things or have a 'blue' content. For the latter, it is argued that brain states can be true or false in exactly the same way that particular tokens of sentences are, since both are physical objects having semantic properties. We will return to these strategies as they apply to the defense of 'token identity' in chapter 11 and 12. We will turn to other objections to type-identity after digressing to draw the distinction.

Type and Token Identity

The identity theory can be interpreted in two ways which are appropriately referred to as theses of type identity and theses of token identity. The type identity thesis has been thought of as required for a reduction of mental states to physical states since it parallels the 'identities' which had been claimed to exist between observable and scientific entities. In the earlier cases, physicalists held versions of the physical identity thesis (PIT):
(PIT): every kind of observable macrophysical object and/or process is identical with some kind of system of scientific objects and/or processes.

Thus, Physicalists asserted, such commonplace identifications as lightning with electrical discharge or heat with molecular motion were claimed to be literally true.

Type-identity theorists supposed that there were parallel identities between types of mental items and types of physical items:

TI (Type Identity): for every mental state type A there is a physical state type B such that (x)(Ax=Bx).

According to this view, a mental state (type) like pain is identical with some type of neurophysiological state or process, say the stimulation of C-fibers.

In contrast, the thesis of token, or particular, identity holds merely that each mental particular is a physical particular, or token, and makes no claims about whether mental properties or kinds can be identified with physical properties or kinds.

Kripke's Essentialist Objection

Saul Kripke has recently raised an objection to identity theories, based on the claim that all identities are necessary; to falsify TI it is sufficient that it be possible for there to be a being with C-fiber
stimulation but no pain, or vice versa. Kripke assumes that such a being is possible since we can surely imagine a sentient thinking creature with a neurophysiological structure different from our own. Here Kripke seems to have ignored his own injunction against conflating that which is imaginable with that which is possible; but this response can only be used as a partial defense of type-identity: without an argument that such an imagined state of affairs is not possible, the type identity theorists can only invoke Kripke's injunction to shift the burden of proof back to Kripke, insisting that he show that his alleged counterexample is a genuine counterexample by showing that the imagined state of affairs is not possible, the type identity theorists can only invoke Kripke's injunction to shift the burden of proof back to Kripke, insisting that he show that his alleged counterexample is a genuine counterexample by showing that the imagined state of affairs is in fact possible. But in the absence of stronger evidence one way or the other, Kripke's essentialist claim does seem at least initially plausible. Another line of response to the essentialist objection is that the identity is merely a contingent identity.
Conventional Necessity and Contingent Identity.

One way to argue for contingent identity is to argue that the essential properties of an object do not determine its identity conditions because 'essential' properties are not intrinsic to objects but are rather based on linguistic convention. This approach is reminiscent of Locke's doctrine of nominal essences, in the sense that it claims that all necessity is verbal (nominal) necessity, i.e., that all necessity is analyzable as de dicto necessity, a matter of necessary truth based on the (merely conventional) definitions of words. Thus there are no genuine (i.e. language-independent) necessities inherent in or intrinsic to (de re) things themselves.

It would follow from this that even though sensations of blue patches are, say, c-fiber firings ('c-fiber firings' is used as a sort of promissory place holder for the as yet undiscovered physiological description of the brain items), they are so only contingently; if things had gone differently, they might have been something quite different. As to other alleged essential differences between the mental and the physiological (physical stuff is essentially locatable, thoughts are not, etc.), the type-identity theorist would claim that they had just been language,
which we should no excise by adopting new conventions. After all, they're our conventions, and if we decide to change them, we can.

The thesis of contingent identity was thought to gain additional support from the fact that many scientific identities seem to have been empirically discovered; hence, the story goes, they are known a posteriori, thus they are synthetic, and thus they must be contingent. But note, first, that although identity is contingent on this account, it makes little sense to say that identities can be empirically discovered if they are at heart a matter of convention. And second, as Kripke points out, it is plausible to think that the apparent contingency of such identity statements is due to confusing the genuine identity statements with contingent empirical counterparts of them. As this rejoinder requires elucidation, we turn to Kripke.

Kripke and Necessary Identity

Saul Kripke has placed the contingent identity approach in jeopardy by providing a systematic account of reference and identity within which all identities are necessary. While a complete examination of Kripke's theory goes far beyond the scope of this occasion, several points need to be made. His account allows for de re necessity and avoids difficulties encountered
by attempts to account for reference in terms of come (tacitly accepted) definition. Briefly, he argues that an individual can be referred to without invoking a description (or the Fregian sense of a definition). If this is so then the identity of an individual does not depend on its contingent descriptive properties or on any merely contingent facts about linguistic conventions based on them. Proper names refer not in virtue of any description or cluster of descriptions or of their descriptive senses, but in virtue of haven been attached to objects by people. Thus, Kripke continues, we can imagine that object in possible circumstances where it has none of its contingently identifying characteristics, by 'building' (or stipulating) a world around it, given that we began by recognizing it in virtue of certain natural features, notably its causal genesis and constitution, i.e. by being able to recognize what kind of thing it was that the name was attached to. Thus there is a connection in res between individuals and natural kinds.

On Kripke's view, proper names and natural kind terms generally designate rigidly, i.e. pick out the same individuals and kinds of things in all the possible worlds where they exist; and if an individual actually is one of a certain natural kind, then it must be of that kind, since it was its kind (causal, genetic.
and constituitive features) that enabled us to pick it out to be named in the first place. Thus, a thing’s natural kind, and in particular, its constitution, is essential to it, de re.

This hurts any identity theory that attempts, as it seems all type-identity theories must, to identify the 'natural' kinds of psychology with those of physiology; for if a thought is a certain kind of brain state or event, identified in the kinds of physics or physiology, then it is necessarily (essentially) of that kind, regardless of our linguistic conventions.

Gibbard's Defense of Contingent Identity

Alan Gibbard has argued that, contrary to Kripke's claim, there are contingent identities. Gibbard argues that we can imagine a quite ordinary case in which 'two' things are identical, but might not have been. His example is this: imagine two lumps of clay, one carved into the top half of a statue of baby Goliath, the other into the bottom half. At time t the lumps are joined creating one new lump (call it 'Lumpl') and a new statue ('Goliath'). Later, at t2, the statue/lump is smashed. Gibbard claims that Lumpl and Goliath are identical, but might not have been. In the actual world Lumpl and Goliath shared all
properties, including spatio-temporal properties; but since Goliath could have been destroyed by squashing Lumpl into a different shape - without destroying the lump - Goliath and Lumpl could have had different properties.

One way of looking at the dispute between Gibbard and Kripke is as a dispute about whether an object's essential properties depend on what sort of object it is considered to be, or whether it has them independent of what one considers it to be; on this view, Gibbard would be seen as defending a sophisticated version of the view that all necessity is verbal necessity. But it seems more profitable here to view Gibbard as addressing the issue of what an object is than that of the supposed primacy of words to properties.

What does Gibbard's account hinge on, if not the reducibility of de re to de dicto necessity? One possibility is this: that what seems to be one thing might belong to different natural kinds and hence have different things it could be. But the thing in question is, on the face of it, just one thing. This approach seems to depend on the view according to which x and y are identical in a world if they share all properties in that world (properties are thought of existing in worlds). In any case, it seems that we are faced with a dilemma; either Lumpl and Goliath
are identical though they might not have been, or there are actually two objects where there seems to be just one. Gibbard takes the former to be the more intuitive horn.

A more radical view is suggested by David Lewis' Modal semantics. He denies that individuals exist in more than one world. Thus the possibility that there could be individuals with C-fiber stimulation but without pain does not constitute a counterexample to the view that actual C-fiber firings are identical with actual pain. Attempts to defend the type-identity theory along these lines must be examined in the context of a wider investigation of individuals, possibility, possible worlds, and conceptual systems, to which I will now digress.
CHAPTER IX

POSSIBLE WORLDS AND CONCEPTUAL SYSTEMS

This chapter begins with a long digression on the status of possible worlds and the things in them, focusing at first on the account given by David Lewis. Following this discussion, we will return to consider the preceding dispute over identity. I will argue that a significant part of the problem lies in not distinguishing between possible worlds and the conceptual structures which underlie them, and will develop this distinction by focusing on a notion of counterpart different from that of Lewis. We will see that a form of token identity theory can survive without requiring this distinction; but in chapter 10 we will find an argument to the effect that no such account will allow psychological explanations to be causal explanations.

Counterparts, Identity and Modal Realism

Among the possible views about possible worlds, David Lewis holds a very strong form of possible worlds realism; his position includes the views that

1) possible worlds exist; they are real in the sense that they are "respectable entities in their own right."
2) Nothing in any possible world is identical with anything in any other world.²

3) Unactualized possibles, i.e. things in possible worlds other than the actual world, exist,³ although

4) they do not actually exist, i.e. exist in the actual world.⁴ Nonetheless,

5) possible worlds are the sort of thing that the actual world is.⁵

In this section, I wish to dispute these claims, at least as they are interpreted by Lewis. I wish to argue that

1) Possible worlds may exist in some sense, but that they are not 'respectable' entities as are actual physical objects

2) objects existing in one world can be identical with objects existing in other worlds

3,4) Some objects existing in merely possible worlds exist in the actual world, and

5) the expression 'Actual World' is ambiguous; in one sense the actual world is not the sort of thing that a possible world is, but is a radically different sort of thing.

The discussion of possible worlds is divided into three sections. In the first section, we will consider Lewis' arguments for his version of possible worlds realism. In the second, these arguments will be brought into question. The third section will explore implications of some of Lewis' premises, and show how a less radical possible worlds realism can explain the intuitively correct parts of them. When we return
to the question of the relation of the objects postulated by psychological theory to those of physics we will draw directly on several of the issues discussed here, and find that this sort of discussion overlooks the importance of the distinction between worlds and the systems in which they are possible.

Arguments for Lewis' View

Arguments for Possible-Worlds Realism.

In chapter 4 of *Counterfactuals*, Lewis presents an argument for the reality of possible worlds:

I believe that there are possible worlds other than the one we happen to inhabit. If an argument is wanted, it is this. It is uncontroversially true that things might be otherwise than they are. I believe, and so do you, that things could have been different in countless ways. But what does this mean? Ordinary language permits the paraphrase: that there are many ways things could have been besides the way they actually are. On the face of it, this sentence is an existential quantification. It says that there exist many entities of a certain description, to wit 'ways things could have been.' I believe things could have been different in countless ways; I believe permissible paraphrases of what I believe; taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called 'ways things might have been.' I prefer to call them 'possible worlds.'

This argument can be analyzed as follows (adding Lewis' notions of real and actual):

---

(R1), things could have been different in many ways;

(R2), thus, there are many ways things could have been besides the way things actually are.

(R3). Possible worlds are ways things could have been; the actual world is the way things actually are;

(R4). Therefore, there are many possible worlds besides the actual world.

(R5). Things that are exist; and things that exist are real (though not necessarily actual).

(R6). Therefore, possible worlds exist, i.e., are real.

Lewis does not attempt to justify R3; he offers it as a stipulative definition, his preferred mode of speech. (R1) is presented as something he believes, and assumes is a reasonable, ordinary belief, while R2 is, he thinks, an acceptable paraphrase of R1. R4 and R5 present my reading of part of Lewis' interpretation of 'real,' 'exists,' and 'actual.'

A slight extension of this argument yields the result that unactualized possibles exist in non-actual possible worlds. Roughly, if a thing could be a way that nothing actual is, then there is a way things could be which includes that thing. While Lewis does not present this argument, he does provide a translation schema\{7\} which allows the translation of

\[ \langle \rangle \; Ra \ ^{\^} (x) \; \sim \; Rx \]
Arguments Against Trans-world Identity

The thesis of trans-world identity (TWI hereafter) in modal logic is the thesis that things in different worlds can be identical. It implies the thesis that one thing can exist in two or more worlds: Lewis denies both of these claims.

The paragraph containing the argument paraphrased as (R) also contains a premise for an argument for the view that things in different worlds are non-identical:

(D)  (D1) Things could have been different than they actually are. (P1)

(D2) But it is not possible for the things that are different than they actually are to be the same things that they actually are. (Lewis argues for this on p. 39)

(D3) Thus, if a thing had been different than it actually is, it would have been a different thing than it actually is, i.e., they would be non-identical.

(D4) But any world could have been the actual world

(5) Thus if a thing had been different than it was in any world, then it would have been a thing distinct from its actual counterpart.
The completion of this argument requires either that no thing exists outside our world, or that if a thing exists in a world, then all the things it would have been if it had been different -- its counterparts -- exist in other worlds. Lewis opts for the latter view; in fact it is the third postulate is of his Counterpart Theory.²

One of Lewis' arguments against TWI rests on one of Quine's arguments against unactualized possibles. According to Lewis, Quine objects to them because it is not clear when they are or are not identical. By getting rid of transworld identity, the only remaining questions of identity are reduced to ordinary questions of individuation within worlds. Lewis introduces a counterpart relation to account for the relation between actual objects and the things they would have been if the world had been different. The counterpart relation is defined in terms of similarity.

In general: something had for counterparts at a given world those things existing there that resemble it closely enough in important respects of intrinsic quality and extrinsic relations, and that resemble it no less closely than do other things existing there.³


That the counterpart relation is problematic is to be expected, since similarity relations are generally problematic; shifting from TWI to counterpart theory saves identity from the vagueness of similarity.

Finally, Lewis notes that trans-world identity theorists would be mistaken to think TWI equivalent to counterpart theory and claims that the latter is to be preferred to the former on the basis of its greater generality. Although TWI-theorists can create an analog to Lewis' world-bound individuals by considering as individuals <thing, world> pairs, the counterpart relation will not in general hold just between such pairs which share first members. This is because the counterpart relation is not an equivalence relation, which he apparently supposes that its TWI analog would be: in particular, the counterpart relation is neither transitive, nor symmetric, nor does it guarantee uniqueness of its relata in worlds.

**Criticism of Lewis:**

**Different Beings vs. Different Ways of Being**

Lewis bases his argument for possible worlds realism on the premises that things could have been different in many ways, and that this implies that there are many ways things could have been. These premises are undoubtedly true; but Lewis errs in assuming that they imply that none of the things in
other worlds that would have existed if the way things are had been different are identical with any of the things in the actual world.

The first step in seeing that there could be different ways of being a thing without there being distinct things for each of those ways, is to consider the difference between what a thing is and what it is for a thing to be a certain way.

As we have seen, Lewis defines a possible world as a way things could be. It is his emphasis on 'things' in this definition that leads him to conclude that the things in other worlds are distinct from actual things. But we must ask, what is a way things could be; and what is the way things are? I suggest that ways things are are classifications of things.

To say, for instance that a way this paragraph is is ill-formed, is to classify this paragraph, to say that it is in the class of ill-formed paragraphs. Alternately, one could say that a way a thing is is a property had by the thing.

Why should we accept this interpretation?

First, because

(P) There are many properties a thing could have had which are different than the properties it actually has.

is an intuitively plausible paraphrase of
There are many ways a thing could be which are different from the way it is. even though the former is, perhaps, less vague than the latter.

Second, because alternate paraphrases are either not intuitively plausible, or are equivalent to our suggestion. For instance, Lewis' argument against TWI suggests

There are many things which are the things that actual things would have been if they had been different.

as a paraphrase of W. This suggestion captures Lewis' view; he says,

we might say, speaking casually, that your counterparts are you in other worlds, that you and they are the same; but the sameness is no more literal than the sameness between you today and you tomorrow. It would have been better to say that your counterparts are men you would have been, had the world been otherwise.4

(T) is obviously not equivalent to (P); but it is not an intuitively plausible paraphrase either.

In the first place we can consider the plausibility of two paraphrases of:

There are many ways a person could be smiling which are different from the way he is smiling.

(W') is an instance of (W). Now consider a paraphrase of (W')

paralleling Lewis' paraphrase (T) of (W)

(T') There are many persons which are the persons that the person who is smiling would have been if he had been smiling differently.

Compare this to the parallel to (P):

(P') There are many kinds of smiles a person could have had which are different from the kind of smile he actually has.

Now it seems flatly absurd to say that (W') entails, or is equivalent to (T'): That a person smiles in one way does not entail that he could have been a different person than he is (unless the consequent is assumed to be necessarily true, which would beg the question), much less that there are other persons he might have been. (P') is clearly a more acceptable paraphrase of (W'), and hence (P) is preferable to (W).

In the second place, if (T) were to capture the meaning of (W) it would do so at the cost of equating things with ways things are. This is equivalent to the view that things just are the properties they have. If worlds are ways things are, and worlds are things, some things (at least) are just ways things are. While this view is not absurd on its face, it is not particularly plausible either -- for what are a thing's properties properties of if a thing just is a set of properties? In terms of Aristotle's distinction between form and matter, this view holds that there is
no matter, only form, no genuine substance kinds, only
adjectival kinds for concrete objects as well as for
abstract ones; our reply is similar to Aristotle's;
what is such a form a form of? But those defending the
view cannot claim that, like Aristotle's abstract
objects, they have intelligible matter, for they have
denied Aristotelian matter altogether.

Finally, this view seems to be inconsistent
with Lewis' position on the status of abstract entities.
For according to Lewis, such entities do not exist
in worlds. But if things are really only sets of
properties, then they would seem to be abstract objects
both in the sense that sets are and in the sense that
properties are. Returning to Lewis' talk of 'ways of
being,' we must note that ways of being are either
universals, if more than one object can share a way
of being, or they are individual essences, if a things
way of being is unique to that thing; but both univer-
sals and essences are abstract entities, and are,
even for Lewis, identical across worlds.\footnote{Counterpart Theory and Quantified Modal Logic, p. 120-23.}

The position I am taking in this issue is not
that there are no things which exist only at single
worlds, i.e., that there are no world-bound individuals,
but rather the view that there are some individuals that exist in many distinct worlds; and that at least ordinarily individuatible objects like tables, chairs, and persons are objects of that sort.

I will also argue that something like Lewis' account is right if one is concerned not with worlds constituted of individuals having different combinations of familiar properties, but rather with different kinds of properties, based on different conceptual frameworks. To put it baldly, Lewis is wrong about trans-world relations, but he may be right about trans-framework relations. Another way to put it is that Lewis is concerned with the counterparts of individuals (counterparts of this table), while I will be concerned with counterparts of sortals (counterparts of tables, different conceptions of what it is to be a table). We will return to an explanation of these mottos and the justifications of these views later in this chapter.

So far, our consideration of (R) has revealed problems in (D). As an argument for possible worlds realism (R) seems intuitively plausible, short of objecting that (R1) is false, i.e. asserting that things could not have been different. But this seems tantamount to an ultra-rigid metaphysical determinism,
claiming that it is metaphysically impossible for things to have been other than they are.

This is not the place for a detailed argument against rigid metaphysical determinism. But one cannot help but wonder whether such a view can be reasonably supported. Some have argued that it is implied by the doctrine of universal causation; but this begs the question by assuming that causation is rigidly deterministic, rather than probabilistic.\(^6\)

If quantum mechanics is considered as providing a complete description of subatomic phenomena, then causation on that level is not deterministic. And since it is conceivable that all causal interaction is mediated on the subatomic level, it is reasonable to believe that no observable case of causation is rigidly deterministic, even though the probabilities of many observable causal relations may approach it.

What evidence is there in support of (R1)? In the first place, the possibility of free action implies the possibility of things being other than they are. And it is much more intuitively plausible to suppose either that there are free actions than that

things could not have been other than they are, or that free action is compatible with rigid metaphysical determinism.

Second, it is at least logically possible that things be other than they are; it is logically possible, for instance, that the person in my shoes is not George. So at least things could be other than they are in a purely logical sense of could. Thus, it seems reasonable to believe that things could have been different than they are in many ways; agreeing to Lewis' stipulation that possible worlds are ways things can be leads to the result that it is reasonable to believe that possible worlds exist.

This does not imply however, that the things which could have had different properties, or which exist in other possible worlds, having different properties there than they do here, are not the same things that exist in the actual world, at least if we mean 'having different combinations of the properties we ordinarily conceive of' by 'having different properties.'

The principle of the indiscernability of identicals (I=, or "leibniz' law") has sometimes been used to argue against TWI. Roughly, if my pen is red in the actual world and green in another world, then they are discernable, and hence non-identical. But
on the present view, properties are relativized to our world, since worlds are ways of classifying things, of ordering things according to the properties (drawn from our system of classification, i.e., our conceptual framework) they have; so for a thing in a world to have a property is for it to be related to other things in that world by the ordering or classification of things as conceived of here. For my pen to be red in the actual world is for it to be classified with, and hence related to a class of objects in this world; and for it to be blue in another world is for it to be related to a class of objects in that world, the ones we would recognize or conceive of as blue if we were in that world. Nothing prevents us from saying that my pen is related to one set of objects here, but that it itself could be related to other sets of objects instead; if the way things are had been different, it would have been. Generally then, properties, or classifications of objects are relativized to the ways classifications can be made within a given theory, to the ways things can be; that is, to possible worlds. My pen is not red simpliciter: it is red in the actual world -- and blue in some other world -- but not discernable from itself in virtue of this fact. It exists in both worlds. Thus it is possible for things to be
different, i.e. to have different properties, than they actually do; so D2 is false and D is hence unsound. 

(D5) May be read in two ways: it is trivially true that if a thing had been a different thing than any thing in another world then the two are non-identical, no matter how similar they seem to be, i.e. whether or not they are counterparts. But it is false that if a thing in one world had other properties than it does in another, then it is really two distinct things — although it may be intimately related to things in those worlds which are distinct from each other, as we shall see.

Thus, Lewis' argument for the reality of possible worlds is not the basis of a sound argument against trans-world identity. But what of his objections to TWI? Quine's claim that it is not clear when unactualized possibles are identical with each other or with actual objects can be met in two ways. First, following Kripke, we could deny that identity criteria are necessary, because it is by stipulating (or assuming) identities that worlds are reached. We do not, says Kripke, look into worlds to find individuals and identify them by comparing their properties to properties of things in other worlds; rather, we pick out an individual and specify changes in his actual
properties to 'reach' a world. This requires the use of terms which do not change reference from world to world, if we are going to refer to the same individual in different worlds; such terms are rigid designators. Kripke holds that proper names can serve as rigid designators, provided that they achieve their reference through some occasion of use, such as a Baptism, in which the name is unambiguously used to designate the object, and that the name used on this occasion is adopted for future use as a referring expression. (This characterization of Kripke's account is rough, but elaboration of detail is unnecessary for our purposes).

Unfortunately, this account has its limits. We often use descriptions to refer to individuals, and descriptions, such as "A thing which is F, and G," or definite descriptions like "The F which is G" are based on properties and relations which can vary from world to world. Thus, even though "The F which is G" may individuate an object, and refer to it and only it in the actual world, there is no reason to think that it will refer to the same object in other worlds. And if a world contains two or more things which are F and G, then we are left with the problem of determining which, if either, of them is identical to our F-G thing.
What sort of view about possible worlds would be consistent with Lewis' premises that possible worlds are ways things could be, as I have interpreted that phrase, and the premise that there are ways things could be which are not actual, i.e., which are not the way things actually are? Any such view would, of course, be committed to saying that there are possible worlds which are not actual. It is plausible to refer to such views as views holding that possible worlds are real.

My interpretation of the phrase 'ways things could be' is analyzable in terms of orderings, or sets, of objects. But what is an object? As I have been using the term, objects are ontological primitives generally capable of being included in several orderings; yet there is room for two sorts of discussion of objects and their orderings.

First, we may speak of different sorts of object.

1) There are what might be called ordinary, observable macro-objects, like tables, chairs, people and statues.

2) There are, if the realist interpretation of contemporary science is correct, micro-objects of which macro-objects are constituted; these may be, in some cases, theoretically observable in the sense
that the theories incorporating provide explanations of how observations of the macro-level constitutes indirect observation, mediated by the instruments of science, of them; although on some accounts it is held that there can be no ordinary observation of them. Similarly, there are scientific objects which are collections of micro-objects and which are usually thought of as identical to macro-objects; but recall our discussion of Sellars on the correspondence between 'scientific' tables and their 'manifest' counterparts in chapter 3.

3) There are abstract objects, which differ from 1 & 2 in that they do not have spatio-temporal properties, sets and numbers, are examples of abstract entities. On some views, all abstract objects are either reducible to or analyzable as sets, but we will not consider the reduction here.

4) Finally, traditional dualists postulate the existence of mental objects, like minds and thoughts. Like abstract objects, mental objects are not said to have spacial properties, but they may have temporal properties. Mental objects are said to be intensional; some theorists regard sets and other abstracts as intensional as well. Again, mental objects are sometimes claimed to be reducible to or analyzable in terms of sets.
The radical nature of Lewis' version of modal realism is reflected in the fact that he postulates the existence of objects which do not actually exist (exist in the actual world) when he postulates world-bound individuals in other possible worlds. The thesis that such merely possible objects exist is objectionable, not only on grounds of parsimony. He claims the contrary, since, he says, he is not postulating a new sort of object; merely possible objects are, on his view, the same sort of object as actual objects. But this can hardly be so; in fact each world will contain only objects of a unique sort; objects which exist only in that world. If Lewis were to embrace this, change the kind of possibility he speaks of and assert that his different worlds were different ways of conceiving one 'world' (or 'universe of discourse' or 'object of conception'), then he would avoid this problem; but at the cost of having to admit that at most one world exists, i.e. that at most one conceptual framework (counting logically equivalent frameworks as one) could correspond to the way things are and could be. This suggests two kinds of possibility, one resting on a given set of concepts (and individuals falling under them), the other based on different ways of conceiving.
So, within Lewis' theory, objects in all non-actual worlds will be of a different sort than real objects, they will be of the possible but non-actual sort. Lewis would reply that it is true, in each world, that the objects existing there are actual in that world, just as actual objects in the actual world are actual in it. But being actual in a merely possible world is not more being actual than being green in that world is being green. What Lewis seems to fail to realize is that to make his account plausible he must deny that it is an account of how the things that we conceive of might have been different, using our way of conceiving of these things and assert that it is an account of how radically different sorts of things might be conceived of, using different concepts than our own. In not allowing different kinds of things Lewis is not radical enough; in claiming to account for possibilities involving our kinds of things, he has been too radical.

Identity, Counterparts and Composite Individuals

On a less radical possible worlds realism, an account of modality within our conceptual scheme allowing transworld identity and restricting the domain of objects to those actually existing, these
problems evaporate; identity across worlds need not be a problem if we use rigid designators and unique individuation within a world is possible. Further, this view specifically rules out the postulation of ontologically basic objects in other worlds, while allowing the construction of derived world-bound objects, consisting of object, world pairs. Further, a logic of counterparts can be adopted for such constructed objects, allowing one to say "the individual which is George in the actual world has twin counterparts in another world." Of course the modal operators for the added counterpart logic would have to be distinguished from the ordinary operators. Counterpart possibility, $c$, would depend, as for Lewis, on similarity of properties; and one's counterpart in a world, allowing 'There is a world where I am more similar to (the actual) Kit Fine than I am (in that world) to myself (in the actual world); i.e. a world in which I am Kit Fine or, more correctly, Kit Fine is my counterpart."

This approach requires no unreasonable primitives. Actual properties, i.e., classifications or orderings discovered in -- or imposed upon -- the actual world, can be used as a basis for property construction in possible worlds, i.e. the determination of which objects have which properties. Doing so
formally would require an analysis of similarity like Lewis', but such an analysis can be stated, as Lewis does, in terms of sets.

In addition to ordinary trans-world objects and world-slice objects (world-object pairs), we can consider composite objects consisting of sets of world-object pairs. This allows the construction of bizarre composites such as the object consisting of George Wood in the actual world, Glenn Ross in the closest world where he exists but George doesn't, John Pollock in the closest world where neither George nor Glenn exist, etc. But it also allows for a natural modal semantics for considering objects insofar as they have, or could have, certain properties. For instance, we can consider Glenn Ross insofar as he is a vegetarian. Following aristotelian usage, this would be considering Glenn qua vegetarian.

A more radical 'possible-worlds' theory, not based on a given set of concepts and individuals is also plausible, as suggested above; each world (or system of worlds of the former type) consists of the objects (actual and possible) conceived of by one conceptual framework. Here, there is no identity of either object or property from world to world; but there may be counterparts, based on the similarity of concepts in one framework to those in another. At most
one of these 'conceptual world-systems' can be real, corresponding to a perfect framework.

Identities and Objects: Contingent, Type and Token

We are now on the brink of returning from our digression to the issues of type and token identity. We will return by applying our discussion first to the issue of contingent identity and then, in stages, to various possible identity and counterpart relations between mental and physical items, finally placing this in the context of relations between explanatory theories in physics and psychology. I will speak of possible worlds in the narrower sense discussed above, i.e. as constructed within a given conceptual framework, except in those places where I explicitly deviate from this standard.

Gibbard Revisited

Our discussion provides a natural response to Gibbard's claim that Lumpl and Goliath are contingently identical. For Goliath qua lump, and Lumpl qua statue can be identical (in Kripke's sense, i.e. necessarily identical), as can Goliath qua actual and Lumpl qua actual even if Goliath (simpliciter) is not identical with Lumpl (simpliciter). This move would require that when one considered an object qua actual one
considered only its intrinsic extensional properties within a given framework. For if one included such properties as being desired by Jones, then it would be possible for them to differ. Notice that Lumpl is not identical with Lumpl qua statue; the latter does not exist at times in worlds where Lumpl fails to be a statue; for instances after the statue is squashed without destroying the lump.

Thus, contingent identity can be viewed as identities holding between world-slice individuals or individuals qua having certain properties. But can this form of contingent identity be used to rescue type-identity theories from Kripke's objection that it is possible for there to be a being with C-fiber stimulation without pain or vice-versa? Two approaches spring to mind.

One approach would identify mental state types qua actual with physical state types qua actual. But unless it can be argued that Kripke's imagined state of affairs is not actual, i.e. that there are no actual (say extraterrestrial) beings with C-fiber stimulation but no pain, or vice-versa, this approach still falls short.
Sortal Identity and Nomologicals

Another approach would resort to the sortal 'human,' or a sortal which applies to all and only animals of the types found on earth as it actually is. Thus it would be claimed that pain, qua occurring in terrestrial-type animals is C-fiber stimulation qua occurring in them, or more generally,

$$\text{For every mental-state-type-qua-occurring-in-terrestrial-type-animals, } A \text{ there is a corresponding physical state type, } B, \text{ such that (x)} (Ax \equiv Bx)$$

This formulation seems able to avoid Kripke's imagined counterexamples, insofar as they rely on imagining extraterrestrials or other creatures sufficiently unlike those actually examined. As such this thesis may be a reasonable basis for research. Of course, the thesis will become more plausible if suitable correspondences are discovered.

Notice that the thesis can be given a natural counterfactual formulation:

$$\text{For all } x \text{ such that } x \text{ is a terrestrial type animal, if } x \text{ were in mental-state type } A, x \text{ would be in physical state type } b, \text{ and vice versa.}$$

Such a formulation suggests that the so called type-identities are not identities or definitions at all, but rather law-like connections. This is clear when we see the types (or properties) in question are not
identical but are merely coextensional, like 'human' and 'naturally featherless bird;' they have the same extensions in worlds having sufficiently similar natural life-forms. And notice that this formulation is a universal statement explicitly supporting counterfactuals, thus being much like a law of science (except that it is so far not supported by scientific evidence).

Thus, this exploration of type identity theories has led, thus far, to an argument for a reductionistic double-aspect theory: the 'laws' state that

1) every item which is a mental state of type A is a physical item of type B,

and that

2) for every type of mental state there is a corresponding type of physical state.

To complete an argument for reductive physicalism one need only argue that physical types are ontologically basic. Note however that if 2) is false there will be irreducible mental types, but 1) could be reformulated as

1') every item of a mental type is an item of some physical type.

1' suggests a form of non-reductive physicalism, assuming again that physical objects can be shown to be ontologically basic.

Note however that 2) is ambiguous; for it does not specify whether the type in question is a physical
type. This suggests the possibility of a theory according to which mental items are physical items having characteristics not within the domain of physics, at least as that domain is ordinarily construed. The crudest version of such an account would make that characteristic one like 'being a thought of Paris.' On this account, saying that a physical state is a thought of Paris would not be identifying two entities, but would be attributing a non-physical property to a physical object. Such properties would be, in a sense, emergent properties of physical objects unless they were grounded, in a law-like way, in purely physical properties. But if these properties are instead thought of as belonging to different conceptual frameworks, the question becomes: how are different frameworks related? Can the relation be law-like? Can it be reductive, either in the traditional positivist sense or in some new sense?

A careful reader has no doubt noted that the alleged rescue of reductionism from the jaws of Kripke has still not established the possibility of the law-like connections required by 2). The most that was shown was that Kripke's objection can be overcome.

Thus, the crucial issue dividing reductive and non-reductive physicalism, is whether or not there are
or can be law-like connections between mental and physical states. This is a question of whether or not there are phychophysical laws. In the next chapter we will examine Donald Davidson's argument for the anomolousness of the mental, which concludes that there can be no such laws, connecting mental and physical events.

Theories and Their Entities

Throughout most of part II, I have been speaking of mental entities in the vernacular; but for the most part, what I have said about thoughts, ordinarily construed, carries over to the items postulated within a psychological theory, in the sense of 'theory' developed in Part I, according to which 'mental' items are postulated in order to provide psychological explanations. Nothing in that account precludes the items postulated by one theory from consisting of items postulated by some other theory. Transposing this theme into the one developed two paragraphs back, the non-physical characteristics of physical entities that makes them mental would naturally be the characteristic that a psychological theory attributes to its postulata; this amounts to the view that mental items are physical items playing the role required by
psychological theories of the items they postulate. What are these roles, and how are they characterized? This theme, and these questions, will take us to the conclusion of part II.

Systemic relations

Before concluding this chapter I will digress a moment on the recurrent theme that possible worlds are ways things could have been and that ways are orderings. Had Lewis gone a step further and said that his worlds were different ways of carving up reality, rather than focusing on things and the ways things could be, he would be at home in a rich tradition with Carnap, Duhem, Goodman, Sellars and Quine; for then his thesis about 'worlds' would be of a piece with the Duhem-Quine thesis that only whole theories with their rules of interpretation meet experience and that such wholes are incommensurable, with Goodman's notion (out of Carnap) that we make worlds by imposing systems of description on experience, and with Sellars' account of alternate conceptual systems carving up reality. This suggests that what Lewis does not properly acknowledge is that outside such systems there are no things. Now this, of course, can be no sooner said than it must be taken back; for it sounds like a version of Hegelian idealism, implying
that the world depends for its existence on mind or minds. But that does not follow, at least if one means by 'there are orderings' something like 'reality can be ordered, or carved, or represented,' and unless one denies that reality has a structure which our science is an attempt to capture. But in presupposing not just things, but the things of everyday experience, and not radically new concepts, or ways of ordering, but using our everyday concepts to describe new things in the same old ways, he has confused the way reality as a whole might be conceived with the ways things might have been given our way of conceiving it.

When we say "it is possible that ... "there are two, radically different things we might be doing. The more ordinary thing we do with this form of speech is to imagine the kinds of things with which we are familiar having the kind of properties with which we are familiar, but just different ones than we think or believe they have. For example, "it is possible that someone is selling computers in central park." Such possibility plays a central role in most counterfactuals: we intuitively evaluate them, as both Kripke and Lewis admit, by imagining what things would be like if the antecedent possibility were the case. This of course, requires that we think using our common
sense garden variety concepts. But when we philosophize, we sometimes use a mode of possibility incompatible with the se, and calling for new concepts which, at first, we understand only metaphorically. It is possible, in this sense, that the world consist of quarks. Here our (or our scientists') dissatisfaction with the old way of conceiving of the world is leading us (them) to develop new concepts; and to the degree that each of them becomes accepted we change the way we think about reality, and in doing so, come to have a new world as the object of our thoughts. Although we begin this by arranging the pieces of the old system, it results as much in a new system as rearranging the 'pieces' of football, soccer and treated the different game of basketball.

That the two sorts of possibility often overlap is to be expected; but their overlap does not mean that we should not attempt to distinguish clear cases of possibilities of changing one. The fact that conceptual change occurs slowly and that differences between the claim that they are nonetheless different ways of serving reality into objects, but rather explains our abilities with to confuse new concepts with their older homonyms and eventually to distinguish them and to relate new ones to their older counterparts. Rules expressing these relations are what
Sellars calls 'correspondence rules' and form the basis of his variety of reduction without identity.

Having spent most of this chapter on identity within a system, we turn back to our discussion of reductive laws relating the mental and physical components of a single conceptual system.
CHAPTER X
NOMOLOGICALS AND REDUCTION

Quick Recap: We have found that 1) it can be reasonable for sciences, including psychology, to postulate entities to explain events in their domain of observation; 2) doing so need not violate the doctrine that all things are ultimately physical, even if definitional reduction and type-identity theories fail, since 3) there are no good reasons for rejecting a sophisticated token-identity theory, i.e., one which says of each individual that it is a structure of physical stuff; however 4) this route seems to violate the intuition that everything must be explainable in purely physical terms unless there is a law-like connection between the predicates invoked in psychology and those of physics. We have also been developing on the themes that 5) there may be a counterpart relation between different conceptual frameworks according to which the concepts of one framework are related to those of another, even though they are neither identical concepts nor, strictly speaking, intertranslatable or interdefinable; and 6) items postulated by psychology are conceived of as satisfying functional roles.
Thus, in order to satisfy physicalist intuitions we must either show how (or at least that) there can be nomological connections between the functionally conceived theoretical entities of psychology and the theoretical entities of physics or find some other sense in which particular psychological phenomena can be given physical explanations. This chapter is intended to examine arguments that, if correct, cut off the former alternative and begin developing the latter. Chapter 11 will continue that development by reviewing the features of an acceptable token identity theory and showing that it is a natural part, though only a part, of an acceptable theory, and chapter 12 will return to the arguments for the realist interpretation of such theories in general, and especially of psychological theory given a realist interpretation of physical theory.

Davidson's Attack on Reduction via Psychophysical Law

In 'Mental Events' Donald Davidson argues that there are no strict psychophysical laws, and goes on to argue, on the basis of this principle and two others, for a form of token identity theory. Briefly stated, Davidson's argument against psychophysical laws is:
1) Psychological explanations essentially rely on mentalistic descriptions of persons and their behavior formulated in terms of propositional attitudes.

2) Physical explanations are causal explanations and rely exclusively on extensional descriptions of the events being described.

3) Thus the vocabularies of physics and psychology do not fit each other: they are not drawn from the same conceptual domain.

4) Even if there were true generalizations relating mental and physical events, they would not support counterfactuals.

5) Laws support counterfactuals.

6) Therefore, even if there were true generalizations relating mental and physical events, they would not be laws.

7) Therefore, there are no psychophysical laws.

The missing premise is that descriptions which do not fit each other because they are from different conceptual domains cannot be connected in counterfactual generalizations. Davidson's argument for this premise is based on his solution to Goodman's grue-bleen problem:

All emeralds are green' is lawlike in that its instances confirm it, but 'all emeralds are grue' is not for 'grue means' observed before time t and green, otherwise blue, and if our observations were all made before t and uniformly revealed green emeralds, this would not be a reason to expect other emeralds to be blue. Nelson Goodman has suggested that this shows that some predicates, 'grue' for example, are unsuited to laws (and thus a criterion of suitable predicates could lead to a criterion of the lawlike). But it seems to me the anomalous character of 'All emeralds are grue' 'is grue' are not suited to one another; grueness is not an inductive
property of emeralds. Grueness is however an inductive property of entities of other sorts, for instance of emerines (Something is an emerine if it is examined before I and is an emerald, and otherwise is a sapphire.) Not only is 'All emerines are grue' entailed by the conjunction of lawlike statements 'All emeralds are green' and 'All sapphires are blue' but there is no reason, as far as I can see, to reject the deliverance of intuition, that it is itself lawlike. Nomological statements bring together predicates that we know a priori are made of each other--know, that is independently of knowing whether the evidence supports a connection between them. 'Blue,' 'red,' and 'green' are made for emeralds, sapphires, and roses; 'grue,' 'bleen,' and 'gred' are made for sapphalds, emerines, and emeroses.

The direction in which the discussion seems headed is this: mental and physical predicates are not made for one another. In point of lawlikeness, psychophysical statements are more like 'All emeralds are grue' than like 'All emeralds are green.'

An obvious reply to this argument can be drawn from Sellars' discussions of the relation of atomic physics and chemistry discussed in Chapter 4 above. For before the development of an inclusive theory defining chemical concepts in terms of basic physical Ones (such as atom, electron, etc.), Chemical and Physical terms belonged to distinct conceptual domains based on distinct methodological and observational presuppositions. Yet chemistry has been shown reducible to physics by developing an inclusive theory. What reasons are given for denying that a theory including
physics and psychology is possible? Davidson seems to have anticipated this objection and provides two sorts of consideration opposed to such a possibility: the irreducibility of intensional idiom and Quine's thesis of the indeterminacy of translation.

The connection between irreducible intensionality and the failure of physical reductionism is made by analogy with an argument to the effect that irreducible intensionality provides an explanation for the systematic failure of reductive behaviorism:

It will sharpen our appreciation of the anomological character of mental-physical generalizations to consider a related matter, the failure of definitional behaviorism. Why are we willing (as I assume we are) to abandon the attempt to give explicit definitions of mental concepts in terms of behavioral ones? Not, surely, just because all actual tries are conspicuously inadequate. Rather it is because we are persuaded, as we are in the case of so many other forms of definitional reductionism (naturalism in ethics, instrumentalism and operationism in the sciences, the causal theory of meaning, phenomenalism, and so on—the catalogue of philosophy's defeats), that there is system in failures. Suppose we try to say, not using any mental concepts, what it is for a man to believe there is life on Mars. One line we could take is this: when a certain sound is produced in the man's presence ("Is there life on Mars?") he produces another ("Yes"). But of course this shows he believes there is life on Mars only if he understands English, his production of the sound was intentional, and was a response to the sounds as meaning something in English; and so on. For each discovered deficiency, we add a new proviso. Yet no matter how we patch and fit the nonmental conditions, we always find the need for an additional condition that is mental in character.
A striking feature of attempts at definitional reduction is how little seems to hinge on the question of synonymy. But the pattern of failure prompts a stronger conclusion; if we were to find an open sentence couched in behavioral terms and exactly co-extensive with some mental predicate, nothing could reasonably persuade us that we had found it. We know too much about thought and behavior to trust exact and universal statements linking them. Beliefs and desires issue in behavior only as modified and medicated by further beliefs and desires, attitudes and attendings, without limit. Clearly this holism of the mental realm is a clue both to the autonomy and to the anomalous character of the mental. 2

The connection between the indeterminacy of translation and the failure of reductionism is that laws are ideally stated in a homonomous vocabulary (one based on a single system of concepts) and that the Indeterminacy of translation ensures that psychophysical generalizations will be heteronymous:

The heteronomic character of general statements linking the mental and the physical traces back to this central role of translation in the description of all propositional attitudes, and to the indeterminacy of translation. There are not strict psychophysical laws because of the disparate commitments of the mental and physical schemes. 3

Of these two theses, the thesis of the indeterminacy of translation can be seen to depend on the thesis of irreducible intentionality, for the indeterminacy thesis requires that extensional or observational
evidence (drawn from observation of linguistic response to environmental stimulus) is not sufficient to determine intentional facts (such as what a particular thought—expressed by a speaker in words—is about). But if intentional facts were reducible to extensional ones, this requirement would not be satisfied.

Thus, Davidson's arguments against reductionism hinge on the irreducibility of intentionality and the claim that psychological explanations are essentially intentional, while physical ones are not, implying that there can be no laws connecting these realms. In the next section we will examine the possibility that a psychological theory which postulates inner episodes can both provide psychological explanations and be part of an inclusive theory whose ontological basis is that of theoretical physics.

While Davidson has taken a slightly different route to the view that psychology and physics differ with respect to their domain, and differ in a way that appears, as he would claim, to be unavoidable, at least given their existing conceptual structures. Of course, there is a sense in which the conceptual structure of physics — our current physics, as a static entity — is as fixed as Sellars' idealized 'manifest image,' and likewise a sense in which the
conceptual structure of psychology is fixed, Davidson is, of course right; these static conceptual systems are irreconcilable, or incomensurable, neither is analyzable in terms of the other, their types are irreducible. The best connection between them is that in each particular case each theory has its own description, both in observable and theoretical terms, of what is in some sense one thing. Just as the manifest table is, strictly speaking, not the table of scientific common sense because of the differing roles of 'table' in those radically different frameworks, the thought of psychology is distinct from its democratean counterpart. What Davidson overlooks (and it is only natural for him to do so) is the possibility of a single successor to these disparate frameworks rather than the discovery of some relationship between them. Now, I think it would be misleading to characterize the development or introduction of such a successor framework as a reduction of psychology to physics; for the new framework would presumably differ from both of its ancestors both in what it counted as observable and in its theoretical commitments. But it does share an important aspect with the reductionist program: the motive, in both cases, is the desire for a unified science.
We turn now to Sellars’ reconstruction of how a new, richer domain of observables can be developed from humble, impoverished, behavioristic beginnings, and how once people have learned to observe others as thinking, sensing and feeling, this leads naturally to the postulation of internal episodes characterized by their functional roles.

**Sellars: From Verbal Behaviorism to Scientific Mentalism**

Sellars’ reconstruction of the conceptual framework of thought proceeds in two steps. First, he develops a theory he calls "verbal behaviorism," according to which thinking is, in its primary mode, speaking, and is, derivatively explained in terms of dispositions to speak. The short-comings of this approach, most notably that it fails to explain these dispositions, leads him to the second step, in which internal states analogous—in some but not all respects—to overt speech are postulated. Sellars goes on to account for many of our ordinary intuitions (as captures by the implications of ordinary usage) in terms of the meanings, or functional roles of the items postulated by his theory, especially when those intuitions seem to conflict with the theory (e.g., the ordinary conviction that persons have privileged access to their own thoughts).
The first step toward verbal behaviorism is to recognize a form of speech Sellars call "thinking-out-loud." People think-out-loud when they say something with no intent to communicate or to have any effect on other persons or the world in, or by, their utterance. Sellars' classic example is that of a man standing on the curb waiting for a bus which, instead of stopping for him, whizzes right by. He says -- to no one and for no apparent reason -- "Damn, the bus went right by." In so speaking, the man thought-out-loud, and according to the verbal behaviorist, an eavesdropper would have heard his very thought.

Sellars recognized that adults, unlike children, do not often think-out-loud, for part of growing up is learning to keep quiet, to speak only when there is some reason for doing so, some purpose to be accomplished. And he is aware of the broad range of actions one may be performing in speaking (illocutionary acts) or by speaking (perlocutionary acts), and of the attempts by Austin, Searle, Grice and others to provide an analysis of meaning in terms of speech acts; but such attempts are limited in that they explain use ultimately in term of speaker intention, which involves the content, or meaning, of a speakers thought and hence must, as Gilbert Harman
has pointed out in "Three Levels of Meaning" presuppose a theory of meaning of content for thought. Sellars provides a single explanation for the content of thought and the meaning of those utterances which are thinkings-out-loud.

The verbal behaviorist recognized that people often think without thinking-out-loud, and explains this at first counterfactually; thus, 'S thinks that p' is equivalent to 'IF S were not inhibited, etc., he would be saying "p". Thus, thinking is having a disposition to speak (to think-out-loud), which is usually overridden by other factors.

Aside from the difficulty of filling out the 'other factors' in this formulation, it does not do justice to the speed with which thoughts can 'come and go' or change. To deal with this problem, Sellars introduces the notion of a short-term proximate propensity. Consider an electromagnet. When power is turned on, it is magnetic, i.e., disposed to attract iron filings, turn compass needles, etc., even if none of those items are in its vicinity. (Note in this regard that difficulty of filling out counterfactuals which fully fill-out the counterfactual definition of magnetism X is magnetic at t iff if iron filings were near X, . . ., then they would be attracted ... Consider also
all the ceteris paribus clauses which such an account would need.) Now, the magnetism of an electromagnet is a propensity or disposition which the electromagnet gains and loses as the power through its coils turns on and off. This kind of dispositional property, which can come and go in moments and may last a very brief period of time is a short-term proximate propensity. By allowing that thoughts are short-term proximate propensities to think-out-loud, the verbal behaviorist can answer the problem of the speed with which thoughts change.

Several other problems can be dealt with by restricting the range of the theory. Thus, when it is objected that not all thinking has propositional content, since, for instance artists and musicians are said to think even though they surely do not think 'in words,' the verbal behaviorist can retreat, restricting his model to thinking by adult humans which is propositional in form; or he can suggest extensions of this theory in the direction of non-verbal behaviorism. Thinking in melody would consist in short-term proximate propensities to hum, play piano, etc. At this level, the analogies become obviously strained. Ryle, behaviorist par excellence, developed these and similar themes in The Concept of Mind and other works spanning much of his career.
But there are stronger objections to verbal behaviorism if it is suggested as a scientific account of thinking, for science is not satisfied with explanation in terms of dispositions or 'powers' which forge a mysterious 'necessary connection' between events. For one thing, explanations of events in terms of dispositions for them to occur are circular if the dispositions are explained in terms of the events. For instance if one were to explain why a sugar cube dissolved in water by saying that sugar is water-soluble, and then explain water-solubility by saying that it is the disposition a thing has to dissolve in water, ones explanations would be going in small, if not vacuous, loops.

The moral here, as in earlier sections, is that scientific explanation requires that causal connections be uncovered through the use of models, in which the dispositional properties of observed objects are explained in terms of the occurrent properties of the theoretical objects of which observables are constituted. Explaining the dissolution of a sugar cube placed in a cup of water by saying "it dissolved because it was soluble" is less than adequate: what is wanted is an explanation of the sugar's solubility;
a model supplied by physical chemistry in terms of the molecular structures of sugar and water, and other occurrent properties of their constituents).

Similarly, explaining why Jones' hand went up by saying "Jones willed that his hand rise (i.e., Jones had a short-term propensity to say "it shall be that my hand go up" accompanied by a propensity to raise his hand) does not of itself explain the causal connection between Jones' propensity to say "It shall..." (is it even an actual, occurrent thing?) and this bodily motion, so long as the willing is only 'explained' by defining it as a propensity. For while there is a logical connection between a propensity for hand raising and hand raising, as between solubility and dissolving, what is wanted is at least a nomological connection between Jones' antecedent state and his consequent state, as between the antecedent state of the sugar, and its dissolved state. Now, the verbal behaviorist can provide a nomological connection by providing evidence that when Jones thinks "I shall now raise my hand," his hand goes up, other things being equal; and he can explain this in terms of Jones' having learned the association between saying "I shall do x" and doing x, and hence having a propensity or to do so. Thus there is a causal connection, in a
Humean or Instrumentalist sense, between the thought and the motion; but what is wanted is a causal explanation in the realist sense, like the realist explanation of solubility in terms of the theoretical entities of which it consists.

At this juncture, Sellars moves from Verbal Behaviorism to the postulation of mental states. It should be pointed out that the initial move here, postulating thoughts as things and not merely as idpositional properties of persons, is perfectly consistent with Cartesian dualism. The problem for the Cartesian, of course, is explaining the connection between the immaterial, non-spatial mental realm and the causal order of physics, neurophysiology and bodily movement. Why did Descartes so steadfastly refuse to identify mental items with bodily ones? In the first place, doing so would threaten the possibility of reconciling science and religion, given the assumption that materialism is inconsistent with both the possibility of freedom and the existence of an immortal soul. But while Descartes sought such a reconciliation, he also saw what he thought was a sufficient reason for thinking that no physical or neurophysiological item or process could be a thought. As Sellars' points out:
he would have rejected this identification on the ground that we had a 'clear and distinct,' well-defined idea of what conceptual thinking is before we even suspected that the brain had anything to do with thinking. Roughly; we know what thinking is without conceiving of it as complex neurophysiological process, therefore, it cannot be a complex physiological process.\[4\]

This is the argument from the introspectable qualities, the intrinsic content, of thoughts. Surprisingly, Sellars takes it seriously enough to devote a great deal of attention to answering it. It is, of course, a version of the argument from the indiscernibility of identicals. Now it should be clear that at least one part of this argument is an obvious non-sequiter: It doesn't follow from the fact that we have not noticed that \(x\) is made of \(y\) that \(x\) is therefore not made of \(y\).

However, on Descartes' view, we have an adequate idea of what a thought is, i.e. an idea which specifies the intrinsic qualities of that thought, since we have direct access to our thoughts. If we do have an adequate idea of something, then we have noticed its essential nature, and hence its composition; so if we have an adequate idea of \(x\) and have not noticed it to have (the intrinsic) quality of being made of \(y\), it is not. Thus there are some introspectable

---

properties by which Descartes thought he could discern thoughts from physiological item or processes; in particular, we have direct access to our thoughts, but not to physiological processes.

Sellars' response to this is twofold. He first attacks that variety of direct access which is supposed to give foundational knowledge. In particular he attacks Descartes' assumption that since we have direct access to our thought, we have an adequate idea, one that specifies all the intrinsic qualities, of thoughts. This attack on the myth of the given is an important part of Sellars' system and will be taken up when we turn to consider some of the epistemological implications of, and constraints on, functional theories. Second, Sellars argues that properly construed (without its givenness) there is a form of direct access to our mental states which provides a direct report, not based on inference, about our mental states. As we shall see, these reports, though direct, are not "given," i.e., intrinsically justified.

The important point for now is that this version of direct access is consistent with a theory according to which thoughts consist of physiological items, contra Descartes. This is because just as we can see
things without seeing all of them or their constituents, as we report our thoughts without knowing their real ontological status. The remainder of the story of direct access is more appropriately placed in the section on Chisholm's recent, but largely Cartesian, epistemology.
CHAPTER XI
TOKEN IDENTITY AND FUNCTIONAL CHARACTERIZATION

So far we have considered a broad range of arguments concerning the requirements of scientific explanation and the possibility of a reduction of psychological explanation to physical explanation. Two points need to be made now

1. Nothing in the present account of explanation requires that psychological explanations be reducible to physical ones. All that is required is that psychological theory provides models that reveal the connection between the event or action to be explained and antecedent events and states of affairs.

2. None of the logical or metaphysical arguments about the nature of mental items has shown that they cannot be "made of" purely physical stuff; what has been shown is that they cannot be strongly identical in the sense of Type-identity. But in this sense it is also false to say that a particular linguistic item, e.g. a token of the word 'red,' is identical with the stuff of which it is made (ink, paper) whether or not structural characteristics are considered.
This is consistent with the view Fodor calls 'Token Physicalism:'

Token physicalism is simply the claim that all the events that the sciences talk about are physical events. There are three things to notice about token physicalism.

First, it is weaker than what is usually called 'materialism.' Materialism claims both that token physicalism is true and that every event falls under the laws of some science or other. One could therefore be a token physicalist without being a materialist, though I don't see why anyone would bother.

Second, token physicalism is weaker than what might be called 'type physicalism,' the doctrine, roughly, that every property mentioned in the laws of any science is a physical property.

It has been strongly suggested that mental items are functionally described states of organisms; the functional descriptions may have dispositional analyses, but the requirements of explanation dictate that the functions (and associated dispositions) be explained in terms of models which, if correct, would picture the structure of the organism in a way which reveals how particular organisms function as they do.

Here, science departs somewhat from metaphysics; for all that psychology requires is an account of mental function in order to explain behavior; while neurophysiology (and ultimately physics) requires an account of the physiological reactions including bodily motion (not described as action) and physical response
to stimulation. Insofar as the motions are actions, the physical responses (or those which satisfy the functional descriptions of psychology) are mental items.

Thus we have a coherent combination of physical, psychological and philosophical doctrines:

1. Bodily events, described in purely physical terms, are explained by causal models.

2. Bodily events, described in intentional terms, are explained by mental items, which are defined by their functional roles.

3. Philosophically, there is no reason to doubt, and every reason to think, that the mental items and events are physical items described in terms of their functional roles, even though those roles do not seem definable in terms of (i.e. reducible to) physics.

This combination of views can be perspicuously called Functionalistic Token Physicalism.

**Objections to Token Identity**

It may be thought that the token identity theory violates scientific intuitions about identity. After all, it might be said, if there is a scientific reason for thinking \( \phi = b \), then it must be based on laws, and support reduction. In this section I will consider several cases of identity wherein no reduction is thought appropriate and will argue that these cases illustrate important features of non-reductive identity theory, even though reduction is a reasonable scientific goal.
First let us consider a standard example of a 'scientific' identity claim = "lightning is electrical discharge." This claim is ordinarily taken to mean that if anything is lightning, then it is an electrical discharge. Assuming this claim is the product of a law relating observable phenomena to theoretical processes, are we to assume that it is necessarily true? Can we not imagine a logically consistent state of affairs which is observationally the same as our physical universe but in which there are no electrons -- a universe with a physical makeup quite unlike our own? Does this possibility show that lightning isn't identical with electrical discharge?

In addition to the rejoinder to this view discussed in chapter 9, several authors have provided what I think are convincing reasons for denying that this case is a counterexample to the necessary identity of lightning and electrical discharge. These authors use the case to argue that the discovery that lightning was electrical discharge may have changed the concept we have of lightning, by changing our conceptual framework, but that it did not change the natural kind, lightning, the paradigms of which were in fact, electrical discharges all along, whether people conceived of it or not.
One could, for instance, adopt Putnam's account of natural kinds and say that even if the lightning-like occurrence in the imagined state of affairs did fit the observational paradigm of lightning, it would not be lightning if our theory about lightning were correct and the imagined 'lightning' did not satisfy our theory, i.e. was not an electrical discharge; it would be 'fools-lightning' or 'manifest lightning.' Or, we could interpret the identity claim as a de-re universal quantification, \((x) [(Lx \iff Ex) \land \text{following Kripke's account of rigid designation, interpret it as "any object (in the actual domain) is necessarily such that no matter what changes, it is lightning if and only if it is an electrical discharge," or as an equivalent identity statement, either}
\[
(x) (Lx \rightarrow \Box (\exists y) (Ey \land x = y))
\]

or
\[
(x) (Lx \rightarrow (\exists y) (Ey \land \Box x = y))
\]

which are equivalent given that \(\{(x)(y) x-y \rightarrow \Box (x=y)\}\). Assuming that our theory is true of the actual world, and that Kripke is right about there being no merely possible objects and about an object's constitution being essential to it, these statements are true. But the same could be said of thoughts and brain states: if an identity theory of the human mind were true then
the imagined states of affairs in which people 'thoght'
but had no brain states would be flawed in the same
way the fools-lightning case was. Here we would
have fools-humans.

This Putnamesque response seems to have the
unfortunate result of requiring that a theory of
thinking be species-specific and implies either that no
non-humans could think or that we would need a different
type of thinking for each species. This point
requires two remarks.

1) When it comes to identifying the particular
theoretical items an 'observed' phenomenon consists
of, it is only right that theories be specialized
to the kind of phenomenon being observed, whether it
be lightning, fools lightning, or human or alien thought.
The lightning-electrical discharge case is, in a sense
ill-suited since the lightning is actually a complex
phenomenon involving discharge through molecules of
oxygen, nitrogen, argon, etc. and the ensuing photon-
shower; lightning on a planet with a methane-amonis
atmosphere would be as different a specific phenomenon
as would the thoughts of inhuman aliens.

2) On the other hand, the main point of identity
theories is not that nothing without terran neurons
thinks or senses but that thinking and sensing are
physical events satisfying appropriate functional and causal relations: so again in comparing lightning and electrical discharge to thinking and (terran, human) neurophysiological states, we are comparing apples and oranges: the former identification is more general than the latter in that it is not restricted to terran lightning, and more specific in that it specifies which atmospheric events lightning is identical with, while the latter only says that thinking is some neurophysiological state or other. The moral of these two points is that if we are going to compare the identity of lightning with some physical event to the identity of thinking with some physical event we should be equally specific or general in both cases to preserve even the vestiges of analogy. We can compare the identities of (a) things satisfying the observable lightning paradigm = things satisfying functional descriptions like 'atmospheric physical event causing sudden retinal stimulation' to (b) Thinking = physical event exhibiting functions $n_1 \ldots n_j$, or we can compare the identities between (c) Terran atmospheric lightning = electrical discharge through an Oxygen-Nitrogen-Argon/etc. atmosphere to (d) human thought = neural state of affairs connecting neural states $\psi_1 \ldots \psi_n$: but it is grossly inappropriate
to compare (a) to (d) or (b) to (c), especially since we have little or no idea of how the specific theory of human thought required of d would go.

This is, however, a digression from the main point of the first analogy, which is that a defense of the necessary identity of lightning and electrical discharge required heavy theoretical presupposition about the nature of identity, reference, natural kinds, and theoretical entities, and that the Kripkean defense required rigid designation over actual objects: that truth of general claim rested on truths about Individual discharges and their (presumed) actual constituency. This leads back to a point raised in chapter 9: that identity is primarily a phenomenon involving individuals and that 'type-identity' is a matter of finding ways of relating the terms in one theory to those in another.

Turning now to a second analogy, consider the relation between a smile and a face. It seems obvious that smiles exist and that smiles are identical to certain facial states. "Ah", I hear a familiar interlocutor asking, "but do smiles really exist?" The reality question is 'are smiles ontologically basic entities'. The hidden assumption is that only basic entities really exist, in the ontological sense of 'basic.' This assumption can be seen to lead to the
consequences that **faces** do not really exist, and so on; but as before, I see no reason to admit only bricks and not the walls they make up unless the question 'what is there?' is interpreted as 'what basic stuff is there?' in the ontological sense, i.e., 'what is everything made of?'

Another question is "but are smiles really states (arrangements, etc.) of faces." Consider the possibility of smile-dualism, according to which smiles are independent immaterial substances, i.e. substances existing in their own right, like the smile of Lewis Carroll's Cheshire cat, which could exist without the cat). No one would ordinarily deny that smiles are arrangements of faces, but hard-core reductionists and smile-dualists would deny the identity of smiles with facial states (and hence the reality of smiles for the former, and the rejection of smile-dualism for the latter) unless there were a reduction of 'smile' to 'such-an-such a facial state.' But how is this 'such-an-such' to be filled in? The smile-dualist would argue that it cannot be (noncircularly) filled-in since, after all there are presumably infinite many human facial configurations which would count as smiles and no guarantee that alien smIlers don't exist, etc., while the reductionist either lays his
trust in the development of science, or says that it is a reasonable research strategy and hence somehow incumbent on good scientists to provisionally accept the claim that the ellipsis will be filled in. I find neither position satisfying: the moral I draw from this is that where there is no firm reason for accepting dialectically opposed ontological theses, skepticism is better, even for the sake of science, than either dogma. It would be nice if either a reduction (or its impossibility) were proved, but nothing, not even ontological parsimony, is saved by assuming a reduction in advance. For even without the reduction, or even if its impossibility were proved, there is no reason to accept smile-dualism other than the thesis, shared by reductionists and dualists, that without a reduction a smile must be an entity distinct from the smiling face, and existing apart from it.

As a third example, consider the relationships between chess-moves and the motion of material objects. In particular, is there any purely physical description $D$ such that the universal statement

$$(x) \ (x \text{ is a chess piece if and only if } x \text{ is } D)$$

is true? Or the more complex version in terms of moves and events.
(x) (X is a chess move if and only if x is D).

This example illustrates two points: first, since almost any physical object can be used as a chess piece, and since even normal chess pieces could be made in an infinite number of ways, it is not surprising that there has not yet been a reduction. But surely chess pieces -- or, to stave off objections, items which can function as chess pieces -- are physical objects. Nonetheless, it is neither the case that anyone believes chess pieces to be non-physical nor the case that anyone expects a general reduction.

Second, in the case of chess pieces the reason for a failure of reduction is relatively clear: chess is a rule-governed, rule-constituted activity, and to say that x is a chess piece is to say that x can play one of the functional roles defined by the rules of chess. Rules are, of course, norms; the failure of reduction in chess is based on the familiar irreducibility of normative to descriptive discourse, with a vengence: for the physicalist reduction, the description can contain only purely physical predicates.

The moral of this example for the philosophy of mind is that if meaning is a matter of rule-governed function and occurrent thoughts are occurrences of sentences occur is neurophysiological, we should no
more expect a reduction of 'thought' to 'Neurophysiological process P' than of 'chese move' to 'physical-type P.' Note that in both cases individual instances of the functionally described items are describable as particular physical items. In the next chapter we will continue our investigation of the relation between objects and their functions.
CHAPTER XII
ASPECTS, STANCES AND SUPERVENIENCE

This discussion points in the direction of what has been called the 'double aspect' theory of mind: that a thought is an item which can be described both a) within the system of intentional language invoked in psychology and b) within the purely extensional language of physics. In fact, the above considerations seem to point to a 'many aspect' theory crossing not just the mind/body boundary but also the observable/theoretical and the syntax/semantics (or form/content) boundaries, as illustrated by the relations between lightning and electrons on the one hand and linguistic expressions and their functional classifications on the other.

Daniel Dennett has recently invoked considerations such as these in arriving at somewhat different conclusions with respect to the nature of mental items and their relation to physical items.

Dennett's Stances

Dennett's view is similar to Davidson's in that both regard psychological explanations as
explanations of a radically different sort than those of physics. According to Dennett, explanations of behavior (as opposed to mere motion) in terms of beliefs, hopes, desires, fears, intentions and other 'inner' psychological states or propositional attitudes, are explanations of an object construed as (qua) an intentional system, i.e. a system capable of intentional states. ¹ Dennett seems, at times, to conflate intentionality (aboutness) with intensionality (non-extensionality, i.e. the feature of a linguistic context in which substitution of coextensional terms fails to preserve truth value), but by and large this conflation in and of itself is not a major flaw in his account.

Dennett's first major point about intentional systems is that "a particular thing is an intentional system only in relation to the strategies of someone who is trying to explain and predict its behavior"² in terms of its beliefs, desires, goals, etc.³ Such a person has taken what Dennett calls the 'intentional stance,' which he contrasts to the 'design stance,'

². Ibid, p. 3.
³. Ibid., p. 6.
which he contrasts to the 'design stance,' which is taken by a person trying to explain the function of an item in normal circumstances, and the 'physical stance,' taken by someone who desires to explain features of an item according to laws of nature. The truth of this tripartite distinction in and of itself would entail that intentional explanations are not reducible to either design or physical explanations unless the intent of the explainer is to produce a single explanation which explains the behavior, function, and lawlikeness at once, and unless behavior is functionally describable and/or lawlike. Dennett's view of the distinction between intentional and extensional seems to imply that this cannot be the case.

The main point of Dennett's distinctions and initial claim is that something is an intentional system only from a certain point of view:

Lingering doubts about whether the chess-playing computer really has beliefs and desires are misplaced; for the definition of intentional systems I have given does not say that intentional systems really have beliefs and desires, but that one can explain and predict their behavior by ascribing beliefs and desires to them ... one can always refuse to adopt the intentional stance toward the computer.

... All that has been claimed is that on occasion, a purely physical system can be so complex and yet so organized, that we find it convenient, explanatory, pragmatically necessary for prediction, to treat it as if it had beliefs and desires and was rational.4

4. Ibid. p. 7-8.
So when, according to Dennett, is an item an intentional system? When (and only when?) we adopt the intentional stance toward it. Thus Dennett abdicates some of the traditional issues of the philosophy of mind. If the Hottentot tribesman feels it necessary to explain the growth of fruit in terms of the desires of plants to feed animals, then, from that point of view, the plants have desires. Dennett does not address the question "when is it **correct** to adopt the intentional stance toward an object x, that is, to explain its behavior by attributing intentional states to it."

Nor does he ask "what is the relation between the states attributed to x in an intensional explanation and those attributed to it by explanations made from within the design or physical stances."

In choosing to approach a subject from one of Dennett's three stances, a person is, in effect, choosing to use a particular system of classification and modes of reasoning appropriate to it. The idea that there are distinct 'stances' or systems of classification' or 'modes of thought' or 'languages' or 'conceptual frameworks' is, of course, not new. In this century, Carnap, Quine, Sellars and Strawson have all invoked this idea, and have made much of it. Davidson's argument against reductionism invokes it.
Carnap's "Empiricism, Semantics and Ontology" bears mention here since both Carnap and Dennett are concerned to deny that such traditional ontological issues as "Do x's really exist" are really any more than practical issues, roughly "does the language (system of concepts, framework, stance) which refers to x's provide useful ways of describing, predicting, and explaining the world." (Dennett explicitly makes this move where x=thoughts, beliefs, et al; Carnap is more clearly concerned with the general issues, considering cases where the language of x is one of ordinary physical objects, numbers, theoretical entities, meanings, etc.)

Thus it should be noted that Dennett's strategy with regard to the "reality" of intentional states can be repeated with regard to the other stances. If something is an intentional state or object only from the viewpoint of the intentional stance, parallel reasoning would seem to imply that an item is physical only from the physical stance. Dennett gives no reason for thinking that any stance is privileged, although I suspect that he takes the physical stance to be the measure of reality [as opposed to Richard Rorty, who explicitly holds (in Philosophy and the Mirror of Nature) that no stance or conceptual framework is privileged or has any claim to correctly represent objective reality.]
While this parallel reasoning may seem to constitute a reductio of Dennett, I am inclined not to regard it as an argument against Dennett's relativism, which when generalized becomes

An item is $\Delta$ only when viewed from the $\Delta$ stance

or

An Item $x$ is $\sigma$ only if $\sigma$ is an attribute which is part of the system of concepts and attributes constituting the $\Delta$-stance, and $x$ is viewed from that stance.

Rather I am inclined to agree with this generalized relativism and ask

1) When is it correct to adopt a particular stance toward something, and
2) When is it correct to make an attribution given that one has adopted a stance, and
3) When is a stance correct?
4) Can stances with seemingly different purposes be combined into a broader stance?

While Dennett does not address the first question he does say that it is permissible to adopt the intentional stance when useful or pragmatically necessary, and he seems to regard this as the end of the issue. If the question were "when is $S$ justified in believing it to be correct to adopt a particular stance $P$", then the answer would be "when $S$ determines that adopting $P$ is necessary to explain the current issue." Presumably 2) is a question which must be
answered from within the stance; in fact, the answers to 2) can be regarded as, in effect, defining that stance.

The third question is the question of metaphysical reality, and is hence the question Dennett steadfastly refuses to acknowledge. Sellars has anticipated this question and answers that a stance is ideally correct when its attributions describe reality to the greatest degree of adequacy possible. Two caveats are necessary here.

First, it must be obvious that it is unlikely, if not impossible, that anyone should ever know that any theory, and hence any stance, is ideally adequate. Thus the concept of the (ideal) correctness of a stance is unverifiable. There may be an infinite series of ever-better theories, or stances, with no end. My guess is that this feature has led Rorty, and perhaps Dennett to regard the question as meaningless, or empirically vacuous, and hence to disregard it.

Second, even if it is impossible in principle to know when or whether we have reached an ideally correct stance, we can nonetheless understand the ideal, just as we understand the mathematical concepts of an infinite converging series, and of the limit of such a series. It is only in terms of such an ideal that
we can compare competing viewpoints. Note that while the fact that some viewpoints have been found to be better than others does not entail that there is a best, it does entail that we have some concept — however indeterminate it may be — of what the best would be. This is, I think, the key to an enlightened understanding of why the notion of a perfect science, what Sellars calls the Piercian Conceptual System, is a regulative ideal.

The final question asks, in effect, whether a viewpoint that can incorporate the purposes of antecedently distinct viewpoints is possible; it suggests again the possibility of a unified science and a mode of reduction distinct from previous attempts to define the primitives of one theory in terms of another. I do not wish to quibble over whether to call the combination of conceptual frameworks into a broader framework a 'reduction' of either to the other, or a 'reduction' at all; but it shares with most recent reductionisms the goal of producing a unified science even while admitting that reduction in the more traditional sense is unlikely or impossible. One thing should be clear; that Dennett, or those who, like him, say that the choice of stances is to be made on pragmatic or 'utilitarian' grounds cannot object to someone
who takes such a stance on the grounds that a combined stance is of greater utility than individual, fragmented stances. Thus if the incorporation of a new range of predicates, say for instance functional sortals useful in psychological explanation, into a new scientific framework, a successor to contemporary physics, were to succeed in explaining the physical basis of human behavior by showing how causal processes of physical items can satisfy these functional sortals. Dennett, at least, could not reject it on the basis of reasons he uses for rejecting traditional reductionism.

Finally, even if the reasons favoring scientific realism seem to fall short, how could someone like Dennett object to the view that scientific realism is a meta-stance according to which, other stances, if adopted, are to be interpreted as securing genuine reference for their singular terms, and as making genuine claims to truth, rather than as merely convenient instruments? Why not take them at face value?

The work is not, of course, done. And perhaps there are no convincing reasons for thinking that it is, in fact, doable; but so long as the goal appears worthwhile and objections are not overwhelming, there is reason enough for the attempt. Not reasons that would sway all to act (I know of no such reasons) but
for those engaged in the attempt to understand the universe, and man in it, the promise of an answer where none stood before, and even were previous promises have failed, may suffice.

Conclusions

What? A conclusion before the third and final part? Well, yes, because what follows applies idea we've developed to other areas, serving a twofold purpose: to show that two traditional philosophical puzzles may be based, at least in part, on principles that our theory of the mental does not require, and to provide additional support for the theory by showing how it can be part of a systematic philosophy that offers solutions to these problems.

We have seen that to interpret a theory so as to provide an explanation is to interpret it as a redescription of the phenomena to be explained; and that this seems to imply that the best explanations (in principle) describe the world as it really is. But nothing in this implies, and many considerations lead us to reject the view that all sciences can use a single vocabulary.

Regardless of whether we must do so, the rejection of the view that all of science can be
reduced to a single, all-embracing theory does not entail that there is more than one domain of basic individuals; for different ways of classifying those (physical) entities and their structures, descriptions of them in terms drawn from other conceptual schemes, can provide new kinds of explanation, perspectives on the nature of the basic entities, perspectives based on features -- notably functional features -- which only structures of entities can possess. While speaking of theories as mere calculational devices divorced from an account of how the calculated connections can be causally mediated can avoid committing oneself to a plethora of unneded entities, nothing is gained and much is lost by ignoring the possibility that functional relations are mediated by -- even if not reducible to -- structures of physical entities.

Thus it can be reasonable for psychology to postulate non-basic entities having functional properties to explain events in its special domain; doing so need not violate the doctrine that all things are ultimately physical, even if definitional reduction and type-identity theories fail, since item.3 postulated by psychology are conceived of as satisfying functional roles.
Psychology, like many of the special sciences, can thus have its own domain of observables; behavior which cannot be defined in purely physical terms because, at bottom, it involves functional concepts which can be neither dismissed nor defined in non-functional terms; and while finer-grained sciences like neurophysiology cannot provide general explanations of these functions, they can explain the causal processes that mediate them in particular cases.

Of course, this route seems to violate the intuition that everything must be explainable in purely physical terms unless there is a law-like connection between the predicates invoked in psychology and those of physics. In attempting to satisfy these physicalist intuitions we have seen that, even if there are no nomological connections between the functionally conceived theoretical entities of psychology and the theoretical entities of physics, there may be some other sense in which particular psychological phenomena can be given physical explanations. The former possibility seems to be ruled out by the anomalous nature of psychological function, a feature stemming from the separation of the physical and psychological domains.
Nevertheless, there may be counterpart relations between different conceptual frameworks according to which the concepts of one framework are related to those of another, even though they are neither inter-translatable or interdefinable. Such relations may lead to the adoption of new frameworks incorporating successors to the concepts of current physical theory and our suggested psychological theory. The successor framework would be based on a broad domain which would be a successor to the previously segregated ones. This offers an approach to the goal of a unified science that allows us to bypass the excesses of logical positivism.

Even without achieving this goal, we have arrived at an account of psychological explanation in terms of functionally characterized physical processes. This is not a new theory, but I never promised a new theory; I hope to have slightly refined it, taken the edge off of recent criticisms, and to have shown it consistent with the requirements of explanation and what had seemed a menagerie unconnected considerations.
PART III

In the next two chapters I consider applications of functionalist philosophy of mind to issues in personal identity and epistemology.

In chapter 13, I argue that functionalism is consistent with a naturalistic approach to epistemology and that neither functionalism nor the desire for a scientific account of knowledge entails that persons have unacquired concepts, and that if functionalism does not reject those aspects of behaviorism it does not supplant, it can provide an adequate explanation of concept acquisition.

In chapter 14 I review recent literature on personal identity and argue that while certain traditional misconceptions about the nature of mind lead to paradoxes, the functionalist account provides a natural solution to problems involving apparent conflicts between criteria of survival and of identity.
Recent versions of the epistemological dispute over whether or not empirical knowledge has a foundation center around the question of whether or not there are either directly evident propositions or immediately justified propositions: contingent propositions which are not justified on the basis of other justified propositions but which are either justified or evident, in the sense that if they were also believed and true, they would be known. They are immediate in the sense that their epistemic status is not mediated by (derived from) the epistemic status of other propositions.

Of course, talk of justified propositions is talk of propositions which are justified for a person $S$ at a time $T$; I will generally abstract from these parameters. Similarly, I am concerned here only with what Hume called "matters of fact" and not with "truths of reason," although I think that my conclusions have interesting corrolaries for such knowledge.

Immediate justification must be distinguished from three of its possible relatives. First, it is sometimes said that there are self-justified propositions.
The very prospect of a proposition's somehow (how?) justifying itself seems so obviously circular that it is seldom given a second look; I will give it a more thorough examination below. Second, propositions are sometimes said to be intrinsically justified; these propositions are such that having them, or believing them, is enough, in itself to justify them. Descartes' "I think" is often thought to be a perspicuous example of an intrinsically justified proposition. Intrinsically justified propositions are immediately justified, but an immediate proposition may fail to be intrinsically justified by being directly evident. There is an air of paradox in putting it this way, but the reason for doing so is simple: directly evident propositions are justified in the sense of having high epistemic status, without being justified in the sense of having a justification. Some authors recognize this fact and expressly admit that the foundation of knowledge is not justified but has a special epistemic status that allows it to be known without justification. Finally, non-inferential justification is a species of immediate justification; a proposition is non-inferentially justified if it is justified but its justification is not inferred from that of another proposition is non-inferential for, as Robert Cummins pointed out to me, it is hard to see what epistemic mediation could be if
not inference, since other sorts of mediation would not be truth-transferring. But this difficulty evaporates when one sees the logical possibility of deriving justification (though not necessarily truth) from true (but not necessarily justified) propositions.

In Theory of Knowledge (TK hereafter) Roderick Chisholm presents an account that utilizes a form of immediate justification which seems, at first, free of the flaws that troubled earlier accounts, especially sense-datum theories. Wilfrid Sellars has long attacked such views on the ground that they were committed to a non-empirical dogma, the "myth of the given." My task is to demonstrate that Chisholm's view is similarly infected and that a version of the alternative offered by Sellars should be attractive to those who seek to escape the myth. Along the way I shall have something to say about a tendency toward innatism in contemporary philosophy of mind.

**Chisholm's Account**

Chisholm believes that one must accept the view that empirical knowledge of the 'outside' world is founded on noninferentially justified beliefs because other alternatives to the regress of justification are absurd. In brief, the regress of justification argument says that a true belief cannot
be known unless it is justified; but if all justification
is a matter of inference from propositions which are
already justified then either a) justification is
circular, b) chains of justification are infinite, or
c) there is no knowledge. But b) and c) are absurd,
and a) must be false. Hence it is false that all
justification is inferential; thus some justification
is nonferential.¹

I am focusing on the account given by Chisholm
in Theory of Knowledge, and will speak of Chisholm as
this historical 'Chisholm stage.' His views have
changed, at least in detail, as is to be expected from
a philosopher who is known for his variety of philo-
sophical analysis. I have been told that his current
account is more amenable to the view for which I argue
than the classic view presented in Theory of Knowledge
was.

Chisholm takes his task -- and that of all
serious epistemologists -- not as that of proving
that skepticism is false, but rather that of
providing an analysis of knowledge which fits our
reflective but preanalytic views about knowledge by
providing principles which will allow the adequate
epistemic appraisal of belief. His strategy for
dealing with radical skepticism is to deny that the
question "How is it possible to have knowledge" must
have an answer, in the sense of a reply sufficient to satisfy the skeptic. In the spirit of G.E. Moore, he takes it to be a matter of common sense that people generally do have the knowledge attributed to them. But in denying that we need to answer the skeptic, Chisholm seems to have -- illicitly -- dismissed the need to explain, rather than merely analyze, knowledge. Assuming that knowledge does exist does not entail that one must cease attempting to explain what it is and how it comes about; and Chisholm's common-sense rejection of skepticism has not shown that an analysis alone can be a fully adequate account of knowledge. Indeed, it is reasonable to insist that any plausible analysis be consistent with a theory that does explain the phenomenon in question. Thus, the "How" questions posed hereafter are not intended as skeptical objections but as requests for explanations, assuming, with Chisholm, that radical skepticism is dead.

Let us pause to rehearse Chisholm's account of immediate justification. First, Chisholm's central principles relating states of a person to the immediate justification of a proposition for that person; these are Chisholm's principles of the Directly Evident.

Note that I will not quibble over whether to count directly evident propositions as justified or as a
having a special epistemic status which is not to be thought of as justification. The former view is consistent with the view that knowledge is justified true belief; the latter with a view that knowledge is true belief which either has this special status or is justified ultimately in virtue of items which do. Chisholm avoids this issue through the use of "evident" rather than 'Justified" in his account.

D 2.1 h is self-presenting for S at t =Df h occurs at t; and necessarily, if h occurs at t, then h is evident for S at t.

D 2.2 h is directly evident for S =df h is logically contingent; and there is an e such that (i) e is self-presenting for S and (ii) necessarily, whoever accepts e accepts h.

(I shall label these and later principles in accordance with Chisholm: they all appear in the appendix of Theory of Knowledge.)

What are the items referred to by "e" and "h" in these definitions? They jointly imply that self-presenting items are evident, and that self-presenting things which are evident are the sort of thing which a person might accept. Further, his definitions D 1.1 to D 1.5 tell us that items which are evident are beyond reasonable doubt and such that only things which are certain can be more reasonably accepted. Two more definitions will serve, I think, to tell us that e and h are:
D 5.1 h is a state of affairs =Df It is possible that there is someone who accepts h.

D 5.2 h is a proposition =Df h is a state of affairs which is necessarily such that either it always obtains or it never obtains.

Presumably then, e and h are either states of affairs or propositions. In chapter 1 of Theory of Knowledge Chisholm equates "believes" and "accepts" and tells us that what one believes is always a proposition. This seems inconsistent with his definitions, however, since it implies that all believables -- states of affairs -- are propositions. In more usual language, propositions are thought of as the objects of propositional attitudes, like belief; Chisholm, a master of distinctions, wants to distinguish between objects of belief which obtain at some times and not at others and those which, like the sentences Quine called "eternal" are true or false independent of time.

I won't quarrel with Chisholm's slightly non-standard usage of "propoition" and "state of affairs," but I'm not adopting it either; hereafter I will use "proposition" for both, since nothing in what follows requires a distinction between propositions and states of affairs. Of perhaps greater importance is the question of whether Chisholm is committed to the existence of propositions as the relate of belief, i.e.
to the view that belief (or acceptance) is a genuine relation between persons and propositions construed as objects. It has been frequently argued that a general problem for foundationalist epistemological theories is that they entail ontological commitment to otherwise unwarranted entities.

Returning now to the relation of self-presenting states to evidence, we can now see that if a proposition (or state of affairs) is self-presenting for a person, that proposition is directly evident to him, as is every proposition that it entails -- in Chisholm's sense of entailment:

\[
D 4.5 \text{ e entails h } \equiv \text{Df e is necessarily such that (i) if it obtains then he obtains and (ii) whoever accepts e accepts h.}
\]

It does not follow from any of this that a person accepts the propositions that are directly evident to him. This is important since, as we will see, it apparently allows Chisholm to avoid an objection fatal to sense-datum theories.

According to Chisholm, self-presenting states provide a special connection between facts, i.e. states of affairs which obtain, and the epistemic justification of propositions. The states of affairs in question are states of the subject, and insofar as the subject is in the world, self-presenting states provide a link between knowledge and the world. It must be
noted, of course, that Chisholm does not construe this link as a causal connection. This is explicit in Chisholm's first principle of evidence:

(A) S's being F is such that if it occurs, then it is self-presenting to S that he is F.

"The predicates that may replace the schematic letter "F" in (A) are restricted to those which would yield a description of a self-presenting state." 2

States of being appeared to in a certain way are, for Chisholm, self-presenting states. Chisholm makes three provisos on this: first, that the 'ways' of being appeared to must not entail the existence of an object which is appearing to the subject, e.g. a sense datum. Second, that the 'ways' be restricted to Aristotle's 'proper and common sensibles.' These two provisos rule out "the wall appears green to me" (because it entails the existence of a wall) and "I sense a green datum now," but allow "I am now sensing greenly and wall-ly," assuming that 'wallness' is a common sensible, i.e. a feature detectable by several senses in concert though not by any single (proper) sense.

The third and most crucial proviso is that the sense of "is appeared to F-ly" used in this account is a non-comparative sense, i.e. its meaning is not to be

2. Theory of Knowledge, p. 139
analyzed in terms of things which are F. Thus, the non-comparative sense of "Jones is appeared to F-ly" is not "Jones is being appeared to in the way he is typically appeared to by things which are F." As we shall see, this proviso is crucial to Chisholm's analysis. ³

In summary, Chisholm claims that certain states (self-presenting states) of subjects are such that if a person is in such a state, he is ipso facto justified in believing that he is in such a state. As previously noted however, a person's being in a self-presenting state does not entail that he does in fact believe that he is in that state; just that if he did so believe, his belief would be justified. One should note that this means that these directly evident propositions are not intrinsically justified in the sense introduced earlier; for it is not in virtue of believing a proposition about a self-presenting state that a proposition has an epistemic warrant.

This formulation of the connection between sensation and knowledge seems intended to avoid objections to foundationalism like that presented by Sellars

³. Theory of Knowledge, p. 30-33.
in his argument against classical sense-datum theories in the opening sections of "Empiricism and the Philosophy of Mind."

There, Sellars argued that classical sense-datum theorists wrongly bridged the gap between sensing particulars and knowing facts by construing

(S) X senses a red sense datum S

as entailing

(B) X senses S as being red

The sense datum theorist construed sensation as an unacquired ability; and the connection between (S) and (B) provided, they thought, a level of immediately justified perceptual belief. But Sellars points out that this attempt commits the sense datum theorist to an inconsistent triad:

A. X senses red sense contents entails x non-inferentially knows that s is red
B. The ability to sense sense contents is unacquired
C. The ability to know facts of the form x is o/is acquired.

Sellars points out that the sense-datum view he criticizes seems to be the result of conflating the idea that there are inner sensory states with the idea that there are inner cognitive states, i.e. states involving judgments, like believings and knowings; he goes on to argue that this conflation is unjustified
since the reasons given for postulating sensory states and those given for postulating cognitive states are independent: in particular, "... there is no reason to suppose that having the sensation of a red triangle is a cognitive or epistemic fact." 4

Chisholm's view apparently avoids this conflation of sensory states and cognitive state and thus, it seems, avoids the sort of criticism Sellars used against sense-datum theorists. It is, however committed a very strong (necessary, according to Chisholm) connection between sensory states and epistemic facts. The present question is whether, and to what degree, Chisholm's connection is objectionable.

Chisholm's arguments for the connection are, in a sense, Kantian: as he sees it, such a connection is the only way to avoid an infinite or circular regress of justification for beliefs about empirical, contingent states of affairs. And yet he does not explain how it can be that a person's being in a given state can make it true that he would be justified in believing something, were he to have the appropriate belief. I use this somewhat awkward formulation, for

Chisholm's principle connects a person's being in a self-presenting state with his being such that the belief that he is in such a state is evident to him; and Chisholm's theory does not require that one believe that which is evident to oneself. Chisholm holds that a belief is evident for $S$ just in case accepting it is more reasonable than withholding it and more reasonable than any but the most reasonable (certain) beliefs.

However, Chisholm's view entails the consequence that a proposition's being evident to a subject entails that the subject understands the proposition. This follows trivially from his analysis of evident belief, his principle (AI), which states that if any proposition $h$ is more reasonable for $S$ than any other, $i$, then $S$ understands $h$, and the premise that there are some propositions which are less reasonable for $S$ than any evident propositions.

Thus, Chisholm's analysis commits him to the view that a person's being in a self-presenting state entails that he understand what it is to be in that self-presenting state. Since he equates being appeared to redly with being in a particular sort of sensory—or phenomenological—state, and identifies at least

5. Theory of Knowledge, p. 27-29.
some of these, including being appeared to redly, as self presenting states, it follows that a person who senses redly -- or is appeared to redly (both ontologically neutral versions of 'has a red sensation') ipso facto understands what it is to sense redly. But since newborn infants, without previous experience, are presumably capable of sensation, this entails that they have an unacquired ability to understand some presumably empirical proposition, a proposition the understanding of which, according to empiricists, must have been acquired through experience. Thus, if Chisholm's theory is to be acceptable within a broadly empiricist tradition, he must explain how it can be that the first time a person is appeared to redly, he understands the proposition that he is appeared to redly. Note that in The Language of Thought, Fodor explicitly embraces precisely this form of innatism; on his view one could not learn a word of a public language unless one antecedently had an internal language which incorporated the concept expressed by that word. We will return to Fodor in section 19.

Of course, the preceding argument does not conclusively establish that Chisholm's analysis is incompatible with empiricism, or that the way in which it is incompatible -- if it is -- is less reasonable than the empiricist view with which it supposedly
conflicts. What it does establish is that Chisholm is seemingly committed to a principle, which taken together with two traditional empiricist principles, is part of an inconsistent triad paralleling the one entailed by sense-datum theories:

(A') 'S is appeared to F-ly' entails S understands the proposition 'S is appeared to F-ly' (or its first-person variant 'I am appeared to F-ly').

(B') The ability to be appeared to F-ly is unacquired.

(C') The ability to understand propositions about experience is acquired.

Of these, Chisholm apparently rejects (C') on the basis of his third proviso that there is a sense of appearance terms such as occur in "I am appeared to F-ly," the understanding of which does not presuppose comparative experience. His claim is made as a rejoinder to a view which holds that the only sense in which 'S is appeared to F-ly' can be understood is given by

'S is appeared to in the way F things appear to S'.

According to this view, appearance terms are to be analyzed in terms of comparisons between present and past experiences where the sensible characteristics of the object of experience is known.

As noted above in Chisholm's account hinges crucially on his third proviso: without it, circularity
would ensur. For in order for it to be evident to a person that he was sensing whitely, he would first have to know what it was like to sense white things, and hence he would have to know when white things were presented to him. Unfortunately, Chisholm's only argument to establish that there is a noncomparative sense as such terms is an ignoratio elenchi; he argues that three existing arguments against claims that such terms can have a noncomparative sense are invalid.

An apparent problem with Chisholm's claim that there is a noncomparative sense of appear words is that it seems to require that the subject understand what it is to sense, say, whitely, independent of previous experience; hence if Chisholm is right, there not be an innate understanding of the ways things can be sensed. Chisholm neither acknowledges nor rejects this implication, and he makes no attempt to explain or otherwise argue what appears to be a variety of innatism.

I must the ability to understand this non-comparative sense appear words be innate? Surely not: for the individual could be acquired it say, via operant conditioning. After a scheme of reinforcement of pairings of, say red objects with the subject's utterances of "red," he may have acquired the disposition
to utter "red" in the presence of red objects, and thus, perhaps, the ability to correctly judge objects to be "red" (under, of course, 'standard conditions'), without doing so by comparing his present experience to memory images or to anything else. This assumes that acquiring the ability to correctly use 'red' entails acquiring the concept of redness.

While arguments against the possibility of a non-comparative sense of appear words have intuitive appeal ('what could we mean by "red" but "things like apples, firetrucks, blood, tomatoes, etc."; or 'how could we learn the meaning of (or the correct use of) "red" except by comparison of such items'), the possibility that humans could acquire a disposition to use "red" when confronted by red objects through operant conditioning may be thought to remove much of the force of these arguments.

A variation on these arguments against non-comparative senses of terms holds that all classification involves comparison; that in order to classify an item as red (i.e., to judge it to be red), one must compare one's present phenomenal state, or the appearance of the object in question, to remembered appearances and what one antecedently knows about their color. But the claim that all classification requires comparison is simply false. Sorting machines can classify punch
cards not in virtue of comparison, but in virtue of their construction and programming. And even very simple biological organisms, are able to react to various stimulus-conditions in seemingly 'rule governed' ways, presumably not depending in any interesting sense on comparison.

Given an acceptable account of non-comparative appear words, can Chisholm use them as a foundation upon which to build a structure of knowledge? In order to do so, he must at least be able to get started: to get from its being the case that a person is in a state of being appeared to redly to the persons being justified in believing that he is appeared to redly. But can Chisholm make this basic step without presupposing other knowledge on the part of the subject? If not, his 'foundation' will fail in its role as a supposed regress-stopper.

At this point some features of the previously sketched account of how a person could be able to use appear words non-comparatively should be examined. Note first, that the mere fact that an individual has acquired the disposition to utter 'red' in the presence of red objects does not imply that the individual understands what it is to be appeared to redly. For the latter requires that the person understand that homogenous red appearances are not simultaneously blue;
and that they are colored. Understanding these features depends on things other than the disposition in question. Hence, in discussing the non-comparative sense of "red" it is more appropriate to treat "red" as a mere sign-design than as an interpreted (English) expression.

Thus, the acquired non-comparative 'use' of "red" discussed above is not a use which itself involves understanding the rich system of categories and implications a person must understand in order to understand the English expression "red." Another way to put this is that this use of "red" is not one in which "red" has its full cognitive status. This is a familiar Sellarsian point; that a person can have a blue-sensation (or be appeared to bluely) without having any beliefs or abilities to believe about, or understanding of color. For to have a belief about a color, say red, one must understand a battery of color concepts.

Turning to a television-signal analogy, it is obvious that even though a t.v. camera circuit may be carrying a blue signal -- having a built-in disposition to do so when confronted by blue objects -- it makes no sense to say that it has a belief involving a color expanse -- even to those who believe that machines which
can pass appropriate turning tests can think; for a television camera cannot pass any plausible test of cognitive ability.

A Brief Digression on Fodor and Innateness

There are strong parallels between Chisholm's analysis of knowledge and Fodor's account of language learning. In the first place, Fodor, like Chisholm, is committed to the view that persons have innate concepts. Fodor's argument for concept innateness is based on his view that the only plausible explanation of how a person learns a natural language. Let's see what Fodor says. First his assumptions:

To begin with, I am going to take three things for granted: (1) that learning a first language is a matter of hypothesis formation and confirmation in the sense explored in Chapter 1; (2) that learning a first language involves at least learning the semantic properties of its predicates; (3) that $S$ learns the semantic properties of $P$ only if $S$ learns some generalization which determines the extension of $P$ (i.e., the set of things that $P$ is true of). 6

Now for the account:

If learning a language is literally a matter of making and confirming hypotheses about the truth conditions associated with its predicates, then learning a language presupposes the ability to use expressions coextensive with each of the elementary predicates of the language being learned. But, as we have seen, the truth conditions associated with any predicate of $L$ can

be expressed in terms of the truth conditions associated with the elementary predicates of L. The upshot would appear to be that one can learn L only if one already knows some language rich enough to express the extension of any predicate of L. To put it tendentiously, one can learn what the semantic properties of a term are only if one already knows a language which contains a term having the same semantic properties. 7

I agree with Fodor's admission that this account is tendentious, (I find it outrageous, in fact) but disagree with his claim that it is the only available explanation of (public) language learning, in any but a stipulated sense of 'learning.' Surely his assumption that

Learning a first language is a matter of hypothesis formation and confirmation.... 8

violates ordinary intuitions about learning. But, to avoid a merely verbal issue, let us grant Fodor this sense as a stipulation.

Now given this definition of Learned, does it follow that there are innate concepts? Sure, if innate concepts are unlearned in Fodor's sense; it would then follow trivially from Fodor's definition of 'learned,' the stipulation that innate = unlearned, and his account of language learning.

7. Ibid., p. 80.
8. Ibid., p. 59
Learning or Acquisition?

The flaw in this account is obvious; for those who deny concept innatism don't take themselves to be denying that unlearned concepts are unlearned; what they, and I, deny is that unlearned concepts are unacquired. What is needed is an account of concept acquisition. What is surprising is that Fodor has ignored the most plausible approach: conditioning.

Creating Conceptual Functions

Well, it's not really very surprising that Fodor ignores conditioning, for it is an idea central to a theory he rejects; Skinnerian Behaviorism. Oddly, I accept Chomsky's rejection of operant conditioning as an explanation of a person's total linguistic abilities. Chomsky's argument is that the theory of operant conditioning cannot explain a person's ability to construct and understand novel sentences. But, importantly, Chomsky's argument does not show that operant conditioning, or something like it in at least the sense that it does not presuppose concepts, cannot explain an individual's acquisition of the (correct) use of individual expressions. What I propose is simple; an individual linguistic competences, including his ability to respond to tokens of "red" with "color."
In order to explain the ability to construct and understand novel sentences, one must go beyond behaviorism; here, Fodor seems right in claiming that a computational model, as provided by functionalist philosophy of mind, is required to cure the shortcomings of behaviorism. Indeed, in a more recent work, his primary criticism of behaviorism is that it cannot account for the interaction between mental states:

Mental causes typically give rise to behavioral effects by virtue of their interaction with other mental causes. For example, having a headache causes a disposition to take aspirin only if one also has the desire to get rid of the headache, the belief that aspirin exists, the belief that taking aspirin reduces headaches and so on. Since mental states interact in generating behavior, it will be necessary to find a construal of psychological explanations that posits mental processes; causal sequences of mental events. It is this construal that logical behaviorism fails to provide. 9

On a more sympathetic reading that I am inclined to accept, Fodor might be thought of as pointing out flaws in the assumptions with which he begins. At least part of the problem with his view seems to lie in his willingness to discard what is good in behaviorism along with what is bad; he's not alone, of course, in throwing out the baby with the bathwater.

My proposal suggests that something like operant conditioning can be used to explain one important aspect of how people acquire (not learn, at least in Fodor's sense) the correct use of words and functional roles that are their meanings; neurophysiology explains how operant conditioning works by describing the causal mechanisms that underlie it. Thus, conditioning is not, by itself, enough; but together with functionalist accounts mind and meaning, and a neurophysiological account of how pathways in neural structures can be caused by repeated conditioning, it explains how a child can be 'programmed' in such a manner as to acquire the ability to recognize objects, words, attributes, etc.

The programming of which I speak here must not, of course, be thought of as 'writing instructions in a language'; but rather as making changes in the physical make-up of the person. This is like the 'programming' of early computers by physically changing the wiring of the machine to create new operational capacities. This form of 'programming' violates the usual 'hardware-software' distinction. In computers today, these changes are sometimes made by altering a read only 'memory' in such a way as to change the machine operations (in effect, changing its wiring diagram), thus changing, and in extreme cases, even creating its
machine language. This sort of 'programming' is called microprogramming and is usually considered to be part of the machine architecture. In Dennett's terms it is appropriately thought of from the design stance, not the intentional stance, since its purpose is to create structures which can function in a manner that can be interpreted from the intentional stance as meaningful.

An Out for Chisholm?

We are not finished yet, though; for Chisholm could avoid the innateness charge by restricting his analysis to persons who have acquired, through something like operant conditioning, an entire battery of linguistic dispositions, rich enough that it makes sense to say that they have acquired the requisite conceptual system, and are also disposed to believe that they are appeared redly to when they are appeared redly to.

This seems very much like the view expressed by Sellars in his Carus lectures. There he sketches the development of a persons mental capacities from preconceptual dispositions through bare abilities to categorize (ur-concepts) to a fully integrated and interconnected system of concepts. This progress would begin with the persons (or proto-persons) being conditioned to be able to respond verbally to objects.
with utterances that sound like predicates and names and end only when the person had a full set of linguistic competences. The training is provided by the responses of the community of persons that surround the learner.

But how could people be conditioned to believe that they are being appeared to redly when they are being appeared redly to if belief and sensation are essentially private, subjective states? How, that is, could other people train (condition) a person to have this disposition (if the trainers did not know when the trainee was sensing redly? The obvious response to this Socratic question is that the trainers do, of course, know when trainees (children) are sensing redly. It is when the trainee is awake, in normal perceptual conditions, and has a red object in front of her. The trainer might be wrong; the child may have a neurophysical defect, or there may be other problems. But notice that the trainer does not begin by training the child to respond to sensations; he begins by training the child to respond to external objects.

It should be obvious to those familiar with Sellars' views that in my attempt to rescue Chisholm's view from the charge that it requires the innateness of many empirical concepts and thus resigns the nature of belief (and other psychological states) to mystery, I
am leading Chisholmian analysis toward Sellarsian explanation. The possibility that persons may acquire such dispositions and abilities is relevant to the justification of belief. Part of the reason they are justified in believing that there is a red object before them is that they have acquired the capacity to make reliable perceptual judgments.

Thus the important difference between Chisholm and Sellars is not that Chisholm embraces and Sellars rejects a basic kind of knowledge; it is rather whether or not the foundation presupposes an unacquired connection between sensory and cognitive states, and, hence between sensory states and epistemic warrant. For on the Sellarsian view, the warrant of foundational propositions is not derived by inference from justified propositions, but is an immediate result of his having acquired a conceptual framework and his being in a state which, given this framework, provides an acquired connection between the world and cognitive states. The connection between belief is thus at once mediated and immediate; it is mediated by (psychologically classifiable) causal processes, but not epistemically mediated by inference from previously justified or evident propositions.
Note that from our 'omniscient observer' point of view, knowing that Jones has acquired a language and the dispositions to say -- or think -- that there are red apples in front of him in ordinary circumstances only when there are red applies in front of him (his acquisition of language makes him a reliable red-apple detector in the sense that it establishes a nomological connection between his saying or thinking "Lo, a red apple" and the presence of red apples), his beliefs that there are red apples in front of him are warranted, though not guaranteed. And this warrant attaches to those beliefs whether or not he is justified in believing he is reliable about such things, so long as he is (in the nomological sense). Thus the justification is not inferred from some antecedently justified belief. Nor need he have the justification, in the sense that he has, or could, actually construct an inference based on what he knows, in order to be justified. The upshot of this is that Jones can know that there are apples in front of him without knowing that he knows.

What then of knowledge of ones own pains, sensations, or thoughts? How could a person come to have noninferential knowledge of these things if they are not a matter of direct acquaintance? And how could one be trained to recognize that one was in one of
these mental states if knowledge of them is a matter of
direct acquaintance with private inner states?

Beginning with pain, the answer should again be

clear. A child's parents do know when the child is in

pain and can hence train the child to respond to it

with "pain." Sellars' rational reconstruction of how
people could learn to reliably report their mental
states -- noninferentially-- is the closing section of
"Empiricism and the Philosophy of Mind." He begins

with an account of how a public, scientific theory of

thought and sensation could begin with a study of

behavior -- much as current cognitive science has,

In this theory, judgments about a person's mental states
are at first justified on the basis of behavioral
evidence alone, and is hence, at that stage, inferential.

But it need not remain at that stage. Thus, people

begin by using the language of impressions to
draw theoretical conclusions from appropriate
premises. (Notice that the evidence for theoretical
statements in the language of impressions will
include such introspectible inner episodes as its
looking to one as though there were a red and
triangular physical object over there, as well as
overt behavior. Finally he (a person who under-
stands a theory of thought) succeeds in training
them to make a reporting use of this language.
He trains them, that is, to say "I have the
impression of a red triangle" when, and only when,
according to the theory, they are indeed having
the impression of a red triangle.

Once again the myth helps us to understand that

concepts pertaining to certain inner episodes—in
this case impressions—can be primarily and essen-
tially inter-subjective, without being resolvable
into overt behavioral symptoms, and that the reporting role of these concepts, their role in introspection, the fact that each of us has a privileged access to his impressions, constitutes a dimension of these concepts which is built on and presupposes their role in intersubjective discourse. It also makes clear why the 'privacy' of these episodes is not the absolute privacy of the traditional puzzles. For, as in the case of thoughts, the fact that overt behavior is evidence for these episodes is built into the very logic of these concepts as the fact that the observable behavior of gases is evidence for molecular episodes is built into the very logic of molecule talk.10

Objections to Sellars that were popular twenty years ago claimed that his account failed at this stage because it makes no sense to think either that knowledge of ones mental states had to be acquired or that it could be both acquired and noninferential. These objections can carry very little weight today, given that the best account of mental states accords them the status of functionally classified theoretical entities constituted of physical stuff.

Thus within the systematic approach to philosophy of which the "new" functionalist theory of mind is a part lie answers to other traditional philosophical problems; answers which are consistent with the requirements of both science and common sense.

In examining Chisholm's analysis we discovered several problematic principles which Chisholm had sought to justify by claiming a dilemma; these or skepticism. These principles are especially problematic in that the analysis they provide seems to deny the possibility of explaining the reliability of belief and even the possibility of knowledge of the kind we so obviously (from the common sense viewpoint) have. But swallowing the claim that the rejection of skepticism entails the acceptance of mystery unexplained like swallowing a cow to get the tenderloin. That a naturalistic explanation of knowledge should come from psychology should have been expected.

In the end, Chisholm's analysis rested on his concept of self-presenting states, which in turn rested on a non-comparative sense of appearance words. Our explanation of the ability to make reliable non-comparative judgments provides an account of non-inferential justification consistent with the view that empirical concepts are acquired, and allows for a theory -- rather than a mere definition -- of knowledge. The theory depends on the sort of functionalist psychological theory that took the spotlight in part II, which thus provides a systematic approach to issues in epistemology and the philosophy of mind.
CHAPTER XIV
TWO ASPECTS OF PERSONAL IDENTITY

The thesis of this chapter is that there are two aspects of personal identity which have been wrongly conflated in recent philosophical literature; these aspects can be distinguished by considering two questions. First, when do two persons have exactly similar personalities? And second, when is it the case that Jones and Smith are identical, i.e., are the same person?

With regard to the latter question, the analysis which emerges below is based on a view similar in many respects to views held by Strawson and Sellars. Following them, I assume that persons are individuals of whom both mental and physical predicates are true, and that persons are always at least physical objects, i.e., objects having such physical properties as extension, mass, and location.

Some may interpret this as inconsistent with certain viable philosophical viewpoints; it may seem, for instance, to be in conflict with classical Cartesian dualism. But this supposed conflict is an illusion,
for no classical dualist holds that a person is only a mind (or soul) though some think persons can continue to exist after bodily death. I admit the possibility of conflict with some forms of idealism, but will dismiss them without argument here; there has been little recent controversy over their rejection. The argument that since for an object to be physical is for it to be subsumed under physical concepts, and since all concepts are themselves mental entities, physical objects are therefore merely metal constructs (as opposed to objects susceptible to mental representations) is at the heart of many idealist doctrines but is clearly invalid. That persons have succeeded in attaining conceptual frameworks for carving up and making sense of the experienced universe is not sufficient to show that the universe is itself nothing but a product of such carvings.

Before turning to the highway of argument, several sideroads deserve our attention. It is sometimes held that persons must be human, i.e., members of our peculiar biological genotype. I see no reason for adhering to this hackneyed dogma, for it would unreasonably exclude the possibility of there being persons (extraterrestrials, perhaps) of different genotype, or of altogether different constitution from ours, being like nothing as yet classified as biological.
Lacking a rigid biological criteria as a characteristic sufficient to distinguish persons from non-person, what is one to use? One answer is that mental criteria provide distinguishing characteristics, but this is not to say that mental criteria are both necessary and sufficient; rather, this is to say that persons must possess mental characteristics, that mental properties are necessary to persons, and perhaps, to their identities. Asking which, and/or how many mental characteristics are sufficient for personhood, much less for personal identity, has misled many philosophers who have forgotten, or not noticed, that what is properly asked for are mental characteristics sufficient to fulfill the conditions necessary for personhood, or personal identity.

This sidestreet has led (like a tedious argument, I'm afraid) to the questions "What are mental properties?", and "What must a thing be like in order for it to have a mental property, or for it to be a mental thing?" Persuaded by the arguments of Geach, Sellars, Fodor and other recent theorists, I assume that mental entities are items postulated by psychological theories as entities which play essential roles in the production of 'intelligent' behavior; mental predicates are predicates definitionally true--within such a theory--of these postulated entities, and are terms serving in
psychologically adequate explanations of behavior. These predicates characterize functions fulfilled by organisms, notably humans, in behaving, or, more to the point, in their intelligent activity (for it is intelligent activity, as opposed to mechanical, automatic, or 'unquided' behavior, which is the special domain of psychology). Mental entities are those entities broadly construed so as to include events and processes, which mediate the functions characteristic of intelligent behavior. Thus, a thought is an entity which can play a role in mediating behavior. Those organisms having the kinds of mental states which typically result in human behavior (whether or not the organism in question is human or is capable of the required behaviors -- as quadraplegics may not be) are to count as persons. I will not go into what counts as 'typically human' behavior except to note that many philosophers have specified this in terms of linguistic behavior.

It is important to note that nothing in this view prevents mental entities from being physical entities; for a mental entity is a thing -- whatever it happens to be made of -- which plays the functionally characterized role which defines that entity. Similarly, while genes are defined in terms of inheritance of traits, nothing in genetic theory precludes
the possibility that a gene might be a biochemical object: one which has the function of transferring a trait.

Lewis and Parfit On Survival and Identity

In several articles spanning the last decade, Derek Parfit has argued that personal identity is not as important a concept as tradition had held, because the important issues which were usually thought to rest on it could be given independent analysis. In particular, says Parfit, the problem of personal survival could be so handled; his implication is that concepts like the assignment of moral responsibility could be given similar (or the same) treatment. Parfit goes beyond this, however, when he argues that the correct analysis of survival is not only independent of the notion of personal identity, but that the two were incompatible: hence personal identity cannot be what is important in such important questions about persons as that of survival. His argument is simple:

P1. Mental continuity and connectedness is what matters in survival.

P2. The relation of mental continuity and connectedness need not be a one-one relation; it can be one-many or many-one; it can also admit of degree.
Parfit would, I think, agree with Lewis' argument so far; his example seems based on very similar intuitions; Parfit describes a case of one person surviving as two (I hereafter refer to this sort of case as a case of fission). Joe is operated on, his brain split and each half placed in a new body. Assume that the operation is successful and that the new individuals, Lefty and Righty both have a high degree of mental connectedness and continuity with pre-operation Joe; Joe's mental life "flows on" in each of them. (I will hereafter refer to this as the Joe case). On Parfit's view, Joe survives the operation as Lefty and as Righty; Lewis agrees (with a slight reservation). But, according to Parfit, Joe cannot be identical with Lefty and Righty, for one thing can never be identical with two.

Lewis accepts Parfit's premises P1 and P2, but denies Parfit's conclusion. He does this in three steps, arguing first that identity is not a relation among the temporal stages of an object, but is a 'relation' the entire object bears to itself. Second, he argues that there is a relation between stages which holds just in case they are stages of a single object; he calls this the I-relation. Third, Lewis argues, the I-relation for person stages just is the relation of sufficient mental continuity and connectedness (which he calls the R-relation). Finally, Lewis replies to
the intuitive objection to identity as a criterion of survival captured in cases like that Parfit describes as "one person surviving as two," by saying that in such cases of apparent fusion, the "one" person who survived as two was not one person at all, but two persons sharing a single body during some of their stages, i.e., during some parts of their histories, just as two highways, U.S. 66 and I40 share pavement in some places.

This argument deserves closer scrutiny. First recall that the I-relation is a relation holding between two states if and only if there is a person including those stages. Persons, then can be reciprocally defined as maximally I-related sets of person stages, i.e., the largest sets such that each member is I-related to every other member. Lewis holds that survival is a matter of identity; thus S survives event e iff there is a person stage Si after e which is I-related to S's present stages. Lewis then equates the I-relation for persons with the R-relation. Recall Parfit's case now: Joe is operated on, his brain split and each half placed in a new body. The fission is successful and Lefty and Righty are each sufficiently R-related to Joe (i.e., each of them has a high degree of mental connectedness and continuity with pre-operation Joe;
Joes's mental life "flows on" in each of them. (I will hereafter refer to this as the Jose case). Lewis and Parfit agree that Joe survives the operation as Lefty and Righty; Parfit claims that Joe cannot be identical with Lefty and Righty, for one thing can never be identical with two. Lewis, however, argues that so far as personal identity is concerned, there were two persons all along; there was a preoperation stage $j$ of Joe; and postoperation stages $r$ and $l$, of Righty and Lefty, respectively, such that $rRj$, $lRj$, and that $r$ and $j$ are members of one maximal set of $R$-related stages while $r$ and $J$ are members of another. There is no maximal set including both $r$ and $l$ stages though, because Lefty's mental life diverge sufficiently from Righty's shortly after the operation. (We will see problems here later). Since there are two distinct maximally I-related sets of person stages, there were two distinct persons all along; before the operation they simultaneously coexisted in Joe. Lewis prefers to say that before the operation, "Joe" is ambiguous, naming two persons, Lefty-Joe and Righty-Joe.

There is, of course, a flaw in this argument, as P, Maddie points out. Imagine that the Joe case is slightly modified; that during the day after the operation, Lefty and Righty remain in the hospital, in identical suites, visited by identical-twin nurses,
and that by coincidence, their experiences are phenomenological duplicates. Then, that night, Lefty is attacked in his sleep, and dies as a result. Had we to expected an attack and asked, before the operation, "Will Joe-Lefty survive the attack?", Lewis' account will give the wrong answer, since before the operation, any stage of Lefty was indiscernable (qua stage) from a stage of Righty. Thus considered before the operation, Lefty was R-related to all the pre-operation Joe-stages, which are also R-related to all of Righty's stages. But, if Lewis is right, they must also be I-related; hence, contrary to our supposition, Lewis must say that Joe-Lefty survives.

This seems to be a genuine counterexample to Lewis' view, since his analysis concludes wrongly that Lefty survives the attack. This conclusion is based on Lewis' conflation of the I and R relations. But if they are not the same, then Lewis' argument fails to establish that what matters in survival is identity.

In a recent paper on this topic Penelope Maddie has argued that branching cases (fission, as described, or its reverse, fusion) interrupt identity, and that in the Joe case there are three non-identical persons, and that Joe survives in Lefty and in Righty. She concludes that the importance of personal identity is
derivative, as Parfit had claimed, but does not note, as I do here, the ambiguity of 'personal identity.'

What has gone wrong in Lewis' and Parfit's analyses? An element apparently common to both is that they take persons to be constituted out of person-stages; perhaps they have erred in taking person-stages to be basic.

While I think that this objection has much to be said for it, the analysis I shall argue for finds an even more basic flaw; for Lewis and Parfit seem concerned not so much by relations between person-stages as relations between the mental stages of persons. Remember, they claim that what is important for survival—and, for Lewis, for identity—of persons across time is mental continuity and connectedness; the R-relation is a relation among the mental characteristics of persons, though Lewis treats it as relations between mental entities simpliciter. The distinction between mental characteristics and mental substances (like Cartesian minds) will play a crucial role in the view of personal identity that I will later defend.

But what is to be said of the strategy of taking temporal stages of objects as basic in the analysis of identity across time? Something seems to have gone awry here; for it is surely objects which exist at various times which are basic, and not their artificially
defined temporal segments. But can we not analyze continuant objects in terms of temporal segments without claiming ontological priority for the latter?

I can see nothing wrong in making a new analysis to solve existing problems so long as the details of the phenomenon under consideration are not lost in the procedure. And in analyzing "identity" across time are we not asking, as Lewis has suggested, whether object A, at t1 is the same as object B at t2 by asking whether the artificially created A-at-t1 is a temporal segment of the same basic entity as the artificially created B-at-t2? And, if it is granted that 'A-at-t1' is a term of art, rather than the name of an ontologically basic item, what objections remain to the analysis? To do justice to this view what needs to be shown is that the analysis can itself be analyzed in a manner not requiring ultimate reference to temporal segments. It has been suggested that they can; just as possible worlds can be dispensed with in terms of primitive modalities, temporal segments can be dispensed with in terms of tenses. That they can be dispensed with to maintain a parsimonious ontology is consistent with the fact that we can speak of them in constructing analyses, and should do so if it makes the analysis understandable. I will leave the job of the ontological reduction of temporal segments to continuants and tense
to those following in the footsteps of Prior and, more recently, Kripke. Assuming that such a reduction is available, I shall continue to investigate the analysis of 'identity' across time with respect to persons and their survival.

**Persons and Personalities**

In the last section we saw that Lewis' attempt to prove that survival is a matter of personal identity was flawed. After divorcing personal identity and survival, Parfit concluded that we can answer questions of survival, and similar importatn questions, without answering questions of identity. Unfortunately, Parfit's ploy has at least one major problem: presumably he intends that we believe that some thing has survived without providing identity conditions for that thing. But this is in conflict with strong intuitions regarding the connection between existence and identity: intuitions regarding the connection between existence and identity: intuitions Quine has encapsulated into the aphorism "No entity with identity."

The natural rejoinder to Parfit's scheme is found in asking, of each analysis, "what is it that has survived, and what are its identity criteria?" Lewis, motivated at least in part by the Quinean intuition that
existence goes hand in hand with identity, and survival with continued existence, apparently thinks that what has survived is the person, and that the appropriate criteria are those of what he has been calling 'personal identity.' The ensuing debate, it seems to me, has generated more heat than light, because an unexamined presupposition has stoked the furnace without generating any electricity.

Let us backtrack and ask again "what is it that has survived?" If not the same person, then what? A premise accepted by both Lewis and Parfit provides a clue: "What I mostly want in wanting survival," says Lewis, "is that my mental life should flow on." In short, what this suggests is not that the same person should survive, but that a person's mental life should survive.

What then is this 'Mental Life' if it is not a person? And what are its identity criteria? Again, Parfit and Lewis supply a clue: Mental continuity and connectedness. A mental life is a mentally continuous and connected set of such mental actualities as thoughts, beliefs, hopes, wishes, desires, memories, experiences, and traits of character (or characteristic dispositions); not just any such set, of course, but one like those a person has throughout his (ordinary)
lifetime. For a person to survive ordinary death is for his mental life to continue, i.e., for the set not to end with the event (perhaps experienced) of ordinary death, but to include after ordinary death similar mental actualities, exhibiting the requisite degree of continuity and connectedness with those before the (ordinary) death.

This talk of 'mental life' may appear to entail something like a theistic view of the soul. I shall argue that this appearance is illusory, after noting that Strawson's position explicitly allows the possible existence of disembodied persons -- though they would have a 'logically secondary' form of existence, would have no contact with other disembodied souls and would be incapable of action, since the capacities for action and experience are bodily capacities. Strawson would limit even knowledge of oneself to memories of one's embodied existence. Aristotle would not have allowed this to provide criteria for individuation, for such memories would necessarily be formulated in terms of universals, sensible particulars having been 'burned away' by the abstractive powers of the active intellect. Aquinas had notorious difficulty in addressing the issue of disembodied survival; I have explored his views in an earlier paper. Though disembodied survival
may be consistent with some accounts of the mental life, I do not think that an account of the mental needs to, or should, entail either the truth or even the possibility of such views.

It will be recalled that in section I our discussion of the nature of mental entities such as thoughts and beliefs assumed that they were theoretical entities postulated as a means of providing psychological explanation. Now to say that a person's thought, at t, that so-and-so, is a theoretical entity is not to say that a person can have no knowledge of his thinking, cannot correctly think that he is thinking that so-and-so. Nor is it to deny that thoughts are about objects. The issues involved in explaining these points so far beyond the compass of this work, requiring a systematic exposition of issues in philosophy of mind, epistemology, and ontology such as that found in Sellars' *Science and Metaphysics*, and extended, to some degree, in his "Metaphysics and the Concept of a Person." I will draw a few important points from the latter before continuing.

First, a thought can be a particular without being an ontologically basic particular. The term 'thought is a verbal noun, and like other nouns leads persons to think that it refers to an object. But as Sellars has argued,
(N) Ann has an occurrent thought now
is strongly equivalent to
(V) John is thinking now.

Even though the latter does not seem to refer to a
mental particular, but to describe an occurrence of a
kind of activity, process, or state in John. This
activity should not be thought of on the model of the
performance, by John, of an action, a piece of conduct.
The point is that (N) is a derivative way of expressing
(V) much as

(H) The ball has redness

and

(E) The ball exemplifies redness

are derivative from

(R) The ball is red

and are to be explained in terms of the peculiar
linguistic functions played in these contexts by 'has'
and 'exemplifies,' a role similar to that of 'is true.'
The point is that these mental entities are not inde­
dependently existing objects but states of (or processes
occurring in) persons much as a smile is not an inde­
pendent object, but a state of a persons face.

Secondly, that-clauses in, e.g. "John thought
that the Eiffel tower is in Paris" are functional
sortals, classifying such processes. A less misleading
parsing of
'John has a thought of Paris'

would be

'John is thinking of Paris.'

which, drawing on the lessons of an adverbial theory of the mental, becomes

John thinks Paris-ly

Finally, to think Paris-ly is to be involved in a mental event which is subject to a functional classification; using Sellars' dot-quoting device for functional classification, we end up with

John 'Paris' es

Thus, to say that a person has a mental item of a certain kind is to offer a functional classification of one of his (often speech-like) inner processes. Similarly, to say that a person has such-and-such a mental life during a certain time period is to give a functional classification of all of his mental processes, states, and dispositions throughout that period.

In outline, then, the present theory of personality says that

1. Mental states are functionally classified states of persons.

2. Since two or more persons can share a functional classification (e.g. "is thinking Parisly", "is sensing sharp pain-in-right-index-fingerly"), they can 'have' the same thoughts, sensations, et. al.
3. A mental stage of a person $S$ at $t$ is the set of mental states had by $S$ at $t$.

4. A mental life is a maximally $R$-interrelated set of mental stages. Thus, a person $S$'s mental life is typically the set of stages associated with $S$'s body.

Thus to say of a set of thoughts, etc., (i.e. functionally classified states of a person), that they constitute his mental life, is to say that he is a certain (specific) type of individual, one whose states have certain dispositional features. Another way to put this is that he has a certain sort of personality, by which I mean 'specific mental life type.'

If one takes seriously the notion that an occurrent thought is an inner sentencing, and construes the medium in which this sentencing qua token, occurs, as neurophysiological, then the type-token distinction can be used to explain what it is for two persons to have the same thought; it is for them to be in states which are tokenings of the same functional type. Just as distinct tokens of "Chicago" can have radically different physical make-up (ink on paper, lightbulbs on blimps) and the same linguistic function can exist even in different languages, as exemplified by 'red' in English and 'rot' in German, the physical descriptions of such neurophysiological tokenings can differ radically.
If two persons can have the same thought-type, why can't they have the same series of thoughts; and beliefs, hopes, desires? The probability of such duplication throughout life is, no doubt, astronomically low, and would require quite similar physiologies, environments, and the like, but here seems to be no reason to think this impossible in principle.

Finally, if two persons can have the same mental-life-type, cannot two human bodies, A and B be so related that A has mental stages M<at<1>, M<at<2>, . . . M<at<n>, and B has Mbt<n+l>, Mbt<n+2>, . . . Mbt<n>, and such that Mat<n>, Mbt<n+l>, . . . Mbt<n> is a maximally R-interrelated set of mental stages?

In other words, cannot two bodies share, over different intervals, a single personality? Spelling this suggestion out yields the following analysis

(MST) A mental-state-type is a complete psychologically consistent set of mental characteristics which could be had at a person at a time.

(MLT) A mental-life-type is a maximally R-interrelated set of mental state types.

(HMLT) S has mental life type \( \psi \) iff S's mental stages are of types Mst<1>, Mst<2>, . . . Mst<n> and these constitute \( \psi \);

(MLTS) Mental-life-type \( \psi \); survives event e iff some individual has \( \psi \) up to e, and some individual has \( \psi \) after e.
(PMLTS) S's mental-life-type survives event e iff S has a mental life of type ψ up to e and some individual has ψ after e.

Now, what is wrong with equating survival with survival of mental-life-type? First, one may object to the abstract and merely possible entities involved; but, as I indicated earlier, it is my assumption that these are merely terms of art, to be eliminated in terms of persons. So I will disregard this objection until it has been shown that abstracta are not eliminable and that abstract terms must be taken as referring to existing entities.

Still this analysis is not without its problems. For cases can be imagined which meet our criteria of the survival of mental life which would not satisfy at least some cogent intuitions about survival). Consider the case of Simon and his mental doppleganger, Simon-Z, an exact mental duplicate of Simon living on, say a 'twin' of Earth. Imagine that in their shared mental lives, each comes to believe that he has a mental duplicate; and each comes to believe himself to be, at t, in danger of being killed at t+n. Would Simon be comforted by someone's telling him that Simon-2's safety was insured, and that thus his mental life would flow on, since they have the same mental life? I think not, though my intuitions boggle just a bit here.
One who thinks that the continuation of mental life is both necessary and sufficient for survival should say that if Simon is killed, i.e. suffers bodily death, and Simon 2 lives on, then Simon has survived as Simon-2. But I doubt that anyone put in Simon's shoes would be much comforted by this claim.

On the other hand, if such mental duplication were commonplace, perhaps the continuation of one's mental life in a body completely distinct from (i.e. sharing no parts with) one's own would count as survival, at least in some cases. Imagine, for instance, a society within which the technique of cloning is perfected, and in which each person has several clone duplicates kept in a device which causes the duplicate to develop to physical maturity without becoming conscious or having psychological states. Then when the person has worn out his original body, a machine transfers his mental characteristics to a 'new' body -- one of his clones -- and destroys the old body. Would this not count as survival?

The cases differ in at least one important respect. In the latter, though not the former, there is a causal explanation of how the survivor came to have the mental states of the person whose body was destroyed.
I find these examples less than decisive at least in the absence of a more complete theory of mind. They seem to suggest that mental continuity and connectedness is not sufficient for survival unless the connectedness is physically, i.e. causally, mediated.

On the other hand, if survival is viewed as persistence of personality -- a phrase found in modern theological accounts -- then Simon could be said to survive in Simon-Z. For the present I will distinguish between two senses of survival; first, the survival of personality, or of a mental life, not requiring a causal link between 'before and after' stages in a mental life. (Simon's personality survives in this sense). Second, personal survival which is like survival of personality except that it does require at least an appropriate form of causal mediation of before and after stages. In both cases, the personality survives: that is, there is a mental life as discussed above including the mental states of a person before an event, and including mental states of a person after that event; but the persons are different.

According to the previous analysis of the survival of mental life, what is important is continuation of a mental life, that is, the continuation of a personality, of a set of mental-event types having the
R-relation. If at a given moment, all of a person's mental states, occurrent and dispositional, were transferred to a body exactly similar -- at least as far as the individual in question would notice -- to his own and the original body was destroyed, the mental life would, by hypothesis, flow on, and his sudden shift in location would be regarded no doubt as somewhat illusory (unless, of course, this transfer of personality were made known to him).

What we have in this case is a personality surviving though not in the same body. Brain transplants present similar cases; only a part of the original body survives but the personality, or mental life, flows on -- assuming, of course, successful surgery and that the transplant operation does not so shock the set of mental states associated with that brain that a breakdown of the R-relation occurs. Such shifts of personality from one body to another have been central to several works of fiction, notably A.E. VanVogt's *The World of Null-A*, in which the protagonist is transferred to a duplicate of his body over a great distance at the moment 'he' (the original body) was being killed, and Daniel Dennett's story "Where Am I", in *Brainstorms*.

If we accept the view that the mental life has continued in such cases, the characteristic functions
of a person having been transferred from one body to another by surgery, magic, or a star-trekish transporter, must we say also that the person has survived, or that the individual having, i.e. instantiating, the mental life before the transfer is identical to the one having it afterwards?

This question must be broken up a bit before it can be satisfactorily answered. If some sort of personality transfer, whether by teleportation or surgery, becomes routine, it might become convenient to alter our ways of speaking, and thinking, about persons. For the present, however, the distinction between a person and his mental-life-type, or personality is already enshrined in the common sense picture of the world; and providing distinct analyses of their identity conditions allows us to leave open the question of whether bodily identity is a necessary condition for personal identity. Ordinary usage suggests that:

Since the bodies in the cases presented are, ex hypothesi, non-identical, and bodily identity is ordinarily considered necessary to personal identity, perhaps it would be agreed that the person did not survive, but that his most important 'aspect,' his personality did.
Notice that the prospect of the transfer of personality raises important moral questions; these are multiplied when the prospects of fission, fusion, and radical change of personality are considered. In fission, for instance, it makes sense to say that Joe, Lefty and Righty share a personality, that existing up to the operation. Assuming that Lefty and Righty go their separate ways -- mentally, that is, having different thoughts, and perceptions, acquiring different beliefs -- then they have distinct personalities which shared some parts. The recent analysis suggests that, three persons are involved. Joe's personality, though not Joe -- his original body having been destroyed -- survives. Fusion can be handled similarly.

The problems that plagued Parfit and Lewis have thus evaporated by distinguishing the concept of identity of personality -- really a form of genidentity -- from that of the identity of persons. This distinction can now be used in treating other problematic cases. In what follows I will mention the problems and suggest solutions suggested by the new perspectives on personal identity. Finally, I will suggest that there is nothing really new about this view.

First, consider cases of radical change of personality. There are a number of situations wherein
we are tempted to say of an individual "he's not the same person any more"; we speak this way of persons who have had lobotomies or other forms of radical neurosurgery, persons who have been brainwashed, and persons who have undergone certain (often religious) conversions and who have come to view their former selves as somehow alien, and have come to view earlier experiences as having different import than their former selves had. If my suggestion is right, we can be speaking the literal truth; for if an individual acquires a new set of dispositions, if there is a severe enough change of the individual's beliefs, hopes, desires, and other 'propositional attitudes,' then the R-relation can be broken. The individual can come to have such different mental states and dispositions that his new set, or stage, is no longer connected and/or continuous with his previous stages. In such a case, the individual has a new personality; and if sameness of personality is necessary for sameness of person, the individual has become a new person -- or in cases of severe neuro-physiological dysfunction, as following radical lobotomy or brain damage, the individual can lose all personality, and become a non-person.

Similarly, suppose Jones and Smith step into a machine and come out with their personalities switched
(or that they undergo a brain-exchange operation with similar results). Using 'survival' terminology, we would say that Jones has survived in Smith's body and vice-versa. But now that these personalities inhabit new bodies, what do we say of the persons, Jones and Smith? I see nothing wrong with saying that they have ceased to exist, but that new persons, one with Jones' body and Smith's personality, have come into being. This possibility raises severe moral -- and legal -- problems; for who is to be punished for transgressions committed by the old Smith ('Smith in Smith's body')? The obvious suggestion is that culpability is a feature of personality, not of the body.

Cases of fission and fusion make this moral problem more obvious and harder to solve. For if Joe robs a bank and then becomes Lefty and Righty, who is to be blamed? Both? (But neither Lefty nor Righty, qua persons, existed at the time of the robbery.) Neither? (But both Lefty and Righty have Joe's personality, at least for a while, following the operation.)

What then can be said of disembodied survival? If the personality is what is said to survive, and is a type, or genus, rather than a particular, the question of disembodied survival rests on the question of the
existence of abstract objects. But notice that even if, per
impossible, types could exist without any tokens of
the type existing, the disembodied personality is not
a person. It is a temporally ordered set of sets of
belief-types, etc., and like all abstracts, is incapable
of action or perception. This squares with Strawson's
account of disembodied survival, as should be
expected.

On the other hand, if enough of a person's
thoughts have been absorbed by his fellows, preserved
on paper or in the actions of others, we tend to say --
loosely -- that his 'spirit' has survived. This view,
that a person can 'live on' in his effects on the world,
is often called 'pagan' immortality. In order for the
present view to validate the theory of pagan immortality,
it must be the case that a sufficient portion of a
persons personality, his mental state-types, be
instantiated to form a set having the R-relation to
members of the set of personality-stages had by the
living person. But if sensory as well as cognitive
states are included in the mental states making up the
stages, this would seem practically impossible. Yet
if we construe a person's 'spirit' as a some how essen-
tial fragment of his personality, then perhaps enough
truth can be found in the notion of pagan immortality
to satisfy romantics, poets, theists, and others who
insist that there is some literal truth in their spiritualistic claims. It should be noted that 'pagan immortality' is the name theists have given this view, indicating its heretical nature by use of the epithet 'pagan.'

Lest one think that this acknowledgement of the possibility of pagan immortality supports any doctrine of supernatural disembodied survival, let me hasten to remind the reader that, strictly speaking, pagan immortality is not a doctrine of disembodied survival at all; rather it is a doctrine which holds that one's personality may be 'embodied' in non-human—even non-biological-form. The catch phrase of the doctrine is, after all, 'He lives on in his (existing) works.' In move from this to a doctrine of really disembodied survival would require at least a strong form of platonism, holding that abstracta exist independent of concrete entities. But even Platonism is not sufficient to insure the individuation of disembodied survivors, given the possibility of two persons with the same mental life. Nothing less than a full fledged theistic or Cartesian doctrine of the soul can guarantee genuine disembodied survival.
Conclusion

By distinguishing two forms of identity, genuine individual identity and gen-identity, or type-identity, and distinguishing personalities, i.e. mental lives as 'types' (without assuming that types can exist independent of their tokens) from persons we have, it seems arrived at a position which has the virtue of being consistent with intuition in ordinary cases, and in most of the contrived cases we have considered survival of personality is a matter of identity—or, rather, genidentity. Continuity and connectedness are, importantly, the criteria for determining if personality-stages are parts of a personality. Personal survival requires physical, or at least causal, connectedness and continuity on a par with that of the continued identity of material objects in addition to the survival of personality; and for all practical purposes, bodily identity is necessary and, — except in cases of radical personality change or cessation resulting from, say brainwashing or lobotomy — sufficient as a criterion of personal identity, and hence of personal survival in ordinary cases. We are left with puzzles regarding moral culpability in cases of fission; some may regard these problems as indicating unsolved problems for the
present analysis; but I, for one, regard them as indicators that something is wrong with retirbutivist accounts of moral culpability. And that is surely a story for a later occasion.
Aune, Bruce; Knowledge, Mind, and Nature, Random House (NY, 1967)


Dennett, Daniel; "Intensional Systems" in Brainstorms.


EPH (Sellars, Essays in Phil. and its History)

EMH (Sellars, Empiricism & the Philosophy of Mind)


Fodor, Jerrold; The Language of Thought, Crowell (NY, 1975).


Harman, Gilbert, and Donald Davidson, eds. Semantics of Natural Language, Reidel (Dordrecht, 1972)


Kripke, Saul; "Identity and Necessity," in Munitz (1973).

Kripke, Saul; "Naming and Necessity," in Harman & Davidson (1972).

Lewis, David; "Survival and Identity" in Rorty (1976).


Lewis, David; "General Semantics" in Harmon and Davidson (1972).

Lewis, David; "Counterpart Theory and Quantified Mode of Logic" Journal of Philosophy LXV, 5 March 1968.


Montague, Richard; "Pragmatics and Intensional Logic" Harmon & Davidson (1972).


Nagel, Ernest; The Structure of Science, Harcourt (New York, 1961).

Parfit, Derek; "Personal Identity" Philosophical Review 80, 1971 pp. 3-27.

Parfit, Derek; "Lewis, Perry and What Matters" in Rorty (1976).

Perry, John; "The Importance of Being Identical", in Rorty (1976).


Sellars, Wilfrid (1965), "Scientific Realism or Irenic Instrumentalism" in Sellars (1976b)


Sellars, Wilfrid; "Behaviorism, Language and Meaning" in *Pacific Philosophical Quarterly*. 