

1-1-1990

Semantic properties and the computational model of mind.

Randall K. Campbell
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Campbell, Randall K., "Semantic properties and the computational model of mind." (1990). *Doctoral Dissertations 1896 - February 2014*. 2062.

https://scholarworks.umass.edu/dissertations_1/2062

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST



312066013576105

SEMANTIC PROPERTIES AND THE COMPUTATIONAL MODEL OF MIND

A Dissertation Presented

by

RANDALL K. CAMPBELL

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 1990

Department of Philosophy

© Copyright by Randall K. Campbell 1990

All Rights Reserved

SEMANTIC PROPERTIES AND THE COMPUTATIONAL MODEL OF MIND

A Dissertation Presented

by

RANDALL K. CAMPBELL

Approved as to style and content by:



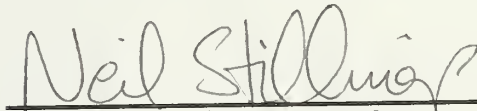
Gareth Matthews, Chair




Bruce Aune, Member



Fred Feldman, Member



Neil Stillings, Member


John G. Robison, Head
Philosophy

ABSTRACT

SEMANTIC PROPERTIES AND THE COMPUTATIONAL MODEL OF MIND

SEPTEMBER 1990

RANDALL K. CAMPBELL, B. A., WILLAMETTE UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Gareth Matthews

Much of the contemporary research in cognitive psychology presupposes an information processing or computational model of human cognitive processes. On this view cognitive states are characterized as relations to internally inscribed representations. Jerry Fodor and Zenon Pylyshyn have argued that those representations have a combinatorial syntax and a compositional semantics, and Fodor has argued that the individuation of representations according to semantic type corresponds, roughly, to individuation according to syntactic type.

I investigate whether this computational model requires us to appeal, directly or indirectly, to the semantic properties of representations when we explain cognitive behavior. I first discuss the requirements of scientific explanation in general, and the constraints of "materialism" and "physicalism" in particular. Then I outline how it is possible for semantic entities to be involved in cognitive explanations, and how Fodor and Pylyshyn think they are involved in explanations on the computational model. I consider whether, given the computational model, references to representations are necessary to explain cognitive processes or whether references to representations can be eliminated in favor of references to uninterpreted formulae. Finally I criticize the argument, suggested by both Fodor and Pylyshyn, that it is our ability to respond to nonnomic or nonprojectable properties of stimuli that requires explanation in terms of the semantic properties of representations.

TABLE OF CONTENTS

	Page
ABSTRACT	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
Chapter	
1. INTRODUCING THE PROJECT	1
1.1 A Question	1
1.2 The Question	6
1.3 What I Do Not Intend to Do	8
1.4 What I Do Intend to Do	10
2. THE RUDIMENTS OF SCIENTIFIC EXPLANATION	13
2.1 Introduction	13
2.2 Explaining the Occurrence of Events	14
2.2.1 Causal Explanations	15
2.2.2 Statistical Laws and Explanation	19
2.3 Explaining Property Instantiation	23
2.3.1 Property Analysis	24
2.3.2 Interpretive Analysis	28
3. PHYSICALISM, MATERIALISM AND THE NON-PHYSICAL SCIENCES	38
3.1 Introduction	38
3.2 Physicalism and Materialism	39
3.3 The Loss of Generalizations Argument	45
3.3.1 The Argument	46
3.3.2 Some Observations	51
3.3.3 More Observations	55

4.	FUNCTIONAL THEORIES AND INTERPRETATION	61
4.1	Introduction	61
4.2	Some Preliminaries	61
4.3	Explaining Dispositions	64
4.4	Functional Theories	68
4.5	Interpretation	74
	4.5.1 Interpretive Functional Theories	76
	4.5.2 A Short Defense	77
5.	THE CLASSICAL COMPUTATIONAL THEORY OF MIND	82
5.1	Cognitive Architecture	82
5.2	The Computational Theory of Mind	83
	5.2.1 Some Motivation	83
	5.2.2 The Theory	86
	5.2.3 Some Consequences	97
5.3	The Classical Computational Theory of Mind	97
	5.3.1 The Theory	98
	5.3.2 Some Evidence	100
6.	THE CLASSICAL COMPUTATIONAL THEORY OF MIND AND STATE-STATE LAWS	105
6.1	Introduction	105
6.2	The Syntactic Theory of Mind	105
6.3	A Problem	109
6.4	Pylyshyn's Defense	112
7.	STIMULUS-STATE AND STATE-BEHAVIOR LAWS	117
7.1	Introduction	117
7.2	Churchland's Argument	119
7.3	The Argument from Stimulus Independence	123
7.4	Nonnomic Properties of Stimuli and Representation	125
	7.4.1 The Argument	125
	7.4.2 Why Nonnomic Properties?	130
	7.4.3 Circularity	133
	7.4.4 More Mystery	138
	7.4.5 Another Try	140

8.	ANOTHER COMPUTER EXAMPLE	145
8.1	Introduction	145
8.2	An Old-Saw Computer Example	146
8.3	Why Interpretation?	151
	BIBLIOGRAPHY	153

LIST OF TABLES

Table	Page
4.1 The input-output laws for circuit C1	66
4.2 The input-output laws for C1's subcircuits	67
4.3 The inputs and outputs of C1 - C4 interpreted as sentences and compound sentences	75
4.4 The inputs and outputs of C1 - C4 interpreted as binary numbers	76

LIST OF FIGURES

Figure		Page
4.1	A vacuum tube triode	65
4.2	A block diagram of the circuit C1 (with batteries, resistors, etc., omitted)	67
4.3	A diagram of an INVERTER circuit (a) and an AND circuit (b)	68
4.4	A block diagram of C4	70

CHAPTER 1 OUTLINING THE PROJECT

1.1 A Question

The demise of behaviorism once again permits psychological theories to refer to the internal states of their subjects. Of course, people have always talked in terms of "beliefs," "desires," "fears" and other notions that seem to presuppose that people have "representational states." Now representations and representational states are common currency in cognitive psychology. Recently there has been a running battle in philosophy and the other branches of cognitive science about the answer to this question:

- [1] Do any psychological theories that appeal to "representational states" provide adequate explanations of human behavior?

In the philosophical literature the dispute has centered on whether what Stephen Stich has perjoratively labeled "folk psychology" can provide adequate explanations of human behavior. This folk psychology is, more or less, our everyday belief/desire "psychology," where representational states are identified by locutions such as 'believes that,' 'desires that,' 'fears that,' and 'hopes that,' and the content clauses that follow them. Stich [1983], among others, has argued that individuation according to content clauses provides a taxonomy of psychological states that is sometimes too fine-grained and sometimes too coarse-grained to provide adequate explanations. Others have worried about whether talk in terms of beliefs and desires can be made "materialistically" or "physicalistically" respectable. These issues have dominated the debate over whether our everyday belief/desire psychology can be reconciled with explanatory cognitive psychology. That debate has, perhaps unfortunately, dominated the discussion of what roles references to representational states play in psychological explanation.

I say 'perhaps unfortunately' simply because the traditional belief/desire psychology is not the only sort of theory to refer to representational states. While many philosophers seem wedded to belief/desire psychology, and argue either that

cognitive explanations must appeal to representational states as they are identified by belief/desire locutions, or that cognitive explanations do not contain any essential reference to representational states individuated in that way, cognitive psychologists have produced a great deal of work referring to "representations" and "representational states" without ever mentioning "beliefs" or "desires." There are many different ways in which references to representational states might be used in psychology. Answering [1] without considering all of those ways seems premature. In the end, whether we discover explanatory psychological theories that refer to representational states may depend solely on the ingenuity of psychologists, so answering [1] may be a purely empirical matter.

While that dispute continues, a similar question has received less attention in the philosophical literature:

- [2] Is it necessary to appeal to "representational states" in order to explain human behavior?

Before dissatisfaction with traditional belief/desire psychology, or dissatisfaction with any other theories, leads us to give up on psychological theories that appeal to representational states (which, for the purposes of the introduction, I will call 'representational theories'), we had better answer [2]. The answer is of some theoretical interest and practical concern to psychologists, who are, after all, trying to explain behavior. But, even though at first this question seems like a matter for science to decide, it may not be possible to answer the question empirically. We certainly might get some idea of the answer by judging the relative predictive success of representational and nonrepresentational theories. But even if the representational theories are more successful than nonrepresentational theories, that would not necessarily be an indictment of nonrepresentational theories; instead, it might be an indictment of our ability to construct nonrepresentational theories. There might be nonrepresentational theories that could predict and explain any behavior that representational theories could predict, even if we never think of them.

The purpose of this dissertation is to start toward an answer to [2]. We can appreciate the difficulty of answering [2] by first considering part of the galaxy of

problems associated with figuring out what the question asks. For example, what states count as representational states? They are purportedly identified by 'believes that' and 'desires that' locutions and they are naturally characterized as relations to *something*. The contenders for what those things are include specific sorts of ordered sets or sets of possible worlds--i.e., "propositions"--and types of sentences, tokens of sentences in our heads, types of images, or tokens of images in our heads--i.e., "representations." Presumably, we stand in the belief relation and the desire relation to these propositions or representations because of our other properties--our neural states, our relationships with the outside world, etc.--so there is not necessarily anything mysterious about these relations. Also, since we stand in many different relations to these sorts of things, there may be explanatorily valuable representational states other than those identified by the 'believes that' and 'desires that' locutions.

If representational states are characterized as (specific kinds of) relations to propositions or representations, we could decide whether a given theory appealed to representational states merely by determining whether people had to bear those relations to those things if the theory were true. But sometimes representational states are not characterized that way. The "orthographic accident" view is the view that when we say someone 'believes that such and such' or 'desires that such and such,' 'believes that' and 'desires that' do not express relations at all: The occurrences of 'believes' in the sentences 'Randy believes that Fred is tall' and 'Randy believes that Fred will win the lottery' do not refer to any sort of relation; the similarity between the sentences is just an accident, and it is not necessary that the states identified by the two sentences are any more similar to each other than they are to states identified by sentences employing '...desires that...' and '...fears that...' locutions. It will be hard to answer [2] if we have to consider every sort of view about representational states to determine whether appeal to those sorts of state is necessary in explanations of human behavior.

Even if we confine ourselves to the relational view, we still have to figure out (1) what relation we stand in to propositions or representations when we have particular representational states, and (2) what propositions or representations are.

(1) Sometimes representational states are identified with "intentional states." Given that identification, [2] would be the same question as

[3] Is it necessary to appeal to intentional states in order to explain human behavior?

Intentional states are often characterized as special sorts of relations to propositions or representations. For example, some people seem to assume that if someone is in a particular intentional state, then he or she is or is capable of "entertaining" that proposition or representation in some way--that he or she is capable of being "conscious" of the content of that representational state. But we might want to exploit certain relations between things and propositions or representations in order to explain the behavior of those things without supposing that those things are conscious or that they are "entertaining" thoughts. When I play chess with a computer I usually reason as if it were a human opponent, and I might think *It thought that I would try to double my rooks*. In effect I have "explained" its behavior by attributing an intentional state to it--the state of thinking *that Randy will try to double his rooks*. Of course I don't believe that the computer is actually in an intentional state, because I do not believe that the computer is actually "thinking" that way; most of us do not believe computers actually "understand" what they are doing, and so we do not believe they have intentional states. Since we can provide some sort of explanation of a computer's behavior without actually appealing to intentional states, we might wonder whether we must appeal to these sorts of intentional states in order to explain human behavior--that is what [3] asks.

Perhaps the relation in question is less exotic. But if representational states are not identified with intentional states, then what sorts of relations to propositions or representations count as representational states? Surely my being within two feet of a token of a meaningful sentence does not count as a representational state. So what sorts of relations count, and which do not? We might avoid the difficulty by making the question a little broader. When I "explained" my computer's behavior by attributing certain "thoughts" to it, I did not

actually believe that it was thinking. But by supposing that it stood in certain relations to "representations" I was able to "explain" its "moves." Even if my computer does not actually think, it might be necessary to suppose that certain relations hold between it and some propositions or representations in order to explain its capacity to play chess. We might ask a similar question about people:

- [4] Is it necessary to attribute to people specific relations to propositions or representations in order to explain human behavior?

This question is sometimes asked as a question about whether it is necessary to provide a semantic interpretation of psychological processes in order to explain human behavior.

(2) A minimal requirement for propositions and representations is that they have semantic properties such as *being true*, *being false*, *meaning that...*, etc. I'm not sure exactly what sorts of properties count as semantic properties, but I think that we probably all have some idea of what the beginning of a list would look like. Many sentences are true or false and are meaningful, and so they seem to be obvious examples of representations. Propositions are apparently either true or false and they are meaningful, even if they are "creatures of darkness."

But this minimal requirement that propositions and representations have semantic properties introduces some difficulties, for, even apart from saying what all the different sorts of semantic properties are, it is a philosophical problem whether things have semantic properties because they are interpreted in certain ways, or whether they have semantic properties because they *could* be interpreted in certain ways. For example, consider the string of symbols: 'v[[T<∃h]]SεθΔ)'. I doubt if this "means" anything to anyone. But we might develop a language in which it does have some sort of "meaning." If we admit that anything that can be interpreted in a particular way is a representation, then everything will be a representation of some sort. We could, for example, develop a language in which coke machines of various sorts are adopted as singular terms.

We might be able to avoid these problems by making the question still broader. When we appeal to representational states in explanations of behavior we refer directly to things that have semantic properties, though perhaps not to the

semantic properties themselves. But if those sorts of things are individuated according to their semantic properties, as sets of synonymous sentences would be, then our explanations seem to depend on the existence of semantic properties, even if they do not refer directly to semantic properties. If we call this sort of commitment an (indirect) appeal to semantic properties, then, instead of asking whether it is necessary to appeal to representational states in explanations of human behavior, we can just ask whether we have to appeal to semantic properties in explanations of human behavior. We might, then, consider this question:

[Q] Is it necessary to appeal to semantic properties in order to provide adequate explanations of human behavior?

If we could answer [Q], we would be on our way to answering some other important questions. For example, if the answer to [Q] is 'no,' then the answer to [3] is 'no': if it is not necessary to appeal to semantic properties in order to explain human behavior, then it must not be necessary to suppose that we stand in any sort of "intentional" relation to propositions or representations in order to explain our behavior. If the answer to [Q] is 'no,' then the answer to [4] is also 'no.' [Q] is more general than [4], since [Q] asks whether any semantic properties need to be appealed to in psychological explanations, not just whether specific relations to propositions or representations must be ascribed to people to explain human behavior. However, since it is hard to see how semantic properties would enter into explanations of behavior other than as properties of propositions or representations, we might as well consider [Q] and [4] the same question.

1.2 The Question

[Q] is a very broad question. To answer it satisfactorily would require an accurate characterization of representational theories. I am not prepared to provide that characterization: there might be a list of the essential properties of those theories, but even if I possessed it, I doubt that many cognitive scientists would agree that it was correct. Answering [Q] would also require some account of how references to semantic properties are involved in explanations of behavior. Though it may seem obvious what semantic properties are, some people have

characterized representational states as relations to sorts of things that just do not seem to have semantic properties. For example, there is a long history of characterizing belief states as particular relations to propositions, where the propositions are characterized as particular sorts of sets or ordered sets. Talking about sets of possible worlds may be useful in a lot of ways, but sets of possible worlds do not seem to have any semantic properties. Of course, various relations between sets are obviously similar to semantic relations--if one set of possible worlds is a subset of another, then we might think of the first set as being entailed by the second, but that is not a good reason to think that those relations are identical. The answer may be that it is the special nature of the believing relation that confers semantic status on these particular sorts of sets: perhaps the sets are meaningful because they are "apprehended" or "entertained" in some particular way. Since "representational states" do not wear their semantic properties on their sleeves, it will be difficult to figure out exactly what role semantic properties play in explanations of behavior.

Since I am not prepared to provide a list of the characteristic properties of representational theories or of the exact role references to semantic properties play in those theories, I will adopt a view about those issues and try to answer [Q] given that view. The view that I take to be the dominant view in the philosophy of cognition has been referred to under a variety of labels almost all of which contain the word 'computational.' Jerry Fodor and Zenon Pylyshyn are the two philosophers who have done the most to define and defend a version of that view, and it is their (largely shared) view which I will address under the moniker 'The Classical Computational Theory of Mind' (CCTM). Thus, this is the question that I will try to answer:

[Q*] If the Classical Computational Theory of Mind is correct, is it necessary to appeal to semantic properties in order to provide adequate explanations of human behavior?

Although both Fodor and Pylyshyn argue that the answer to this question is 'yes', I am assuming that the answer is not *necessarily* 'yes'--in other words, I am assuming that their claim is not part of CCTM.

I should be clear about what sort of theory CCTM is. It is not a theory about human behavior, nor is it (primarily) a view about the correct philosophical analysis of typical "mentalist" predicates, such as 'believes that....,' 'desires that....,' etc. In his most substantial work dedicated to defining CCTM, Fodor [1975] claims to be doing neither experimental psychology nor philosophical analysis. Instead, he claims to be doing what used to be called 'speculative psychology.' According to Fodor, speculative psychologists 'thought about such data as were available about mental processes, and they thought about such first-order psychological theories as had been proposed to account for the data. They then tried to elucidate the general conception of the mind that was implicit in the data and in the theories'[1975, p. viii]. This description seems to fit the behavior of many of those who argue about the "cognitive architecture" of psychological processes. CCTM thus comprises a conception of the mind, or a view about the architecture of cognitive processes.

Unfortunately, Fodor and Pylyshyn are themselves not always forthcoming about how semantic properties are involved in explanations of behavior, so the problem is not exactly clear cut. However, we may be able to come from the other direction by determining when only genuinely nonsemantic properties are involved in particular sorts of CCTM-style explanations. For example, as Fodor [1980] pointed out, being syntactic is a way of not being semantic. If a psychological theory refers to representational states, and so to things that might be semantic, we can try to figure out if the explanations the theory provides could be replaced by explanations that refer to only the "syntactic" properties of those states.

1.3 What I Do Not Intend to Do

The conditional form of [Q*] will limit the sorts of questions that I ask and answer. Since the Classical Computational Theory of Mind is a view about the "cognitive architecture" of cognitive processes, CCTM entails some things about what good cognitive theories will look like and suggests methodological claims about what cognitive psychologists should be trying to do. Of course, CCTM has been criticised on both of these counts. For example, there is the (too) often stated claim that "computationalism" has not produced any "successful" theories. Robert Cummins [1988] suggests that this sort of criticism comes from non-psychologists

who either have their sights set unrealistically high or who simply do not know what is going on in cognitive psychology. There are, I am sure, more worthy criticisms of the ability of CCTM style theories to provide explanations of cognitive behavior, but, since I do not want to reveal my ignorance, I will leave those problems to Fodor and Pylyshyn, to their critics among the cognitive psychologists, and to the philosophers [Paul Churchland, 1981, Patricia Churchland, 1986, and Stich, 1983] who know more about psychology than I do. I am not a psychologist and I do not intend to do any psychology.

There are, of course, more "philosophical" criticisms of CCTM--for example, that CCTM explains nothing about consciousness or intentionality [Dreyfus and Dreyfus, forthcoming, and Searle, 1980]; that CCTM cannot be true given the correct analyses of mentalistic language; and that the notion of "content" that CCTM entails cannot be made "physicalistically respectable" [Baker, 1987]. I intend to avoid those issues too. For one thing, those problems are extremely difficult. But more importantly, I think that the impulse to criticize CCTM, whatever the motive, has led some people to overlook the question 'just what role do semantic properties play in CCTM-style theories?' and instead concentrate on showing that, whatever their role is, CCTM is not right. CCTM still seems to be the dominant (philosophical) view about cognitive architecture, and Fodor and Pylyshyn are the most visible defenders of representational theories (in the philosophical ranks), so the question seems to be worth asking.

My project might be put into perspective by an analogy with the logicist project of Frege and Russell. They tried to show that classical mathematics was just logic in disguise, and that mathematical statements could be translated into (much more complicated) statements in a purely "logical" vocabulary. They certainly had no intention of either defending or attacking any of classical mathematics. They just wanted to figure out whether mathematics could do away with the somewhat mysterious notions of "zero," "number," and "successor." I have no intention of either defending or attacking CCTM, but I am concerned with discovering exactly what role semantic properties play in explanations produced by CCTM-style theories and whether that role is essential to cognitive explanation.

1.4 What I Do Intend to Do

To answer [Q*] we first have to know a little bit about scientific explanation. The first task, and the aim of chapter 2, is to provide a general account of what scientific explanations are like. Of course, I will not be concerned with describing scientific procedures and methodology, or the reasoning of individual scientists. Instead I will be trying to provide a sketch of the conditions successful explanations satisfy. Scientific explanations come in roughly two forms: Explanations of events through subsuming them under a causal or statistical law, and explanations of properties/capacities through property analysis. The explanation of events is by far the most familiar of these two types of explanation, and the elements of my account will be culled from several well received accounts. Until relatively recently, however, it has not been emphasized that many scientific explanations do not fit the pattern of event-explanation. Robert Cummins [1975 and 1983] has introduced the term 'functional analysis' for explanations of how certain sorts of properties/capacities are instantiated in systems. Cummins is not the only person to recognize the need for functional analysis, but he has made the most determined effort to say just what functional analyses look like. My account of functional analysis will be largely parasitic on Cummins' account.

Since many, if not most, philosophers and scientists will not tolerate an "explanation" that is neither "physicalistic" or "materialistic," my task in chapter 3 is to discuss the constraints of physicalism and materialism on scientific explanations. Terms like 'materialism' and 'physicalism' have a lot of associated baggage, and often they are defined in ways that make them extremely unattractive. Apparently physicalists and materialists believe that *in some sense* all of science can be "reduced" to physics. Of course, they often disagree in what sense it is that the reduction can be achieved. I will try to provide a formulation of "physicalism/materialism" that is widely shared through the scientific and philosophical community.

Adopting physicalism/materialism has some consequences for what is considered an adequate psychological theory: many people seem to think that it entails that everything that can be explained can be explained in the vocabulary of physics. Since many philosophers who are committed to this kind of

physicalism/materialism also claim that physics alone cannot explain everything worth explaining, and that we have to appeal to representational states to provide adequate explanations of human behavior, it is of some interest to see how the following two claims can be compatible: (i) physicalism/materialism is true, and (ii) appeal to semantic properties is *necessary* in adequate explanations of human behavior. The most familiar argument to that effect is what I call the 'loss of generalizations' argument. It has been presented in various forms by Donald Davidson [1970], Daniel Dennett [1978a], Hilary Putnam [1973], Zenon Pylyshyn [1984], and received its best known presentation by Fodor [1974]. I will devote the last part of chapter 3 to that argument.

In chapter 4 I will be concerned with saying a little more about what explanations of complex systems are (supposed to be) like. If the "loss of generalizations" argument is correct, then there must be different "levels of explanation," depending on the phenomena or capacity being explained. So it is important to see how interpretive analysis--a type of functional analysis, and the typical explanatory strategy in cognitive explanation--provides us with these different levels of explanation and captures the relevant generalizations.

Fodor and Pylyshyn have defended, both independently and together, the Classical Computational Theory of Mind as a theory of cognitive architecture, and they have claimed that it is the dominant view in cognitive psychology. Roughly, the view is that human psychological processes (or at least the ones involved in the execution of cognitive capacities) are "computational" processes, and that human cognitive states are computationally related. These cognitive states are "representational" in a specific sense. Chapter 5 will be devoted to presenting the Classical Computational Theory of Mind and some of the motivations for it.

Stephen Stich [1983] contends that there could be no adequate cognitive explanations if cognitive states are individuated according to their semantic properties in the way required by CCTM. However, Stich suggests that we could give adequate explanations by appealing only to the "syntactic" properties or "formal" relations between cognitive states. In chapter 6, I will try to show that *if* several of the main claims of CCTM are correct, *then* we could give adequate explanations of the relations between cognitive states by appealing only to the "syntactic"

properties or "formal" relations between cognitive states. I will also discuss considerations that Pylyshyn seems to think count against Stich's view, and, perhaps, against my claim.

Even if it is possible to explain the relationships between cognitive states without referring to the semantic properties of those states, it may not be possible to express generalizations linking those states to stimuli and behavior without referring to the semantic properties of those states. In chapter 7, I will discuss an argument for the claim that generalizations linking stimuli and behavior to cognitive states are best framed in terms of the semantic properties of those states. I will consider some reasons to believe that generalizations of that sort are not available, and I will discuss Fodor's [1986] and Pylyshyn's [1985] contention that it is because these sorts of generalizations are unavailable that we must explain human behavior in terms of representational states.

Finally, in chapter 8, I will consider whether an adequate explanation of cognitive behavior must include whatever information it is that we learn when we attribute semantic "content" to internal states.

CHAPTER 2

THE RUDIMENTS OF SCIENTIFIC EXPLANATION

2.1 Introduction

There is not much readily available evidence about the nature of psychological processes. "Direct observation" is the primary source of empirical information, and introspection seems to be the most direct way to "observe" psychological phenomena. However, I think it is safe to say that no one really knows how much value to put on the evidence that introspection provides, since no one really knows what to say about consciousness.

A natural suggestion is that we could observe psychological processes first-hand simply by looking at the microstructure of brains. Of course this brings up the question 'is it possible to observe directly all of the important properties of psychological processes?' For instance, there is a long tradition that claims that many important psychological processes are, in some sense, "non-material." Others argue that, since no two brains have exactly the same microstructure, just observing the microstructure of brains will not provide evidence about what makes most of us behave in significantly similar ways: things that "think" very much like us might have thinkers that are radically different from our own.

This, I think, is the natural answer to the question: If we need to appeal to properties that we cannot observe directly in order to explain whatever phenomena we want to explain, then we cannot observe directly all of the important properties of the phenomena; if we do not need to appeal to properties that we cannot observe directly in order to explain the phenomena that we want to explain, then we can observe directly all of the important properties of the phenomena. Scientists often ascribe properties to things simply because there is no way to provide a good explanation for some phenomenon without making those ascriptions. This is especially apparent in psychology, where beliefs, desires, intentions, memories, and other sorts of representational states have been attributed to people in order to explain their overt behavior. Psychologists, and the rest of us, seem to assume that if we can explain someone's behavior only by attributing a particular property to the person, then the person has that property.¹

Philosophers of psychology are fond of making claims about what sorts of properties we have to attribute to people in order to adequately explain their behavior. Of course, we cannot evaluate those claims without first understanding what adequate explanations are like. Since my goal is to answer the question 'If the Classical Computational Theory of Mind is Correct, is it necessary to appeal to semantic properties in order to provide adequate *explanations* of human behavior?', my first task is to give an (introductory) account of scientific explanation.

2.2 Explaining the Occurrence of Events

Most of the philosophical work on scientific explanation has been concerned with giving an account of explanations of the occurrence of events.² In this century, the predominant view has been that the occurrence of an event is explained by subsuming it under a "universal" law. For example, suppose we were wondering why the volume of a particular gas increased in a particular instance. We might point out that the pressure on the gas decreased, and we might refer to the "law" that gasses expand when the pressure on them decreases. "Showing" that this event had to occur consists in giving a valid deductive argument with a statement of the initial conditions--the decrease in pressure--and the law as premises, and a statement of the type of event to be explained as the conclusion. This is the *deductive-nomological* account of explanation: The occurrence of a certain event is explained by giving a valid deductive argument with statements of initial conditions and at least one universal law as premises, and a statement of the event to be explained as the conclusion. In every case where we have a successful deductive-nomological account of why an event occurred, that event was expected given the explanatory facts: if we had known what the initial conditions were before the event in question, we would have been able to predict the occurrence of that event. Events of the same type can be explained in the same way if the initial conditions were of the same type in each case.³

Most of the standard works that presuppose the deductive-nomological account of explanation claim that occurrences of events are explained by subsuming them under laws of the following form:

[1] For all x , if Px , then Qx .⁴

Here the variable ' x ' ranges over certain sorts of "systems" and ' P ' and ' Q ' are predicates which attribute properties to those systems.⁵ For example, the statement

[3] For all gases, if their temperature is n , then their volume is m ,

is of the form 'For all x , if Px , then Qx '. Statements of this form are called 'universal conditionals.' The standard scheme for explanation is supposed to be

[4] i. For all x , if Px , then Qx .
ii. Pa
iii. Therefore, Qa ,

where the first premise is a statement of the relevant law, the second premise is a statement of the initial conditions and the conclusion is a statement of the event to be explained. The schema for the prediction of events is the same, except the occurrence of Qa is in the future instead of the past.

2.2.1 Causal Explanations

Unfortunately, defenders of the deductive-nomological account often provide examples like the following explanation of a change in the volume of a gas:

[5] i. $P_1V_1 = P_2V_2$
ii. $P_1 = 2$ atmospheres, $V_1 = 1$ liter, $P_2 = 1$ atmosphere.
iii. Therefore, $V_2 = 2$ liters.

The first premise is Boyle's Law, which tells us that when the pressure on a gas is changed, the original pressure on the gas, P_1 , multiplied by the original volume, V_1 , equals the new pressure on the gas, P_2 , multiplied by the new volume, V_2 . The second premise states the original volume and pressure on the gas, and that the pressure was halved. From that the new volume of the gas can be deduced.

However, [5] is not of the same form as [4], because the statement of Boyle's Law is not a universal conditional. Boyle's Law is actually stated as an equation that

allows us to derive (i) the volume of any gas, given its original volume, the original pressure on the gas and the current pressure on the gas, and (ii) the pressure on any gas, given its original volume, the original pressure on the gas and the current volume of the gas, as well as (iii) the original volume of any gas, given the original pressure on the gas, its current volume, and the present pressure on the gas, and (iv) the original pressure on any gas, given its original volume, its current volume, and the current pressure on the gas. Thus Boyle's Law specifies a correlation between the different values of a particular system's variables; it does not just tell us what the volume of a gas is if its original volume was one liter, the pressure on it was two atmospheres, and the current pressure on it is one atmosphere.

Charles' Law is another law that is stated in the form of an equation:

$$[6] \quad V_1/T_1 = V_2/T_2$$

It allows us to determine the change in the volume of a gas given a change in its temperature, and to determine the change in the temperature of a gas given a change in its volume; it does not *just* tell us how the volume of a certain gas changes when its temperature is doubled. Of course both Boyle's Law and Charles' Law state more than just matter-of-fact correlations. They also tell us what would be the case in counterfactual situations. For example, Charles' Law presumably tells us that if we heated so much air it would expand to fill so much space, even though we will never actually heat that much air.⁶ If they did not support counterfactual claims, equations like Boyle's Law and the Charles' Law would not be *laws*. Statements of this kind are better called *nommic correlations*, and they *are* laws, but because they are equations they are not universal conditionals; [5] and [6] are not of the same form as [4].

It may seem as if I am straining at a gnat, since the following sorts of universal conditionals can be inferred from Boyle's Law

- [7] For all gases, if their original volume is 1 liter and the original pressure on them is 2 atmospheres, and the current pressure is 1 atmosphere, then they have a volume of 2 liters.

- [8] For all gases, if their current volume is 2 liters and the current pressure on them is 1 atmosphere, and the original pressure on it was 2 atmospheres, then the original volume was 1 liter.

and the following sorts of conditionals can be inferred from Charles' Law,

- [9] For all gases, if their temperature triples, then their volume triples.
[10] For all gases, if their volume triples, then their temperature triples.

But these conditionals are *not explanatory* if they are just derived from Boyle's Law and Charles' Law. [8], for instance, would not be accepted as an explanation of why the volume of a given gas was originally 1 liter, and [10] would not be accepted as an explanation for why the temperature of a certain gas (whose volume had just tripled) had tripled in temperature. They wouldn't be accepted as explanations because, although they are true, they do not assert the right sort of connection between the events: The "reason" that a gas was originally one liter is not that the original pressure on it was 2 atmospheres and the current volume is 2 liters and current pressure on it is 1 atmosphere; the "reason" that a gas tripled in temperature is not that its volume tripled. [7] and [9] are no more explanatory than [8] and [10] if all they do is assert certain sorts of correlations.

Historically, the claim that scientific laws are just statements of correlations, and nothing more, has accompanied the claim that scientific theories are merely descriptive. If 'explanation' is to mean anything more than 'description,' then there must be additional constraints on what sorts of laws provide explanations. Requiring explanatory laws to support counterfactuals is not enough, since presumably [7]-[10] are all true when read as subjunctive conditionals. There is a grand tradition according to which only *causal laws* are explanatory. [7] and [9] have the appearance of causal laws since the 'If..., then...' locution is used to indicate causal connections between events as well as logical and subjunctive connections between statements. But if universal conditionals are derived only from nomic correlations like Boyle's and Charles' Laws, then they *only* express correlations. Thus, according to this tradition, in an explanation of why the volume of a gas is two liters, given its original volume, the original pressure on the gas and

the current pressure on the gas, an appropriate law has to generate statements of the form

- [11] For all gases, if their original volume is 1 liter and the original pressure on them is 2 atmospheres, and the current pressure is 1 atmosphere, then_c they have a volume of 2 liters,

where 'then_c' indicates that the event(s) stated in the antecedent *cause(s)* the event(s) stated in the consequent.⁷

Most of the philosophical and methodological literature concerning scientific explanation adopts the view that events are explained by subsuming them under causal laws. For example, the methodological concerns of both philosophers and scientists seem to be dominated by three different concerns about causation: (a) One major concern is the problem of distinguishing mere correlations from genuine causal connections. Statistics classes and introductory texts on "scientific method" are full of 'how to' material on the subject, and, ever since Hume, philosophers have been trying to figure out if we *can* justify distinctions between correlations and causes. (b) Another concern is with the justification of causal claims where there is little or no evidence for the occurrence of an event except that its occurrence would "explain" the occurrence of a more obvious event. This problem is notorious in every branch of the sciences: in medical research where "humors" were once referred to in explanations of illnesses, in physics where the existence of massless particles are hypothesized to explain certain phenomena, as well as in psychology where the behaviorists reacted against what they believed were unjustifiable claims about unobservable mental processes. The philosophical problem has been well known since Hume, and it has been the major point of dissatisfaction with appeals to "souls" or "non-physical minds" in explanations of behavior: claiming that souls account for human behavior is unsatisfactory if the only way we can identify souls is as the things that account for human behavior. (c) Perhaps the most widespread methodological concern is evident in the unease with which most scientists and philosophers contemplate "random" events. Prior to quantum mechanics, scientific theories that hypothesized random events were considered completely unacceptable, and there is still an obvious and widespread

desire to replace those sorts of theories with theories that do not countenance random events. Presumably this is because there is an implicit assumption that since random events cannot be explained through subsumption to causal laws, their occurrence cannot be explained at all.⁸

2.2.2 Statistical Laws and Explanation

Unfortunately, the existence of statistical laws in science muddies up the tidy picture of explanation that is provided by the deductive-nomological account. There are hundreds of examples of statistical laws. Suppose, for example, we wanted to explain why 25% of the members of a particular (static, randomly breeding) population have sickle cell anemia. We might point out that sickle-cell anemia occurs only in people who are homozygous for a particular allele at a particular gene locus. But we have not begun to explain the frequency of sickle-cell anemia until we note that the frequency of the sickle-cell allele in the population was 0.5, and refer to the Hardy-Weinberg Law which tell us that, probably, 25% of the population will be homozygous for an allele if the allele frequency is 0.5. If we had known prior to the event that the frequency of the sickle-cell allele in the population was 0.5 and if we had known of the Hardy-Weinberg Law, we would have been able to predict what portion of the population would suffer from sickle-cell anemia. To give an adequate picture of scientific explanation, it seems we have to account for explanations that appeal to laws of the form,

$$[11] \quad p(Q|P) = r,$$

which tells us that the probability (relative frequency, chance) of the occurrence of an event of type Q under conditions of type P is r.⁹

It may seem as if we might avoid having to give an account of statistical explanation by observing that in cases like the explanation of sickle-cell frequency via a statistical law, we could have explained the frequency without the statistical law: 25% of the population got sickle cell anemia because 25% were homozygous for the sickle-cell allele, and we can explain each instance through molecular genetics. Explanations that appeal to statistical laws like the Hardy-Weinberg Law are "reducible" in this sense. If every statistical law were reducible in this way we

might claim that they are convenient predictive devices, but that they do not actually provide complete explanations. This seems especially plausible given the fact that statistical laws do not "explain" why some particular P's are Q while the other P's are not. Unfortunately, quantum mechanics provides an example of a predictively successful theory with statistical laws which shows no immediate prospect of being "reduced" to a non-statistical theory. It seems as if we have to take the possibility of irreducible statistical laws seriously.¹⁰

If non-statistical explanations of events are construed as valid deductive arguments, it is natural to assume that statistical explanations are valid inductive arguments. That is the assumption of the *inductive-statistical* account of statistical explanation: The occurrence of an event can be explained by giving a valid inductive argument with statements of initial conditions and at least one statistical law as premises, and a statement of the type of event to be explained as the conclusion. Just as in deductive-nomological explanations, in every case where we have a successful inductive-statistical account of why an event occurred that event was expected given the explanatory facts: if we had known what the initial conditions were before the event in question, we would have been able to predict (with some confidence) the occurrence of that event. Events of the same type can be explained in the same way if the initial conditions were of the same type in each case.¹¹ This then is the schema for inductive-statistical explanations:

- [12] i. $p(Q|P) = r$
- ii. Pa
- iii. Therefore, Qa

If the argument is a *valid* inductive argument, then the occurrence of Qa is very probable given the occurrence of Pa, and so r should be close to 1. For example, The Hardy-Weinberg Law allows us to conclude with great confidence the frequency of homozygote-dominants, homozygote-recessives and heterozygotes at particular gene loci given the frequency of the individual alleles in a (static, randomly breeding) population.

However, there are some problems that require some additional constraints on which of these sorts of inductive arguments count as explanations. One problem,

referred to as *the ambiguity of inductive-statistical explanation*, is that by specifying different (though true) initial conditions and using different statistical laws much different probabilities can be given for the occurrence of the event being explained. For example, if we are trying to explain why our population has a 25% rate of sickle-cell anemia and we point out that, say, only 2% of all (static, randomly breeding) human populations have such a high frequency of sickle-cell anemia and that the population in question is a (static, randomly breeding) human population, we will be able to infer that it is very improbable that the population would have such a high rate of sickle-cell anemia. But if we appeal to the Hardy-Weinberg Law and the frequency of sickle-cell alleles in the population, we will be able to infer that it is very probable that the population will have a rate of sickle-cell anemia close to 25%. This problem might be avoided by adopting a *principle of maximal specificity*: in an adequate statistical explanation, the class, C, to which the individual case is referred cannot be divided up into subclasses in such a way that if the individual case were referred to one of the subclasses the probability of the event occurring would differ from its probability under C. For example, if we refer to the statistical generalization that only 2% of all (static, randomly breeding) human populations have such a high frequency of sickle-cell anemia and that the population in question is a (static, randomly breeding) human population, then we have violated the requirement, because the population belongs to a subclass of populations--those with sickle-cell allele frequencies of 0.5--for which the probability of having a sickle-cell anemia rate of 25% is not 2%.¹²

There is another problem analogous to the problem of supplying explanations by subsuming events under nomic correlations. Statistical generalizations express statistical correlations. There are many true statistical generalizations about events of a certain types that do not explain the occurrence of those events. For example, the fact that there is a high correlation between regularly consuming birth control pills and avoiding conception does not explain why Jones did not conceive while he was taking his wife's birth control pills.¹³ Given the inductive-statistical account, there has to be another restriction on the types of inductive arguments that count as explanations, such as a *requirement of strict maximal specificity*, that would rule out the use of statistical laws that

introduce nomically irrelevant properties. Of course, its difficult to give an account of what a nomically irrelevant property is.¹⁴

The biggest problem with the inductive-statistical account is that it does not seem to allow for a large number of statistical explanations that actually appear in science: many statistical explanations do not rely on premises that would provide a valid inductive argument for the occurrence of the event to be explained. For example, paresis occurs in a small percentage of cases where people have syphilis. It is not probable, then, that someone would get paresis if they have syphilis, so, according to the inductive-statistical account of explanation and its high probability requirement, we could not explain a case of paresis by pointing out that the subject has syphilis. But paresis *only* occurs in syphilitics, so pointing out that someone has syphilis is a common "explanation" for why he has paresis.¹⁵

This is not a small problem. Many familiar examples of statistical explanation fall outside of the inductive-statistical schema in the same way: cases of lung cancer are often explained by pointing out that the cancer victim smoked heavily. But no one smokes so heavily that they have a 90% chance of getting lung cancer. Instead, the point is that heavy smokers are much more likely than nonsmokers to get lung cancer, so if a heavy smoker gets lung cancer, heavy smoking is probably its cause. This explanation relies on a comparison of one class with a control group, and it is a familiar explanatory strategy. This observation motivates the *statistical-relevance* approach: Events are explained (not by giving a valid inductive argument, but) by showing what conditions are statistically relevant to the occurrence of the event.¹⁶ A condition R is statistically relevant to the occurrence of Q under condition P if and only if

$$[13] \quad p(Q|P \cdot R) \neq p(Q|P)$$

(i.e., the probability of the occurrence of Q under conditions P and R is not the same as the probability of the occurrence of Q under P alone).

Many of the problems with the inductive-statistical account are easily dissolved on this view. It is easy to see that being a population with a sickle-cell allele frequency of 0.5 is statistically relevant in cases where populations have sickle-cell anemia rates of 25%: there is rarely a case where a population has a

different sickle-cell allele frequency and has a sickle-cell anemia rate of 25%. In the case where Jones avoided conception by taking his wife's birth control pills, we can see that being a man is the statistically relevant condition: the probability of avoiding getting pregnant given that one is male and takes birth control pills is (slightly) greater than the probability of avoiding getting pregnant given that one takes birth control pills; the probability of one avoiding getting pregnant given that one is male and takes birth control pills is no greater than the probability of getting pregnant given that one is male. In cases of paresis it is easy to see that syphilis is the only relevant factor: in the absence of syphilis there is no condition that raises the probability of getting paresis to greater than zero.

Of course the statistical-relevance account requires a lot of elaboration. It is not without competition. But no view has clearly become the established account of statistical explanation, and the immediate prospects for a "received" account are poor.

2.3 Explaining Property Instantiation

Philosophical discussions of scientific explanation usually exhibit a preoccupation with the explanation of events. The material that I have reviewed so far does not cover any new ground, and may seem a little pointless. But event-explanation is not the only sort of explanation, and the review may make it a little easier to appreciate the difference between explanations that are intended to explain the occurrence of events and explanations that are intended to show how properties are instantiated in particular sorts of systems.¹⁷ When we ask a question of the form 'why does x have the property P?' we are usually asking for an explanation of the occurrence of an event, say, 'why did the gas expand (to such and such volume)?' But another question of the form 'why does x have the property P?' is 'why is ice less dense than water?' It is hard to construe this as an inquiry about what brought an event about. The appropriate answer would not include a reference to any sort of event; instead, it would refer to the constituents of both ice and water--water molecules--and to the relationships between those constituents in the two different forms of water. A similar question about property instantiation is 'Why are do some people have darker skin than others?' This might be a question

about how people come to have the skin color that they have, and so be correctly answered in terms of genetic and evolutionary laws, or it might be a request to know how skin color is instantiated, in which case the correct answer would refer to the amount of melanin in our skin. These sorts of questions about property instantiation are common in science. Following Robert Cummins (1983), I shall use *property analysis* to refer to the explanation of how properties are instantiated in systems.¹⁸

2.3.1 Property Analysis

The strategy of property analysis is relatively straightforward. Suppose we are trying to explain how a property, P , is instantiated in a system, s . First we would introduce a *composition law* which specifies the components and organization of the system in question:

[14] s has components C_1, \dots, C_n , and they are organized in manner O .

An *instantiation law* of the form,

[15] For all x , if x has components C_1, \dots, C_n , organized in manner O , then x has property P ,

can be derived from *nommic attributions* specifying the properties of C_1, \dots, C_n . [14] and [15] provide us with what we wanted to explain:

[16] s has property P .

I have construed property analyses as valid deductive arguments that have as premises at least one universal law and a statement of "initial conditions" as premises, and a statement of what was to be explained as the conclusion. Thus property analysis seems to conform to the deductive nomological account of explanation. Unfortunately, the similarities between event-explanation and property analysis are usually obscured in the literature. For example Salmon provides a good example of property analysis when he explains the deductive-nomological account of explanation: 'In providing a D-N explanation of the fact that

this penny conducts electricity, one offers explanans consisting of two premises: the particular premise that the penny is composed of copper, and the universal law-statement that all copper conducts electricity' [1984, p. 27.]. But Salmon does not point out that this example is very much different from examples of event-explanation: 'If something is copper, then it conducts electricity' is a universal conditional, but in it the 'if...., then...' locution in does not indicate a causal or statistical relation between events any more than the locution does in the sentence 'if 10/2 is 5, then $5 + 5 = 10$.' In property analyses, premises of the same form as [15] are not transition laws. Similarly, though typically events are designated by ascribing properties to things (at times), a request to know why *s* is *P* is not always a request to know what brought about a particular event.

The importance of property analysis may come out best in explanations of *transition laws*. Laws, according to most of the standard works that assume the deductive-nomological account of explanation, are explained by deducing them from more general laws. For example, while he discusses several instances of explanations that might be appropriately classified as examples of causal laws, Nagel [1960, pp. 33-37] presents the explanation of the following law:

[17] For all *x*, if *x* is ice, then *x* floats in water.

(Note that [17] is not actually a law under which events can be subsumed: though the 'if...., then...' form of the statement makes it look like a statement about transitions, it is not about transitions or types of events. Surely, no one wants to claim that something-being-ice-events cause that-something-floating-on-water-events. An appropriate causal law would instead state that ice-being-placed-in-water-events cause ice-floating-events. [17] is a nomic attributions: it attributes (dispositional) properties to certain sorts of things. The causal law that I have suggested is true if and only if ice has the disposition to float in water, so it is easy to confuse the two sorts of laws. How do we explain that causal law? We could refer to [17], but it just says that things with the property of being ice have the property of floating on water. [17], then, just says something about the dispositions of ice, and that is precisely what we have to explain.) Nagel deduces [17] from Archimedes' Principle--that a fluid buoys up a body with a force proportional to the weight of

the fluid the body displaces--along with references to the relative densities of ice and water. Nagel, in effect, shows that [17] is an instance of the more general law that less dense things float in denser substances, and thus the explanation conforms to what is almost accepted as "conventional wisdom" in the philosophical literature: transition-laws are themselves explained by subsuming them under more general transition laws.

However, there are many examples of transition laws that cannot be explained through subsumption to more general laws. Consider, for example, the standard explanation of why salt is soluble. First we refer to a composition law to the effect that salt is composed of NaCl molecules with such-and-such bondings and such-and-such relationships to each other. Then we appeal to an instantiation law to the effect that anything with those sorts of components and that organization will dissolve in water. From the composition law and the instantiation law we can infer that salt dissolves in water. Of course, there would have to be an explanation of the instantiation law: it would have to be derived from nomic attributions specifying the properties of the NaCl molecules. Those laws might in turn have to be explained by referring to the composition of the molecule and the properties of Na and Cl.

Typically, transition laws that specify the dispositions of complex systems will resist explanation thorough subsumption to more general transition laws. For example, suppose we were trying to explain the kneejerk reaction to a blow on the front of the knee. We might derive that "law" from a more general transition law to the effect that events that propagate impulses at one end of a reflex arc cause muscle twitch events at the other end. How do we explain that? Eventually we run out of more and more general transition laws, but there is still something left to explain: why do reflex arcs behave the way the do? The obvious strategy for finding the answer is to determine how the components of the reflex arc work. Our analysis of a dispositional properties include composition laws that refer to components of the system that have dispositional properties. Thus, though the way systems instantiate transition laws may be explained through property analyses, those analyses may refer to other transition laws. The important point is that in these sorts of explanations it is not just transition laws to which we must appeal.

Cummins [1983, pp. 15-16] suggests that, historically, atomist theories are the most important examples of property analysis. When those theories were first developed people lived in a macrophysical world; the physical things of which they were aware were things that they could see and feel. The events that they needed to explain involved macrophysical objects, and those events could be explained in terms of other macrophysical events: They could see that the salt dissolved because it was put in water, and they could see that this causal relation held generally between salt-dipping-in-water-events and salt-dissolving-events. But of course there were other questions to answer, such as 'Why does salt dissolve when it is put in water?' They could not answer that question by merely saying that it is a law that salt dissolves in water, or by saying that salt is soluble. What was required was an explanation of how solubility is instantiated by soluble materials.

This is presumably the sort of thing that atomism was developed to explain. The hypothesis was that macrophysical objects exhibit the properties they have as a result of their constituents and the organization of those constituents. The constituents were elementary particles called 'atoms'. These atoms were organized in certain ways because they had certain properties that made some atomic structures possible and others impossible. Of course this needed to be explained too, and the earliest atomists explained these possible relationships by hypothesizing that atoms have "hooks" and "eyes". Different sorts of atoms have different numbers of hooks and eyes, and their relations with other atoms are determined by the number of hooks and eyes that they have. With that the explanation is relatively complete. There is no need to explain how atoms have hooks and eyes-- i.e. how they instantiate the properties of having hooks and having eyes, because atoms have no constituent structure. The properties of having hooks and having eyes can only be explained by explaining the "concept" of having hooks and the "concept" of having eyes.¹⁹

The important point to notice here is that if all scientific explanation was devoted to explaining the occurrence of events, then there might not have been any need for atomistic theories, ancient or modern. All the events that originally needed to be explained were events in the macrophysical world. Most, if not all, of those events could be explained by referring to causes that also belonged to the

macrophysical world. But laws about the relationships of events, in effect, just specify dispositions of systems. If there had been no need to explain how some of these dispositions were instantiated, there may have never been any need for hypotheses about the existence of atoms.

2.3.2 Interpretive Analysis

Disposition analyses are usually "descriptive." An analysis is descriptive when the dispositions to be analyzed are specified by a "physical" descriptions of their precipitating conditions and manifestations. For example, suppose we had a device that always produces a card with holes arranged in such-and-such a manner when a card with holes arranged in such-and-such other manner is put into a slot in the right side of the machine. Since the disposition is specified descriptively, we would probably give an analysis that appealed only to "physical" facts about the constituents of the machine and their interrelationships: it might be analyzed by referring to a part of the system with a disposition to carry a current when and only when both gates A and B are closed, etc.

There are, however, many examples of dispositional analysis that are "interpretive": they specify the dispositions to be analyzed via their *semantically interpreted* precipitating conditions (inputs) and manifestations (outputs). For example, we might describe the card that was put into our device as asking what the sum of four and four is, and the card that came out as the answer that the sum of four and four is eight. In this case where the disposition (or "capacity") is specified via semantic interpretation it would be appropriate to analyze the disposition in terms of other semantically interpreted dispositions. Instead of analyzing the disposition by referring to the parts of the device and their own (electrical) dispositions--for example, carrying a current when and only when both gates A and B are closed--we might analyze it by referring to subdispositions such as the disposition to infer that if A and if B, then A & B. These dispositions might in turn yield to interpretive analysis. When we reached elementary semantically specified dispositions that do not yield to interpretive analysis, then we would try to provide descriptive analyses of the dispositions, i.e., we would try to

show how the elementary semantically specified dispositions were physically instantiated in the system.

Transitions (or causal relations) between events cannot be entirely ignored in interpretive analyses. If we analyze a disposition into several other dispositions, we might still have to explain why the manifestation of one disposition precipitates another.²⁰ This is straightforward in dispositions that have a descriptive analysis, because the relations between elementary capacities can be explained by referring to the causal relations between them. It is more complicated when we provide an interpretive analysis of a disposition. Suppose that we analyze the capacity of a calculator to do arithmetical operations into capacities to add and subtract, and then into capacities to "carry numbers", etc. Since numbers and other sorts of abstract entities have no causal properties, we can explain how the "output" of one semantically specified capacity can lead to the "input" only by continuing our analysis of the capacities until their elementary semantically specified capacities yield to descriptive analysis and then we can explain the causal relationships between the elementary capacities. In other words, before we can explain the causal relationships between semantically specified capacities, we have to see how they, or their analyzing capacities, are instantiated in the physical system.²¹

Interpretive analysis often allows us to achieve a generality in explanations that descriptive analysis does not provide. For example, a Macintosh computer can play a reasonably good game of chess using Sargon III. We might give a interpretive analysis of its disposition to play chess by referring to its subcapacities to rank board positions, to sort through variations, and so on. Those subcapacities might also yield to interpretive analysis. Eventually we should be able to provide descriptive analyses of the elementary semantically specified capacities. IBM computers also play a reasonably good game of chess using Sargon III. The analysis of an IBM's capacity to play chess when it executes Sargon III might be exactly the same as the analysis of a Macintosh's capacity to play chess when it executes Sargon III, down to the point where the elementary semantically specified capacities are descriptively analyzed--in fact, we would expect them to be the same since they are executing the *same* program. The instantiation of the elementary semantically specified capacities would differ because the computers are physically different,

and so the physical instantiations of those dispositions cannot be the same. However, if we explained the same dispositions (to "play chess") via purely descriptive analyses we would get two completely different analyses for the two machines because of their physical differences. Thus interpretive analysis allows us to account for some obvious similarities between the dispositions of the two computers that descriptive analysis would overlook. This may allow us to explain how, for example, cognitive capacities might be instantiated in systems with radically different physical properties.

It is not hard to find accounts of interpretive analysis in the cognitive science literature, though, of course, those accounts differ in detail, emphasis, and terminology according to the preferences of their individual authors. This sort of sketch of interpretive analysis seems to provide an attractive picture of how we explain the behavior of certain sorts of system--especially those that perform "cognitive" processes. Though it may only be a philosopher's worry, these accounts are conspicuously silent about what, exactly, locutions like 'interpreting dispositions as...' mean. I will try to clear up some of these worries in chapter 4.

NOTES

1 Of course I am not claiming that it is generally accepted that property attributions are justified if they provide psychologically compelling "explanations." There are many restrictions on what count as scientific explanations.

2 Here I mean "concrete event," where a single event can be referred to by several logically unrelated sentences. For example, the sentences 'Kripke is writing a paper' and 'Kripke is writing a philosophy paper' might pick out the same concrete event. Usually we do refer to events by ascribing properties to individuals, such as population p having gene frequency P (at t). But events are not just individuals with properties (at times). For example, the event *population p being of such and such size* can be the very same event as *system s having a predator population of such and such size*, even though the individuals and the properties mentioned in the descriptions are different.

Events, as we shall see, are subsumed under laws because of their properties. They can be subsumed under causal laws in virtue of different properties. For example, an event can be subsumed under a law about populations under the description 'population p being of such and such size' and under a law about some other sort of system under the description 'system s having a predator population of such and such size.'

States of systems can also be described in many different ways. Suppose, for example, that we construe belief states as relations to propositions. The psychological state of A standing in the belief relation to P might be the same as the state of A having such and such neural properties. (That does not entail that the property *standing in the belief relation to P* is the same as the property *having such and such neural properties*.) In the scientific literature the distinction between events and states of systems seems to be largely arbitrary, depending, for the most part, on the duration of the referent.

In science events are usually individuated according to their causal properties. Of course, claiming that causal indiscernability is the criterion of event identity is not much help since an event's causal properties can be determined only through its relationships with other events. Thus, the criterion of causal indiscernability may presuppose a criterion of event identity.

3 There is actually some disagreement about whether we have to refer to universal laws in order to explain individual events. Michael Scriven [1958] and Robert Cummins [1983] argue that we can explain why individual events occur

without referring to scientific laws. For example, Cummins [1983, p. 5] argues that we can explain why the front window shattered simply by pointing out that a ball hit it, without mentioning any sort of law. We *justify* our explanation by pointing out that glass is so-and-so much fragile, and that so-and-so much fragile things shatter when struck with so-and-so much force, but, Cummins claims, justifications are not explanations.

The predominant view is that the reference to the law is essential to the explanation, since we have to refer to general laws at some point to show why the ball striking the window explains why the window shattered. (See, for example, Hempel and Oppenheim [1948], Carnap [1966, p. 7], and Nagel [1961, p. 31].

- 4 Hempel and Oppenheim [1948] is the classic presentation of the deductive-nomological account of explanation. For other examples, see Carnap [1966, p. 7], and Nagel [1961, p. 38].

Carnap [1966] has actually claimed that (non-statistical) universal laws have the form ' $\forall x(Px \supset Qx)$.' I'm sure that he assumes some constraints about which statements of this form count as laws, including some to ensure that vacuously true universal statements are not laws.

- 5 Because of the structure of our language we usually refer to events by ascribing properties to individuals--for example, the 'assassination of Archduke Ferdinand'--and only rarely by name--the big bang, D-Day. In science events are described in the same way. Thus, it is common to see transitions represented as

[a] If $P a$, then $Q b$,

and to see transition laws represented as

[b] For all x and y , if Px , then Qy .

There is a small problem here, and a large one. The small problem is that the variables range over individuals, and the implication is that two individuals are involved. Many generalizations in science describe changes in systems that do not have identifiable parts--population genetics, for example, describes transitions in the gene frequencies of single populations, not relationships between different things. To the best of my knowledge, generalizations of the form of [b] never occur in population genetics. We might, then, represent laws

as if they describe changes in certain systems (which are, actually, just individuals):

[c] For all x , if Px then Qx .

This shares a rather large problem with [b]: representing scientific generalizations as in [b] and [c] might lead us to think that events are just things having properties (at times). Usually we do identify events by one distinguishing feature, such as population p having gene frequency P (at t). But events are not just individuals with properties (at times). For example, the event *population p being of such and such size* can be the very same event as *system s having a predator population of such and such size*, even though the individuals and the properties mentioned in the descriptions are different. Needless to say, this can lead to considerable confusion. (Fred Feldman [1980] argues that Kripke's argument against psycho-physical identity theory rests on such a confusion.)

6 Of course there are notorious problems involved in saying just what the truth conditions for subjunctive conditionals are. It seems as if an account of subjunctive conditionals would have to say something about the truth of statements in different but "similar" cases to the actual case. But it is difficult to see how we could explicate similarity without appealing to scientific laws. If we do explicate similarity by appealing to scientific laws, then the requirement that laws support counterfactual claims is not much help.

The currently fashionable semantics for subjunctive conditionals in terms of similarity orderings of possible worlds simply leave the difficult questions about similarity unanswered.

7 Of course, there have to be many other restrictions as well. See, for example, Nagel [1961, chapter 4].

8 I owe this taxonomy of methodological concerns to Cummins [1983, pp. 11-13].

9 Unfortunately, the classic examples of statistical explanation are much more complicated than the examples of non-statistical explanation that I use here. Part of the complication is that the classic examples of statistical theories--statistical mechanics and quantum mechanics--have laws that are not of the form of [11]. In fact, those theories are often called 'deterministic'. To sort through this mess in the most elementary way possible, consider classical

mechanics. In classical mechanics the (mechanical) state of a system at a time is completely described by the three position coordinates for each element at that time and the three momentum components for each element at that time. These coordinates are called the 'state variables' for the system. An equation, called the 'force function,' can be used to determine the mechanical state of any (isolated) system given the mechanical state of that system at some earlier time. The explanation of why a system was in a particular state at t' would presumably refer to its cause, the state of the system at an earlier time, t . Since a unique mechanical state for a system at t' can be deduced from the laws of mechanics and the mechanical state of the system at t , mechanics is often called 'deterministic.'

Statistical mechanics was gradually developed for the obvious reason that it is beyond human capabilities to actually determine what the mechanical state of any complicated system actually is: We cannot determine, for instance, the instantaneous position and velocity of elements that a complete state description requires; we can only give average values over some (small) interval. In order to avoid the difficulties of not being able to determine the individual motions of elements in systems, new hypotheses were added to classical mechanics that concerned the probability of elements in certain systems being in various mechanical states during any small interval of time. With the new hypotheses it became possible to define a mechanical state in terms of the statistical properties of the individual elements. The state variables are thus *statistical* state variables. Given this conception of a state, a unique (statistical) mechanical state for a system at t' can be deduced from the laws of statistical mechanics and the (statistical) mechanical state of the system at t . Thus, statistical mechanics seems to be deterministic in the same way that classical mechanics is. An explanation in statistical mechanics of why a system was in a particular state at t' would presumably refer to the (statistical) state of the system at an earlier time, t , and a universal conditional statement of the following form

For all x , if Px , then Qx

where 'P' expresses the property of being in a (statistical) mechanical state of such-and-such kind (at t) and 'Q' expresses the property of being in a (statistical) mechanical state of another kind (at t').

Even though statistical mechanics is deterministic in this way, its laws are really just statements of probabilities in disguise. For example, if we were trying to explain the movement of a satellite at t' we would refer to the probability distribution of the state variables for the elements that make up the satellite at some earlier time, t , and then use the laws of mechanics to determine

a unique probability distribution for those state variables at t' . Quantum mechanics is similar to statistical mechanics in that it does not require that a state description be complete in the way that classical mechanical state descriptions are, but only requires state descriptions which provide probability distributions for some of the variables of position and momentum. The important difference between these two types of theories is that statistical mechanics is used under the assumption that it could in principle be replaced by non-statistical classical mechanics and is used only as a convenient tool, while quantum dynamics is "irreducibly" statistical: Under the assumptions of quantum dynamics it is *theoretically* impossible to give precise values for every state variable as classical mechanics requires. (See note 10.)

- 10 The statistical nature of quantum mechanics rests on Werner Heisenberg's "uncertainty relations" which can be derived from the laws of quantum mechanics. One of these relations is expressed by ' $\Delta p \Delta q \geq h/4\pi$ ' where ' p ' and ' q ' are variables for the instantaneous coordinates for momentum and position of particular elements and ' h ' stands for Plank's constant. ' Δp ' is the coefficient of dispersion (or deviation, or "uncertainty") from the mean value of the momentum at a given instant. So that statement says that the product of the dispersions of momentum and positions is never less than $h/4\pi$. This, in effect, says that if the position of an element is made precise or is specified with a high probability, then its momentum can only be specified with a relatively low probability. If the position of the element is specified with a low probability, then its momentum can be specified with a higher probability.

Usually these relations are paraphrased as saying that if one variable is made (near) precise, then the other one cannot be "determined" with more than r probability. This makes the principle sound like an epistemological principle about how much we can know. But if we strip the principle of its various interpretations, it may be more accurate to say that the theoretical apparatus of quantum mechanics precludes giving precise values for both the position and momentum of any element. Nagel [1961, chapter 10] says that it would be, in effect, "ungrammatical" to provide precise values for both the position and the momentum of an element.

- 11 The inductive-statistical account is due to Hempel [1962 and 1965a].
- 12 Both the problem and this proposed solution were advanced in Hempel [1962 and 1965a].
- 13 The example is due to Salmon [1984].

- 14 James Fetzer [1981, pp. 125-126] has tried to give this sort of requirement.
- 15 The example is due to Scriven [1959, p. 480].
- 16 This view was first suggested in Salmon [1965], and a more comprehensive account appears in Salmon [1984].
- 17 Much of the credit for this emphasizing the difference goes to Robert Cummins [1983, chapters 1 and 2]. My sketch is to a large extent parasitic on his account.
- 18 Fodor [1975] points out that there is another way to construe questions like 'why does x have the property P?', 'in virtue of what does x exhibit the property P?', and 'what makes x P?' These might be requests to know what caused x to be P, they might be requests to know how x instantiates P, or they might be (poor ways to ask) 'what does it mean to say that x is P.' On this last reading these sorts of questions are requests for conceptual analyses, and not empirical at all.
Fodor argues that Gilbert Ryle's [1949] attack on "mentalist" psychology, and a good deal of the motivation for behaviorism, is based on this sort of ambiguity. For example, Ryle [p. 33] scoffs at references to a clown's mental processes in explanations of why that clown's behavior is witty. Instead, Ryle suggests, the clown's behavior is witty because it is unexpected, creative, interesting, etc. Fodor [1975, pp. 3-9] claims that Ryle may provide the correct answer to the question 'what does it mean to say that the clown is witty?' but he does not answer 'what causes the clown's witty behavior?'
- 19 If we could always demand more, then no explanation could ever be complete.
- 20 This would not be necessary for systems with dispositions that could be analysed in terms of subdispositions that were not systematically related to each other. Cummins [1983, p. 33] suggests the example of fiber optic bundles where the disposition to transmit a signal can be analysed in terms of the dispositions of the independent fibers, without any reference to systematic relationships between them.
- 21 If it seems odd to say that semantically specified capacities are dispositions to manipulate abstract objects, just consider that if we describe an operation of a

calculator as adding $2 + 2$, then we have described it as performing an operation on numbers, not numerals.

If it seems odd to say that semantically specified capacities can be descriptively analysed (i.e., can be instantiated in physical systems), just consider the calculator which instantiates a semantically specified capacity by manipulating symbols which represent numbers.

CHAPTER 3

PHYSICALISM, MATERIALISM AND THE NON-PHYSICAL SCIENCES

3.1 Introduction

Every attempt to discuss the relative explanatory capacities of the various sciences seems to plunge inevitably into a philosophical jungle that is neatly disguised by the two terms 'physicalism' and 'materialism'. As with most '...ism's, there is not much hope of defining a univocal sense of either term that will please even a majority of those who proudly label themselves with one or the other of those terms. In spite of that, philosophers continue to use those terms to define their positions and those of their fellow philosophers--with predictably confusing results. Some philosophers have adopted one or the other of the labels primarily from a conviction that some sort of atomism is true: that everything in the world is "made up" out of microphysical entities, and that everything that happens results from the properties of those microphysical entities. Others seem to have adopted one or the other of the labels from a conviction that "science is the ultimate arbiter of reality," as Baker [1987] put it--that whatever happens can, in principle, be explained through science and through science alone. Many of the conflicts about physicalism and materialism thus decay into arguments about what sorts of things really exist or what sorts of things demand explanation and can be explained. Thus the positions often just seem to be methodological principles stating one's unwillingness to recognize the existence of things that cannot be decomposed into microphysical parts or things that cannot be explained by some science.

Many philosophers seem to believe that the ontological thesis that "everything decomposes into microphysical particles"--however difficult that is to make clear--entails the thesis that "everything can be explained by physics." Physicalists and materialists are thus often supposed to support the view that all of science ultimately "reduces" to physics. Of course, the notion of reduction is as difficult and unclear as any in the philosophy of science. Presumably the claim that all of science reduces to physics entails, at least, what I will call the 'minimal reductionist claim':

MRC: Everything that can be explained by the other sciences could, in principle, be explained by appealing to (true) physical theories.

Since explanations in physics generally do not depend on appeals to semantic properties, if we knew whether the minimal reductionist claim is correct, we would know the answer to the question 'Is it necessary to appeal to semantic properties in order to provide adequate explanations of human behavior?' Of course, there are many philosophers--including Fodor, Pylyshyn and other adherents of "computational" models of cognitive processes--who claim to be either physicalists or materialists and who reject the claim that physics could, in principle, provide all of the explanations provided by the other sciences. It might pay, then, to try to describe a position that appeals to most of the various physicalists/materialists, and then to see what consequences this sort of position has for explanations outside of physics. I will try to describe the position without relying on the notion of reduction. I hope that it is not *too* confusing.

3.2 Physicalism and Materialism

The world is made up of individuals ("substances" or "continuants"), events (or "occurrents"), and properties (or "types"). Though some of us may have some philosophical qualms about putting it that way, those are the sorts of things that are typically referred to or expressed by the terms, variables and predicates in the vocabularies of scientific theories. A "physicalist/materialist" might make any or all of these claims:

- (I) Every individual is a physical thing.
- (II) Every event is a physical event.
- (III) Every property is a physical property.

These require some interpretation. Since I am primarily interested in the truth of MRC, and since I do not want to get embroiled in a debate about what sorts of things exist and what sorts of events actually occur, I will treat these principles as if the

are restricted to the individuals, events and properties that are referred to in the sciences.

Of course, 'referred to in the sciences' is pretty vague. Presumably physicalists/materialists think that in some sense I am a physical thing, that my behaviors are physical events, and that my biological "adaptedness" is a physical property. But no scientific theories of any sort mention *me* or *my* behaviors or *my* biological adaptedness. I am subject to certain laws because of my properties, such as *having such and such mass*: my actions are subsumed under scientific laws because of their properties, such as *being utterances of such and such sort*. We can say, then, that biological things are things that have properties that are expressed by predicates in statements of biological laws, that chemical events are events that have properties that are expressed by predicates in statements of chemical laws, etc.¹

- [I'] For any individual x , if x has a property P that is expressed by a predicate Π of a true scientific theory, then x is a physical thing
- [II'] For any event e , if e has a property P that is expressed by a predicate Π of a true scientific theory, then e is a physical event.

What are physical things and events? We could try the same approach: physical things and events are things and events with properties expressed by predicates of physics.

- [I''] For any individual x , if x has a property P that is expressed by a predicate Π of a true scientific theory, then x has a property P^* that is expressed by a predicate Π^* of a true physical theory.
- [II''] For any event e , if e has a property P that is expressed by a predicate Π of a true scientific theory, then e has a property P^* that is expressed by a predicate Π^* of a true physical theory.

These principles are incompatible with Cartesian dualism: if everything has some physical property, then there can be nothing without some property such as

mass, extension, location in space, etc.² But these principles are not strong enough for most physicalists/materialists. According to these principles individuals and events could be subject to scientific laws in virtue of their non-physical properties. But if the microphysical constitution of things and events is what makes them do what they do, then, in some sense, it is because individuals and events have physical properties that they do what they do.³

[I'''] For any individual x , if x has a property P that is expressed by a predicate Π of a true scientific theory, then P is expressed by a predicate Π^* of a true physical theory.

[II'''] For any event e , if e has a property P that is expressed by a predicate Π of a true scientific theory, then P is expressed by a predicate Π^* of a true physical theory.

Since individuals and events are of certain types if and only if they instantiate the properties that characterize those types, then if all the *properties* expressed by predicates of one theory are *properties* expressed by predicates of another, all the *types* "recognized" by the first theory are also *types* "recognized" by the second. So if [I''] and [II''] are true, then the types of individuals and events recognized by scientific theories must be types of individuals and events recognized by physics. Thus, [I'] and [II'] together provide a statement of the notorious "type-type identity thesis."

The type-type identity thesis has received some well-deserved criticism. There are numerous examples of types of things that are typically referred to in the vocabularies of the other sciences that are not characterized by a property that is expressed by a predicate in the vocabulary of any physical theory. For example, the type individuated by the property *being of such and such temperature* is not recognized by physics. Of course, the defender of [I] might respond that the property *being of such and such temperature* is the same as some property which can be expressed in purely physical terms of some such form as *having atoms of such and such sort organized in such and such manner*. However, the temperature of a gas is the mean kinetic energy of its constituent molecules, but the temperature

of a plasma is not, since plasmas have no constituent molecules. This is an example of "reduction relative to a domain of phenomena." *Being of such and such temperature* has not been identified with a physical property. Instead *being a gas of such and such temperature* and *being a plasma of such and such temperature* have been identified with different physical properties. If there are lawlike generalizations that can be made using the predicate 'is such and such temperature,' then we have a case where [I''] seems to be false. The problem seems to be even more obvious with properties like *being photosynthetic*, which can be instantiated by organisms with vastly different physical descriptions. If *being photosynthetic* is a physical property, then in the vocabulary of physics it would have to be expressed by an indefinitely long disjunctive predicate that applied to all the existing photosynthetic things (as well as all of the possible ones). That just is not the sort of property that physics deals with; I am sure that a predicate expressing that property will never appear in a physics paper.⁴

Though multiple instantiation causes problems for [I'''] and [II'''], I think that most physicalists/materialists would be satisfied with a weaker position. Usually the position is stated by saying that the properties in virtue of which individuals and events are subject to scientific laws "supervene" on the physical properties of the individuals and events. Unfortunately, it is not easy to explicate the concept of supervenience. Jaegwon Kim [1982] suggests that a set F of properties is supervenient upon a set G of properties with respect to a domain D just in case 'any two things in D are such that necessarily if they differ with respect to F then they differ with respect to G' [1982, pp. 51-52]. I am not sure what sort of necessity is supposed to be important here. Kim might mean "metaphysical" necessity--the sort of necessity that is sometimes equated with "truth in all possible worlds." But in that case the concept of supervenience will not be of much interest: Scientific laws are supposed to be contingently true, so though it may be true that such and such sort of electrical device transmits such and such sounds, it might not have (in some other "possible world"). Do we then want to deny that the device's capacity to transmit sound supervenes on its electrical properties, since electrically identical devices might have performed differently (in some other possible world)?

Perhaps Kim means "physical" or "nomic" necessity--the kind of necessity that scientific laws describe. But this does not seem to work either. On this reading, the bonding properties of molecules (i.e., their dispositions to bond with other sorts of molecules) supervene on their compositional properties (i.e., their properties of being composed of certain atoms organized in particular ways), since scientific laws require that any two molecules that differ with respect to their bonding properties also differ with respect to their composition. However, on this reading, it is also true that in many cases the compositional properties of molecules supervene on their bonding properties. For example, scientific laws require that any molecules composed of something other than two hydrogen atoms and one oxygen atom have bonding properties that are different from those of H₂O molecules. But an adequate account of supervenience has to take into account the fact that the bonding properties of H₂O molecules "depend," in some sense, on the composition of the molecules, though their composition does not depend on their bonding properties. So, on this reading of Kim's principle, supervenience does not require the right sort of connection between the two sets of properties.⁵

Presumably, if some properties of an individual or event "supervene" or "depend" on other properties of the individual or event, then we can explain (through property analysis) why that individual or event has the former properties by referring to the latter properties. Sometimes we will be able to explain why individuals and events have certain sorts of properties simply by referring to the physical properties of those individuals and events. I shall call these sorts of properties 'physically instantiated.'

- [D1] A property, P, is physically instantiated in (an individual or event) x, if and only if it is possible to explain how P is instantiated in x by referring only to properties of x that are expressed by predicates of true physical theories.

There are good reasons to think that not every property is physically instantiated (in something or other). For example, it is now popular to claim that things are "intelligent" (or exhibit some specific kind of "intelligence") in virtue of their functional organization: intelligence is supposed to be instantiated by things

that have certain "functional properties." Those functional properties are not physical properties. Of course, some of the things that have those functional properties have them because of the electrical properties of their parts, and those parts have those electrical properties in virtue of their basic physical components. But those who claim that things are intelligent because of their functional organization might not want to say that we can explain how intelligence is instantiated in a particular system without referring to the functional properties of the system. So, though it may not be possible to explain how each property is instantiated in individuals and events by referring only to the physical properties of those individual or events, it may still be possible to explain how they are each instantiated by referring only to properties that are themselves physically instantiated, or by referring to properties the instantiation of which can be explained by referring to properties that are physically instantiated, etc. To avoid prejudging the point, I am providing this (recursive) definition:

- [D2] A property, P, is physically instantiated* in (an individual or event) x, if and only if either P is physically instantiated in x, or if it is possible to explain how P is instantiated in x by referring only to other properties that are physically instantiated* in x.

I suspect that when people claim that certain sorts of properties supervene on physical properties they often mean to say that the former properties are physically instantiated* (by something or other). This suggests the following claims:

- [I*] For any individual x, if x has a property P that is expressed by a predicate Π of a true scientific theory, then P is physically instantiated*.⁶

- [II*] For any event, e, if e has a property P that is expressed by a predicate Π of a true scientific theory, then P is physically instantiated*.

Providing that individuals and events are the only sorts of things that have properties, [I*] and [II*] entail this principle:

[III*] For any property P that is expressed by a predicate Π of a true scientific theory, P is physically instantiated*.⁷

This, I hope, captures the spirit of the claim that "everything is physical." I think that [III*] provides a plausible account of a position that is widely held, if not explicitly defended. I am not sure how one argues for this sort of claim, since its truth depends on what the true scientific theories turn out to be like--and no one knows what that will be. It is attractive, I suppose, because it implies that nothing is more than a collection of its basic parts and that we can explain why things have the properties they have by referring to the components of those things. I think that most philosophers and scientists would regard as suspicious any scientific explanation that entailed that [III*] is false. If [III*] is a constraint on explanation then we should find out whether it entails the minimal reductionist claim:

MRC: Everything that can be explained by the other sciences could, in principle, be explained by appealing to (true) physical theories.

3.3 The Loss of Generalizations Argument

I have been liberally referring to the "vocabularies" of the various sciences. The sorts of things about which a particular science is supposed to provide explanations is, roughly, determined by its vocabulary: Biology explains biological phenomena (and properties) because that is what the vocabulary of biology ranges over; chemistry explains chemical phenomena (and properties) because that is what the vocabulary of chemistry ranges over. In fact, the vocabularies of the various sciences seem to define different levels of explanation. For example, biology is of no help to us in explaining the chemical properties of chemical compounds, because in the vocabulary of biology there is no way to refer to the sorts of things that we would have to appeal to in such an explanation.

Of course, if MRC is true, sciences like psychology, chemistry, and biology will be distinguished by what they cannot explain because of limitations in their vocabularies, while physics will have none of those limitations. Those physicalist/materialists defending the "autonomy" of non-physical sciences must say why physics is limited in its explanatory power in virtue of its particular

vocabulary, regardless of the truth of [I*], [II*] and [III*]. The closest thing to a "received view"--which Hilary Putnam [1960, 1973], Donald Davidson [1970], Daniel Dennett [1978b], Zenon Pylyshyn [1984] and many others have suggested or championed, and which Jerry Fodor [1974, 1975] put in its most famous form, is that physics fails to "capture generalizations" that other sciences can capture. In my discussion I will follow (roughly) Fodor's presentation.

3.3.1 The Argument

Presumably this is the sort of case that is of concern: Suppose we want to explain why particular muscles in the arm of a human subject twitch when a stimulus is applied to a particular neuron in the brain of the subject. The explanation in neurological terms might refer to a spike potential being realized in a certain neuron and activating other nerves according to neurological laws, with the end result that the muscle in the arm of the subject receives a certain sort of stimulus. This might be "explained" in physics by referring to a whole chain of causal relations between physical particles. But suppose we were trying to explain why that *type* of muscle twitch occurred in every human subject that was stimulated in that same way. An explanation in neurology might refer to the type of spike potential that was realized in every case, with the result that certain kinds of nerves were activated in every case, with the result that the same kind of muscles were activated in every case. But an explanation in physical terms would give a completely different story in each case, because the neurons and muscles of each of the subjects would be constructed out of elementary physical particles in different ways, and the spike potentials of each neuron would be physically instantiated* in different ways. Thus physics might provide an explanation of sorts, but there would be no indication of what was common to the different cases.

Consider the sorts of explanations that might be given for types of events that are not typically explained in physics. Suppose we were trying to explain the occurrence of events of a certain type in a particular system, *s*. According to conventional wisdom, we might explain the occurrence of those events by referring to some other (earlier) events and a law which says that events of the latter type cause events of the former type. Presumably, the law will be a law of a

science, \mathcal{S} , that is intended to explain events in systems like s , and that includes among its typical predicates 'S¹' and 'S²', which express the properties in virtue of which the two types of events are subsumed under the relevant law. A causal law would be of this form:

$$[1] \quad \text{For all } s, \text{ if } S^1s, \text{ then}_c S^2s. \text{ }^8$$

Thus our explanation might look like this:

- [2] 1. For all s , if S^1s , then_c S^2s .
2. S^1s .
3. Therefore, S^2s .⁹

But consider what happens if we try to give this explanation in the language of physics. Events that are typically described by 'S¹' in \mathcal{S} might have multiple instantiations in physics. Though we might truly say that

$$[3] \quad S^1s,$$

in the vocabulary of \mathcal{S} , an "equivalent" statement in physics would probably be much more complicated.¹⁰ For example, if we were originally trying to explain the propagation of a particular impulse in a particular nerve, then 'S¹' might express the property of *being depolarized at a dendritic synapse*. That property would certainly be instantiated in a huge variety of ways. It is also certain that none of those instantiations could be adequately described by only one typical predicate of physics. So a statement "equivalent" to [3] in the vocabulary of physics might look like this:

$$[3'] \quad (P^1a \text{ and } \dots \text{ and } P^d d) \text{ or } \dots \text{ or } (P^e e \text{ and } \dots \text{ and } P^h h).$$

The law that we referred to in our explanation might have this as its equivalent in the language of physics:

$$[1'] \quad \text{For all } x_1, \dots, x_n, \text{ if } [(P^1x_1 \text{ and } \dots \text{ and } P^d x_d) \text{ or } \dots \text{ or } (P^e x_e \text{ and } \dots \text{ and } P^h x_h)], \text{ then}_c [(P^i x_i \text{ and } \dots \text{ and } P^k x_k) \text{ or } \dots \text{ or } (P^m x_m \text{ and } \dots \text{ and } P^n x_n)].$$

If we suppose that ' $P^l a$ and ... and $P^d d$ ' is a description of a particular physical instantiation of the property of *being depolarized at a dendritic synapse*, then ' $P^l x_1$ and ... and $P^d x_d$ ' certainly is not a "typical" predicate of physics. If we accept that [1'] is a law of physics, and that physical types are characterized by the properties expressed by predicates that appear in the laws of physics, then we will be committed to claiming that predicates like ' $P^l x_1$ and ... and $P^d x_d$ ' do pick out physical types.

Of course, when we explain the occurrence of a type of event by appealing to a scientific "law," we appeal to statements that are presumed to hold for possible worlds that are like ours in the relevant respects--i.e., we appeal to laws that support counterfactual claims. But if [1'] is only extensionally equivalent to [1], [1'] will not hold in every "possible world" where [1] holds. We might, then, have to strengthen [1'] by adding an indefinite number of disjuncts to the physical descriptions of the types of events in order to cover the counterfactual instances.

Fodor adds another reason to doubt whether statements like [1'] could be laws. He points out that though we might call a statement ' φ causes ψ ' a law, and call a statement ' α causes β ' a law, we generally would not consider the statement ' φ or α causes ψ or β ' to be a law. Fodor's own example is that, though it might be a law that irradiation of green plants by sunlight causes carbohydrate synthesis, and it might be a law that friction causes heat, it is not a law that (either irradiation of green plants by sunlight or friction) causes (carbohydrate synthesis or heat) [1974, p. 109].

The three previous points are reasons to doubt whether [1'] is really a law of physics, or whether it is even a law. If we were willing to accept that statements like [1'] really are laws, we could then supply this "explanation":

- [2']
1. For all x_1, \dots, x_n , if $[(P^l x_1 \text{ and } \dots \text{ and } P^d x_d) \text{ or } \dots \text{ or } (P^e x_e \text{ and } \dots \text{ and } P^h x_h)]$, then_c $[(P^i x_i \text{ and } \dots \text{ and } P^k x_k) \text{ or } \dots \text{ or } (P^m x_m \text{ and } \dots \text{ and } P^n x_n)]$.
 2. $[(P^l a \text{ and } \dots \text{ and } P^d d) \text{ or } \dots \text{ or } (P^e e \text{ and } \dots \text{ and } P^h h)]$
 3. Therefore, $[(P^i i \text{ and } \dots \text{ and } P^k k) \text{ or } \dots \text{ or } (P^m m \text{ and } \dots \text{ and } P^n n)]$

Fodor claims that arguments like [2'] are not explanations because they fail to satisfy a pragmatic constraint on explanations: They do not show what it is among the events of the same type that makes them the same types of events. A law like [1] shows very clearly that all the events of a certain type are causes of events of another type, but a statement like [1'] neither tells us what the events described in the disjuncts of the law's antecedent have in common nor tells us what the events described in the disjuncts of the law's consequent have in common. Because of this, explanations in a physical vocabulary will "fail to capture generalizations" that are "captured" in explanations provided by different sciences.¹¹

Putnam's version of the objection also stresses pragmatic constraints on explanation rather than purely formal ones [1973, p. 132-134]. He points out that in cases where we try to explain the occurrence of an individual event the "relevant features" of the situation may not be brought out if we gave an "explanation" in terms of physics instead of in the vocabulary of another science. It seems as if this would be even more pronounced in explanations of types of events. Putnam claims that it is a pragmatic constraint on explanations that they point out the relevant features of causes and effects, and do not leave them 'buried in a mass of irrelevant information' [1973, p. 132]. Unfortunately he does not elaborate about how much "relevance" is required to provide an explanation. It is clear, however, that his position is similar to the claim that Fodor has made--not to mention Davidson, Dennett, Pylyshyn, etc.

Though he does not flesh out the example, Fodor [1974, pp. 103-104] mentions that Gresham's law is the sort of law that we cannot replace with an extensionally equivalent statement made in purely physical terms without losing an important generalization. Gresham's law (restricted to a bimetallic standard) says roughly that

[GL] If two different metals are freely coined at the mint and the ratio of the value of one metal to the other in the market diverges from the mint ratio, coins of the metal which is underrated will disappear from circulation.¹²

Suppose, for example, a country minted a gold coin and a silver coin and redeemed the coinage at a ratio of one gold coin for every 10 silver coins. Gresham's law

states that if the gold was actually valued in the "market" at 11 times the value of silver, then the gold coins would disappear from circulation--presumably because it would be more lucrative for individuals to use the gold coins in some other way than as a standard of exchange.

Of course, a physical description of the different metallic coinage that might be involved in monetary exchanges would probably not fit into all of the books in all of the libraries of the world. Trying to describe the "markets" in which these exchanges take place would be an even more daunting business. But Fodor is not claiming that providing explanations of events that fall under Gresham's law would be *difficult* in physical terms, he is claiming that it would be *impossible*. There are lawlike generalizations that can be made about monetary exchanges when they are described in terms of currencies, market structures and exchange ratios, but, according to Fodor, there are no lawlike generalizations that can be made about the monetary exchanges governed by Gresham's law when they are described in physical terms. First there are the formal problems: suppose the antecedent conditions of each instance of Gresham's law were each specified in the language of physics, and that the consequent conditions were also. The world market in 1835 and the discrepancies between the Spanish, English and American mint ratios at that time do not seem to be instances of a natural kind in physics, and so cannot be governed by anything like what we have traditionally considered the laws of physics. And, of course, the disjunction of all the antecedent conditions and the disjunction of all of the consequent conditions certainly are not related by anything like what we have traditionally considered laws of physics.

Even if we set these formal complaints aside, it seems like the original version of Gresham's law tells us something that its extensional equivalent in the vocabulary of physics does not: If coins of different metals are minted and their mint ratio is different from the ratio of the value of one metal to the other in the market, then the underrated coins are going to disappear from the market. The "law" in a purely physical vocabulary will tell us only that if one of a huge number of physical descriptions is satisfied, then one of another huge number of different physical descriptions will be satisfied. One of the disjuncts of the antecedent will provide a purely physical description of the market and the exchange of heavy and

light pieces of eight in the Colonies, another of the disjuncts will provide a purely physical description of the market and the exchange of undervalued gold coins in the reign of James I, etc.; one of the disjuncts of the consequent will provide a purely physical description of the disappearance of the heavy pieces of eight from the Colonies, another of the disjuncts will provide a purely physical description of the disappearance of undervalued gold coins from the market during the reign of James I, etc. There is no clue about what is common among the disjuncts in the antecedents and the disjuncts in the consequent, and the observation that Gresham made is lost in the tangle.

3.3.2 Some Observations

Fodor makes two claims for his view: (1) 'It allows us to see why there are special sciences at all', and (2) 'it allows us to see how the laws of the special sciences could reasonably have exceptions...' [Fodor, 1974, p. 113]. The first claim, I think, is a bit strong. The second claim leads into some difficult territory.

If Fodor is right, and statements like [1'] are not physical laws, and arguments like [2'] are not explanations, then we cannot rely on physics to provide certain sorts of scientific explanations that the non-physical sciences can provide. But everyone already knows one reason why we have the non-physical sciences: even if we could, in principle, rely on physics to supply the sorts of explanations that the non-physical sciences provide, those sorts of explanations would be too complicated for mere mortals to provide.

Fodor's second claim is more interesting, and he explains it this way:

We allow the generalizations of the non-physical sciences to *have* exceptions, thus preserving the kinds to which the generalizations apply. But since we know that the *physical* descriptions of the members of these kinds may be quite heterogeneous, and since we know that the physical mechanisms which connect the satisfaction of the antecedents of such generalizations to the satisfaction of their consequents may be equally diverse, we expect both that there will be exceptions to the generalizations and that they will be 'explained away' at the level of the reducing science. This is one of the respects in which physics really is assumed to be bedrock science; exceptions to *its* generalizations (if there are any) had better be

random, because there is nowhere 'further down' to go in explaining the mechanism whereby the exceptions occur.[1974, p. 112]

Presumably this is the sort of case that Fodor has in mind. Gresham's law tells us that an undervalued currency will disappear from circulation. But there might be exceptions to Gresham's law: There might be a culture, for instance, that values beauty more than gold, and so instead of hoarding (monetarily) undervalued coins, they might hoard (monetarily) overvalued and more beautiful, coins. This sort of occurrence is merely an aberration, so we can safely ignore it rather than redescribe all the different currencies and monetary exchange systems in physical terms and thus lose the generalization that Gresham's law does provide (when it holds). Psychology or sociology can explain the aberration.

But treating exceptions this way seems much too cavalier. When Fodor first suggested Gresham's law as an example, he asked his readers to suppose that it was 'really true.' But presumably Gresham's law is a universal conditional (that supports counterfactual claims), so, if Gresham's law is actually true, it does not have any exceptions. Indeed, how are we supposed to test whether a generalization is a law if we are allowing laws to have exceptions? It seems as if we have grounds to question whether generalizations that have exceptions are really laws at all.

Perhaps I am naive even to wonder whether laws can have exceptions: certainly we know of a good many laws that have exceptions, and they are frequently used to provide scientific explanations. Even the laws of physics have exceptions. The pendulum law--which states that the period of a pendulum is the square root of the pendulum's length divided by the constant of gravitation, multiplied by 2π --is false for pendulums made out of iron and under the influence of strong magnetic forces. Laws about the conductivity of certain metals are false if the metals are supercooled or superheated, etc. Physical laws are usually stated on the assumption of ancillary hypotheses to the effect that the system governed by the law is not in such and such sorts of environments so the law in question will not fall victim to wildly counterfactual situations. One of the interesting aspects of working science is that scientists are often faced with situations where they can either dismiss an apparent exception to a law with an ancillary hypothesis about the sorts of systems in which the law is intended to hold (or with an ancillary

hypothesis about the testing procedures, or with an ancillary hypothesis about the accuracy of the instruments, etc.), or they can treat the apparent exception as a falsifying instance.

Exceptions to laws, when not considered falsifying instances, are dismissed with *ceteris paribus* clauses that provide conditions on the applicability of the laws. Gresham's law, for example, might have a rider to the effect that it does not hold when the individuals involved in the monetary transactions are not motivated solely by personal profit, and another rider to the effect that they are at least minimally intelligent, etc. A law about human behavior might have a proviso to the effect that the subjects in question have not suffered from a massive blow to the head, and they might have another disclaimer for alcoholics, etc. Of course, these clauses are usually implicit, and they are stated only when we find a case that we would rather dismiss as an odd case rather than treat as a falsifying instance. If we can dismiss the odd cases that way, then we can preserve the generalizations that we would lose completely by relying on physics alone.

It is, however, in dispute whether laws with "exceptions" are explanatory if we do not know the (implicit) *ceteris paribus* clauses that dismiss those exceptions. For example, it is easy to assume that, though Gresham's Law fails to explain why the undervalued currency fails to disappear in the case where the beauty-lovers hoard the overvalued coinage, Gresham's law does explain the cases where undervalued currency does disappear. But surely it is reasonable to ask, 'What is the difference between these cases? In each case the antecedent of the law is satisfied, so why the different outcomes?' There are some grounds for arguing that Gresham's law is not explanatory as it stands, since it does not explain the difference between the two cases, and that an appropriate amendment has to be added.¹³ Thus, on this view no generalization is explanatory unless it has a full complement of *ceteris paribus* clauses telling us in which cases it is not intended to hold.¹⁴

Though this requirement may be much too strong, and though it is not a requirement that we would actually make of working scientists, who work with a largely implicit understanding of the domain in which their generalizations are supposed to hold, we do ask for these kinds of explanations. The more exceptions

there are to a law, the more reasonable such demands are, and the more reasonable the claim that references to the law do not provide explanations. The requirement is interesting, because if we do provide a complete account of the kinds of cases where a law is not supposed to hold, then we have, in effect, provided a complete description of the sorts of systems in which the law is supposed to hold.¹⁵

There is an important difference between physics and the other sciences when it comes to supplying *ceteris paribus* clauses for their laws. Presumably the provisos about when the laws of physics will and will not hold would be stated in physical terms. But the provisos of a non-physical science would often not be stated in the vocabulary of that science. We cannot, for example, say without begging the question that a monetary exchange system is one in which all of the laws of economics hold. Instead, some consideration has to be given to what sorts of things are going to count as monetary exchange systems, and that will probably be done in terms of the psychological and epistemological properties of the individuals in the system, as well as a good variety of other properties. In turn, most psychological laws do not hold for every human, so there has to be some understanding about what sorts of people or things are governed by those laws. To define them explicitly, we might have to refer to neurological properties and laws or perhaps functional properties and laws.

In fact, I doubt that there are any "exceptionless" psychological laws. Some of Pavlov's laws about conditioned responses seem like good candidates for exceptionless laws, since they are relatively uncomplicated and they don't require particularly intelligent subjects. But, of course, we have to point out that the laws do not hold for individuals whose brains have been damaged in certain ways, who have certain sensory handicaps, or who have been previously conditioned in some confounding way. Since the military has recently discovered that most standard brain functions can be disrupted (temporarily, I hope) with a good dose of microwaves, we have to discount those cases also. Of course, those sorts of cases are so obvious that a psychologist typically is not going to have to worry about them, and will not even bother to write down the appropriate disclaimer. Those cases do show that, if we were to say in exactly which cases psychological laws are supposed to hold, we would have to retreat to terms that are not in the vocabulary of

psychology, and we might have to refer to some properties that are typically considered physical properties.

I think that these common-sense observations show that by appealing to the laws of economics, psychology, biology, chemistry (and maybe even physics) we never explain *anything* without some prior understanding about what systems these laws are about--and that sort of understanding will only be provided by some less "special" science. However, it would be a *non sequiter* to argue that Fodor, Pylyshyn, and Putnam, et al., are wrong on the basis of that observation. Even if the non-physical sciences are not largely autonomous, they may still be required to explain why certain sorts of phenomena occur. I have gone on at length because I think that the "loss of generalizations" argument has lead some to take the talk of "levels of explanation" too seriously, and to occasionally forget that the "special sciences" are dependent on the not-so-special sciences (a point I will exploit in chapter 8). After all [Fodor, 1974] is titled 'Special Sciences (or: The Disunity of Science as A Working Hypothesis)'. My ecumenical conclusion is that the non-physical sciences must rely on their less special counterparts. Talk about particular "levels of explanation" is mostly heuristic, and making distinctions between the "vocabularies" of the different sciences is merely convenient for those writing chapters on scientific explanation.

3.3.3 More Observations

Since I have discussed the relationship between physics and the other sciences and the loss of generalizations argument at some length, I should say something about whether that argument is successful and whether MRC is true. Unfortunately, for at least two reasons, the argument is difficult to evaluate: First, the loss of generalizations argument relies on a collection of considerations that weigh against MRC, but, which, I think, is not even considered conclusive by the philosophers who have presented versions of the argument.

The second problem is that the loss of generalizations argument ultimately seems to support only the claim that "capturing generalizations" is a pragmatic constraint on adequate explanations. Even if we knew exactly what capturing generalizations amounts to, it is always hard to determine whether a pragmatic

constraint is warranted. If I claim that something is a necessary condition for the adequacy of a scientific explanation, then I imply that there is a univocal sense of the phrase 'scientific explanation' to be analyzed--or at least that there is some predominant usage that I am analyzing. Surely though, even in standard scientific work the phrase 'explanation' is used in many (rather imprecise) ways. There may not be any correct analysis of 'scientific explanation,' so the best we can do, I think, is recognize that in some sorts of cases physics cannot provide the gratifying explanations that the non-physical sciences may be able to provide.

However, even if we assume that psychology cannot be replaced by physics, it is still possible that it could be replaced by neuroscience, or some other "lower-level" science. Neuroscience does not deal in beliefs, desires, intentions, representations or anything that admits of semantic properties, so if psychology could be replaced by neuroscience, there would be no need for appealing to semantic properties to explain human behavior. Of course, we could object to the prospect of neuroscience usurping the domain of psychology in the same way that the non-physical sciences have been defended against the advances of physics: we can claim that neuroscience cannot capture the generalizations that typical psychological laws can capture. However, arguing this way is not so easy in the case of neuroscience. It seems fairly likely that there is an important sense in which physics will miss the generalizations that psychology can capture, but it is not so clear that neuroscience will miss those generalizations--at least, not as long as we are talking only about organisms and not machines. In any case, we will have to wait until both psychology and neuroscience are much more sophisticated to see what the prospects are. This is an empirical question, not a philosophical question.

I am limited to a non-empirical investigation, so, if I am to say something about whether appeals to semantic properties are necessary in order to explain (human) behavior, I can only comment on whether semantic interpretation can provide explanations that cannot be provided without them, or whether the appeal to semantic properties is theoretically superfluous. Particularly, I will discuss whether appeals to semantic properties are necessary given the Classical Computational Theory of Mind. My investigation, then, will be purely *theoretical*.

I do not intend to say anything about what sorts of theories psychologists typically adopt or what sorts of theories psychologists should adopt. I will try to avoid making any empirical claims beyond those about what is to be found in the literature.

NOTES

- 1 True generalizations sometimes contain references to properties of individuals and events that are not relevant to the truth of the law. We have to rule out properties with that sort of "non-essential" role from counting as "being involved in scientific laws."
- 2 Of course, Descartes caused some confusion by maintaining that minds are immaterial and also maintaining that they are located in our pineal glands and that they are involved in bringing about our behavior.
- 3 It might be natural to try to count as physical things and events just those things that have as constituents only things that are referred to in (true) physical theories (or have as constituents things that have as constituents only things that are referred to in true physical theories, etc). Usually this sort of claim is made by saying that every individual and event has a "purely physical description." But there will be problems figuring out when two descriptions count as descriptions of the same individual or event. For example, in the sciences events are usually individuated according to their causal roles in certain systems. Suppose an event that is specified non-physically--for example, Jones believing that such and such--has an extensionally equivalent physical description. Presumably, its causal role will be defined in terms of other non-physically specified events. Those events must also have extensionally equivalent physical descriptions. But, of course, the same problems occur there.
- 4 Mark Wilson [1985] attempts to show that these sorts of properties may indeed be physical properties by giving a sort of "slippery slope" argument: The property *having a temperature of n* is a disjunctive property, and it is a physical property. The property *being hard (to degree n)* is a disjunctive property, and though it is more disjunctive than *having a temperature of n*, the difference is a difference of degree rather than of kind. The same sort of observations can be made about other "multiply instantiated" properties.
Alan Nelson [1985] suggests some ways to draw a principled distinction between various sorts of properties in order to avoid this sort of slippery-slope. He also points out that any position that takes it to be almost trivially true that *is in pain* is a physical property, as Wilson's position does, will reduce the dispute to a merely verbal one.

- 5 The definition that I quote from Kim [1982] is for what Kim call's 'weak supervenience.' Kim [1989] offers criticisms of weak supervenience (and similar attempts to define 'supervene') that are different from mine, though they do center on the definition's failure to capture the relevant dependence relationship. In those same papers he suggests a definition of strong supervenience that is supposed to capture that dependence relationship. However, strong supervenience seems (to me) to require type-type identities of a dubious nature.
- 6 A scientific theory might attribute relational properties to things--such as being a specific distance from something or other--or events--such as causing some other event. That sort of property cannot be physically instantiated* in whatever has that property, but it might be physically instantiated* in something or other.
- 7 If [III*] is true, then the "doctrine of emergence"--the view that some properties of things and events are unpredictable and inexplicable given only information about the physiochemical properties of those things--is false.
Nagel [1961, pp. 368-372] points out that in some sense the doctrine of emergence is truistic: Physical theories can be used to explain no more than their vocabularies express. Presumably, physical theories do not include predicates for properties like *being photosynthetic*, so in a purely physical theory we cannot deduce that certain structures of molecules will be photosynthetic. *Being photosynthetic*, then, is emergent with respect to purely physical theories. But there might still be a theory containing true instantiation laws to the effect that certain structures of molecules will be photosynthetic. *Being photosynthetic* will not be emergent relative to that theory.
- 8 Remember that 'then_c' indicates that the conditional is causal, not material.
- 9 Of course, this ignores many subtleties, such as temporal relations.
- 10 The assumption that we can make sense of equivalent event-descriptions introduces a lot of problems. (See note 3.)
- 11 The phrase is from [Pylyshyn, 1986, chapter 1].

- 12 Gresham's law has been stated in many different ways in many different textbooks. Most standard introductions to economics have versions similar to this one.
- 13 The parallel between explanation and prediction may be misleading here. According to the deductive nomological account of explanation, events are both explained and predicted by deriving statements of those events from statements of initial conditions and at least one law. Thus it may seem as if the only difference between accurate predictions and acceptable explanations is that predicted events are in the future, not in the past. But suppose that Gresham's law has so many exceptions that it yields correct predictions only in about 60% of the cases where the law is supposed to apply. Is it correct to say that because we correctly predict those 60% that we understand why they occur in those cases rather than some other events? Clearly, if we base our predictions only on the observations that certain currencies are undervalued and on applications of Gresham's law, then we cannot distinguish beforehand the cases where the undervalued currency disappears from the cases where they do not.
- 14 J. L. Mackie [1974] argues for this sort of view (with many important refinements).
- 15 There might be a case for MRC--the claim that everything that can be explained by a (true) scientific theory can be explained in the vocabulary of some (true) physical theory--on the basis of these observations about exceptions. If the laws of the non-physical sciences are riddled with exceptions--if they must be amended with great lists of (implicit) provisos--then it may be that there really are no interesting generalizations left to be made by those laws. If, for example, we could include only a pitifully small percentage of states of the world as monetary exchanges, then it is plausible to suppose that neither Gresham's law nor any other economic laws about monetary exchanges are genuinely explanatory (though they may be adequate predictive devices). It may be that all of the non-physical sciences are this way, and that they do not provide explanations. MRC, then, might be vacuously true. But, of course, the best evidence against this sort of argument is that, in every day life as well as in science, we seem to operate quite well explaining events by referring to laws that have exceptions.

CHAPTER 4

FUNCTIONAL THEORIES AND INTERPRETATION

4.1 Introduction

If the loss of generalizations argument is sound, physics cannot completely explain every type of event: We may (in principle) always be able to provide a purely physical account of why an individual event occurs, but we cannot always show what is similar about different events of the same type by appealing to only the laws of physics. Apparently we need chemistry to explain chemical phenomena and biology to explain biological phenomena and neurology to explain neurological phenomena; the different scientific disciplines divide up the explanatory work. In this chapter I will say a little about how different theories can "capture generalizations at different levels." In particular, I will suggest how psychological theories might provide explanations about types of behavior and dispositions that "lower level" theories, such as neurology, cannot provide, and I will discuss the role of semantic interpretation in those explanations.

4.2 Some Preliminaries

When we try to explain human behavior, we are confronted with a "black box problem": Our physical behavior is obvious, and the environment in which our behavior occurs is observable, but, aside from introspection, there is little evidence of what was going on inside our heads. Predictions about human behavior have to be based largely on our relationships with our environment.

People who know very little about computers are also confronted with a black box problem. If we know nothing about what goes on inside of our computers, it seems that the best we can do is describe their dispositions with lists of *input-output laws*. Input-output laws state relationships between the precipitating conditions of a particular system and the easily observable states of that system, without any reference to any mediating states the system might go through. Of course, theories can differ about what types of events their laws are about. One theory, for example, might individuate the relevant types of events in a coarse-grained way, where, say, each punch of the 'k' key on the computer counts as the

same type of input, while another theory might individuate the relevant types of events in a very fine-grained way, where, say, keystrokes on the same key made with different amounts of force count as different types of inputs. The laws can also differ in what the 'then' in them means: They might express causal relations between events of particular types, or they might express only statistical relations between events of particular types.

Input-output laws for computers are analogous to *stimulus-behavior laws* for humans, which describe relationships between the properties of stimuli and the behavior of human subjects. With few exceptions, stimulus-behavior laws for humans will be extremely complicated. How someone responds depends on how they were conditioned earlier, previous experiences, etc. For example, it seems that whether I jump around and yell after I drop a large hammer on my toe will depend on how I have been taught, conditioned, etc. There won't be a generalization about my response to hammer-dropping events, but there may be generalizations about hammer-dropping-on-toe events after Spartan-training-by-father-events and generalizations governing hammer-dropping-on-toe events after sissy-training-by-all-my-aunts-events, etc. Stimulus-behavior laws will usually have huge conjuncts in their antecedents specifying many different sorts of events.

The theories sought by "radical behaviorists" are obvious examples of theories that have only stimulus-behavior laws. Apparently the original motivation for these sorts of theories was a philosophical prejudice against scientific theories that introduced entities or events that were neither "observable" nor specifiable in terms of observable entities or events. When the prejudice was applied to psychology, one type of behaviorism was the result: Suppose a theory that refers to internal states of its subjects succeeds in relating types of stimuli to behavior. We could only know what those intervening states are on the basis of the observations we have made about the subject's history of stimulation and behavior. But then we might as well eliminate the references to the intervening states and relate the subjects behavior directly to its history of stimulation and behavior.¹

This view accompanied some independent positions that were also grouped under the rubric of 'behaviorism.' For example, if we are going to explain the behavior of most people with theories that consist only of stimulus-behavior laws,

then people must be fairly uniform in how they respond to stimuli. Thus the view that the genetic differences between people are usually psychologically unimportant is sometimes referred to as 'behaviorism.' If referring to unobservable phenomena is not allowed, and if we want to account for everyday talk in terms of pains and desires and beliefs and other internal states, then we had better find a way to construe that talk in terms of observable phenomena. Thus 'behaviorism' is used to describe the view that "mental" states like being in pain, and having a belief that such and such, are really "dispositions to respond to stimuli."

Radical behaviorism fell out of favor as the associated forms of behaviorism were abandoned. However, the most compelling reasons to abandon radical behaviorism are perhaps the most obvious. The first is a practical problem: There are too many different possible stimuli (and too many different ways to taxonomize the stimuli) to even get a very good start on discovering the correlations between stimuli and behavior. Certainly it might be fruitful to suggest hypotheses about "unobservable" intervening states because that would provide a way to simplify the complex relationships between stimuli and behavior. The second problem is that stimulus-behavior theories simply do not explain much of what we would like to know about behavior: Even if we did have a complete set of stimulus-behavior laws for humans, or some portion of human behavior, those laws would only specify dispositions. Those dispositions need to be explained. There has to be some account of why people have the dispositions they do have, and that account must refer to the internal states of people.

Suppose that we were going to describe the "behavior" of a computer (running a particular program). We might give a complete account of what the computer would do given any particular sequence of keypunches. That, of course, would be an immense project, but once we were done we would have a list of input-output laws for the computer. But those laws would just specify the "dispositions" of the computer. If we wanted to explain why the computer had those dispositions we would probably try to discover something about how the computer worked on the inside: we would probably try to figure out what "internal" states the computer goes into when certain keys are punched, how the internal states of the computer

are related to each other, and how the "internal" states of the computer are related to what the computer displays on the screen. Thus we would have several more kinds of generalizations about the computer: "input-state laws," which specify the relationships between keypunches and other precipitating conditions and the states of the computer, "state-state laws," which specify the relationships between the states of the computer, and "state-output laws," which specify the relationships between the states of the computer and what it displays on its screen or prints on its printer.

I will divide up psychological laws in roughly the same way: (1) *stimulus-state laws* specify relationships between the properties of stimuli and the "internal" states of people; (2) *state-state laws* specify relationships between a person's internal states; (3) *state-behavior laws* specify relationships between the internal states of people and their behavior. This taxonomy is crude, at best, but it will provide a convenient way to organize the discussion in chapters 6 and 7.

4.3 Explaining Dispositions

Stimulus-behavior laws specify human dispositions, or, in other words, dispositional properties of people. (Recalling part of chapter 2) we can explain how a nondispositional property, *P*, is instantiated in something, *s*, by referring to the other properties of *s*. First we would introduce a *composition law* which specifies the components of the system and the way in which they are organized:

[2.14] *s* has components C_1, \dots, C_2 , and they are organized in manner *O*.

Nomic attributions specify the properties of the components, C_1, \dots, C_2 , and from them we can derive an *instantiation law* of the form

[2.15] For all *x*, if *x* has components C_1, \dots, C_2 , organized in manner *O*, then *x* has property *P*.

From [2.14] and [2.15] we can derive what we wanted to explain:

[2.16] *s* has property *P*.

We can analyze a simple disposition in the same way. Suppose that we were trying to explain why copper conducts electricity. We would first introduce a

composition law to the effect that copper is composed of atoms with so many electrons, neutrons and protons, and we would describe how those atoms bond together. Then we would appeal to an instantiation law to the effect that anything with those components and that organization will conduct electricity. Of course, we would also have to provide an explanation of the instantiation law; it would have to be derived from nomic attributions specifying the tendency of electrons to dissociate from the nucleus of copper atoms and move toward protons which are not already "paired" with electrons, etc.

When we refer to 'tendency of copper to dissociate from the nucleus...' in our explanations we refer to the dispositions of a component to help explain the disposition that we were originally trying to explain. When we are trying to explain how complicated dispositions are instantiated in certain systems, the components mentioned in the composition laws will usually have other, simpler dispositions. Consider vacuum tube triodes, which change the charge of an electric current (figure 4.1).

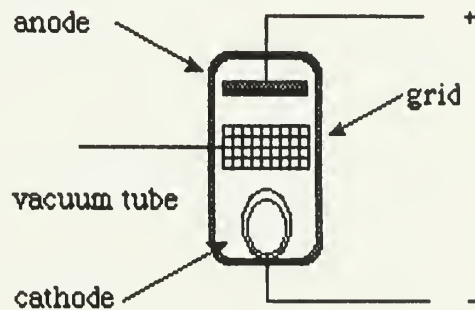


Figure 4.1. A vacuum tube triode.

We can begin to explain why they change the charge of a current by pointing out that they are vacuum tubes that contain a filament, the cathode, and a metal plate, the anode, with an electrode, the grid, in between. The cathode, anode and grid are each attached to separate copper wires. We can also point out that the wires conduct electricity well, and that the cathode emits electrons in a vacuum, while the grid and the anode do not conduct as well, and do not (ordinarily) emit electrons. Of course, when we note the dispositions of the various components of the vacuum

tube triode, we are specifying transition laws for the components that tell us what the final state of each component will be, given an initial state. When we know the transition laws for each of the components and the organization of the components, we can derive what the final states of the vacuum tube triode will be given various initial states: If the grid has no charge and if the cathode has a negative charge with respect to the anode, the cathode will emit electrons and there will be an electric current across the tube. Since the anode does not emit electrons, if it has a negative charge with respect to the cathode, then there will be no current across the tube. So when the grid has no charge, the triode acts as a diode and only conducts electricity in one direction. However, when the grid has a negative charge relative to the cathode, electrons emitted from the cathode are repelled and the flow of electrons from the cathode to the anode is inhibited. When the grid has a positive charge the electrons emitted by the cathode are accelerated towards the anode, and the current is amplified.

Typically, more complicated dispositions have more interesting explanations. Consider the disposition, specified by the "input-output" laws in table 4.1, of a particular circuit, C1, to turn a red light and a blue light on and off depending on the various positions of two switches.

Table 4.1 The input-output laws for circuit C1.

Input		Output	
<u>Top switch</u>	<u>Bottom switch</u>	<u>Red light</u>	<u>Blue light</u>
up	up	off	off
down	up	on	off
up	down	on	off
down	down	off	on

If we took C1 apart we might discover that it looks like seven boxes attached together with wires, as in figure 4.2 (page 67).

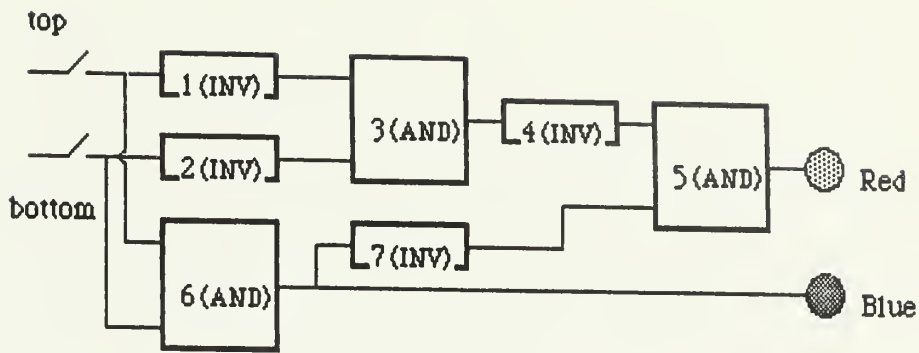


Figure 4.2. A block diagram of the circuit C1 (with batteries, resistors, etc., omitted).

We could begin to explain the disposition of C1 by noting that the switches closed the circuit (and allowed current to flow) when they were down and opened the circuit (and did not allow current to flow) when they were up. By noting the dispositions of the boxes to pass on high or low charges when they were supplied with high or low charges, we could determine whether the lights would turn on or off given how the switches were set. If these subcircuits had the dispositions indicated by their individual input-output laws in table 4.2., then C1 would have the dispositions specified in table 4.1.

Table 4.2. The input-output laws for C1's subcircuits.

3, 5 and 6			1, 2, 4 and 7	
<u>Input</u>		<u>Output</u>	<u>Input</u>	<u>Output</u>
high	high	high	high	low
high	low	low	low	high
low	high	low		
low	low	low		

Since 3, 5 and 6 all have high outputs when and only when they have high inputs through both wires leading to them, it would be appropriate to call them 'AND' circuits. 1, 2, 4 and 7 all give a high output when they receive low inputs, and they give a low output when they have high inputs, so it would be appropriate to

call them 'INVERTERS.' To give a complete explanation of C1's disposition to turn on or turn off lights, we would have to explain the dispositions of each of the components. We would do that by referring to their parts and to the properties of those parts. For instance they might be constructed of vacuum tube triodes, such as in figure 4.3., so we could explain the dispositions of the AND and INVERTER circuits by referring to the dispositions of vacuum tubes.

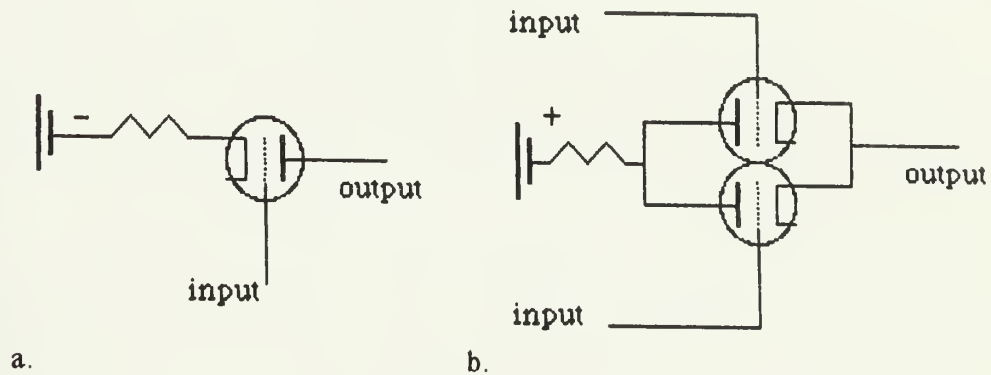


Figure 4.3. A diagram of an INVERTER circuit (a) and an AND circuit (b). Both are constructed of vacuum tube triodes. A high (negative) input to the control grid in the INVERTER circuit prevents electrons from flowing from the cathode to the anode, and results in a low (relatively positive) output. A high input to both control grids in the AND circuit prevents electrons from flowing from the cathode to the anode, and results in a high output. If either control grid receives a low input, the flow of electrons is amplified and the output is low.

The dispositions of the vacuum tubes can be explained by showing how those dispositions are instantiated in the properties of their components, as I suggested earlier. The properties of those components can in turn be explained, until finally it has been shown how the disposition of C1 is physically instantiated*. (Remember chapter 3.)

4.4 Functional Theories

It will be easier to appreciate talk about "levels of explanation" if we consider another circuit, C2, which has the same dispositions as C1, has components organized just like those of C1, and has components that have exactly the same dispositions as C1's components because they are constructed of vacuum tube triodes

just as C1's components are constructed. Regardless of all of these similarities, C1 and C2 will have some physical differences--their components will have slightly different molecular structures. However, it is obvious that the difference between C1 and C2 is irrelevant to the explanation of why the vacuum tube triodes have the same dispositions to carry current, and why all of the INVERTERS and AND circuits have the dispositions specified in table 4.2, and why C1 and C2 have the same dispositions to turn on and off the lights. We cannot explain the similarity between C1 and C2 in a purely physical vocabulary.²

Consider another circuit, C3, which has the same dispositions as C1 and C2, has components organized just like those of C1 and C2, has components that have (roughly) the same dispositions to carry current as the components of C1 and C2, but which has components constructed out of semiconductor diodes instead of vacuum tube triodes. A physical description of C3 would be very different from C1 and C2. C3's wiring diagram would look different from the wiring diagram of C1 and C2, and the voltage of the inputs and outputs of the semiconductor triodes would be different from those of the vacuum tube triodes. Because of the differences we could not explain the dispositions of the subcircuits of C1, C2 and C3 in the same way. However, the differences in the construction of the subcircuits are irrelevant to the explanation of why C1, C2, and C3 all have the same dispositions to turn lights on and off: any circuit that has subcircuits with the same dispositions as 1 - 7, organized in the same way as they are in C1 will have the same dispositions as C1.

This distinguishes C1, C2 and C3, from C4, which has the same dispositions as C1, C2 and C3, and which is composed of INVERTERSs and AND circuits, but which has its components organized as figure 4.4 illustrates (page 70).

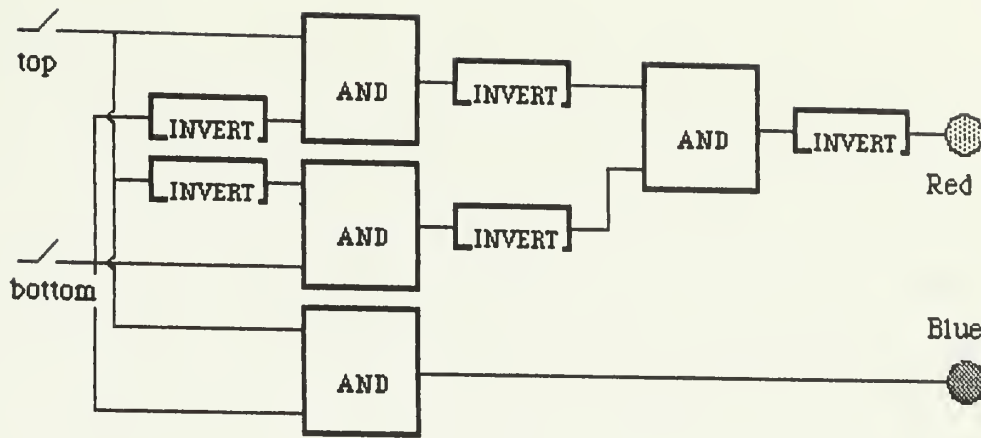


Figure 4.4. A block diagram of C4.

There is obviously something common to the explanation of the dispositions of C1, C2 and C3 that is not common to the explanation of C4's dispositions.

The similarities between C1, C2 and C3 are often called "functional" similarities. C1 and C2 are physically different, and C3 is electronically different from C1 and C2, but they are all "functionally equivalent" at the electronic level because, in some sense, their electronic components have the same functions. C4 has the same dispositions as the other circuits, but it does not seem to be functionally equivalent with any of them at any level.

Functional theories exploit the sorts of relationships that these comments about "functional equivalence" suggest. They posit the existence of certain sorts of internal state-types and posit relations between those states and the possible inputs, between those states and the possible outputs, and between the internal states themselves. If they are correct, then they will provide correct predictions of the outputs given the inputs, in addition to providing some explanation of why those input-output relationships exist.

Consider a system, s_j , that has well defined inputs and outputs, and a set of well defined internal states, and suppose that we have a theory, T, about the relevant causal or statistical relations between those states. Presumably the language of T includes some predicates, ' I^1, \dots, I^j ', that express the properties by which the (possible) types of inputs are individuated, some predicates, ' O^1, \dots, O^k ', that express the properties by which the (possible) types of outputs are

individuated. If we knew enough about the internal states of the system, it would also include some predicates, ' S^1, \dots, S^n ', that express the properties by which the (possible) internal state-types are individuated--perhaps physical properties, biological properties, or electronic properties, etc.³ Let ' \rightarrow ' represent the transitional relations (causal or statistical) between the types of inputs and outputs, and the internal state-types. Thus T would look something like this:

$$[1] \quad [(I^1 s_j \rightarrow S^1 s_j) \& (S^1 s_j \rightarrow S^2 s_j) \& (S^2 s_j \rightarrow O^1 s_j) \dots].$$

As a convenient shorthand, let's represent [1] with,

$$[2] \quad T(S^1 s_j, \dots, S^n s_j).$$

If we replace the occurrences of ' S^1, \dots, S^n ' with variables ' p_1, \dots, p_n ' *ranging over properties or types*, and if we replace the occurrences of ' s_j ' with a variable over systems, we get the open sentence,

$$[3] \quad T(p_1 s, \dots, p_n s).^4$$

[3] specifies a number of functional roles: a state-type performs the i^{th} functional role *in* α , if it is the i^{th} member of an n -tuple that satisfies [3] when all of the occurrences of ' s ' in [3] are replaced by occurrences of ' α '. For example, a state type, S^1 , performs the 1st functional role *in* s_j if [2] is true. (Though it is common to talk about state-types performing functional roles without mentioning the system in which they perform that role, it is important to notice that state-types might perform different functional roles in different systems.) [3] also specifies a number of ("second order") functional states: a system, α , is in the i^{th} functional state if it has a property that is the i^{th} member of an n -tuple that satisfies [3] when all of the occurrences of ' s ' in [3] are replaced by occurrences of ' α '. If we bind all of the variables in [3] with existential quantifiers, we get a statement of this form:

$$[4] \quad \text{There are } p_1, \dots, p_n \text{ such that } T(p_1 s_j, \dots, p_n s_j).$$

[4] entails that there are "first-order" state-types that have functional roles *in* s_j in virtue of which s_j 's inputs and outputs stand in the particular relationships to each other; it does not entail anything about what those state-types are. This sort of claim can be important when we do not know the exact nature of the states that intervene between the inputs and outputs of a particular system.

[4] is a claim about one particular system, s_j . We might also want to claim of several theories that they each satisfy this open sentence:

[5] There are p_1, \dots, p_n such that $T(p_1s, \dots, p_ns)$.

In other words, we might want to claim of several systems that they each have some state-types that perform particular functional roles in those systems. If the following is true,

[6] For every system s , if Fs , then there are p_1, \dots, p_n such that $T(p_1s, \dots, p_ns)$,

then all the systems that are F are "functionally equivalent." Of course, [6] *does not* entail that there is a state-type that performs the same functional role in each of the systems. When philosophers refer to "functional theories," they usually mean theories of the same form as [6].⁵

C1, C2, and C3 satisfy this open sentence

[7] There are $p_1 \dots p_{20}$ such that $T(p_1s, \dots, p_{20}s)$,

where ' $T(p_1s, \dots, p_{20}s)$ ' is a theory about the relationships between the four possible states of the switches and the four possible states of the lights. The relevant internal states will be the twelve different states of the wires that lead into and out of the various circuits--states of the wires having either a high or low charges--and the transitions that are mentioned by the theory will specify the dispositions of the subcircuits. Since C1, C2 and C3 each satisfy [7], they are functionally equivalent.⁶ C4 does not satisfy [7].

Functional theories are useful for illustrating the similarities and differences between complex systems. The notion of functional equivalence allows us to state generalizations that we could not otherwise state. Beyond the input-output laws for C1, C2 and C3, they don't have very much in common: They are not of the same physical type--if they any of them are of any physical type at all--and they are not all of the same electronic type. But we can explain why they have the same input-output laws: they are functionally equivalent. This is the same sort of observation that we make about different computers that have the same capacities. The computers may be different physically, but it is their functional equivalence which explains why they have the same capacities.

I should make a relatively simple observation that is often overlooked in the literature: Functional theories can vary a great deal in the way that a system's inputs and outputs are individuated and in their specificity concerning the transitions between the internal states that connect those inputs and outputs. Thus claims about the "functional equivalence" of two systems only make sense relative to a particular functional theory. For example, suppose that C1 - C3 were parts of three different machines, M1 - M4, which have the same sorts of (relevant) dispositions. When we tried to explain why those machines have similar dispositions we would refer to the dispositions of their components, including C1 - C3. But we might suggest two different sorts of theories for M1 - M3: we might suggest a functional theory that describes the dispositions of C1 - C3 as transitions between certain state-types, say switches being up or down and lights being on or off, but which ignores transitions between the internal state-types of C1 - C3; or we might suggest a functional theory for M1 - M4 that doesn't ignore the internal state-types of C1 - C3. In other words, we might suggest a theory that does not entail that C1 - C3 satisfy [7], or we might suggest a more detailed theory that does. Now, suppose that M4 is like M1 - M3 except that in it C4 does what C1 - C3 do in M1 - M4. Relative to the less specific functional theory M4 is functionally equivalent to M1 - M4, while relative to the more detailed theory M4 is not functionally equivalent to M1 - M4.

Because functional theories are neutral about which types perform the functional roles they specify, we don't need to know nearly as much to explain the

dispositions of systems as we would have to know if we were going to try to describe the internal events of each token of each type in a completely physical, chemical or neurological vocabulary. This makes functional theories useful in solving black box problems where a system receives many different inputs (or types of stimuli) and provides many different outputs (or types of behavior), but where we know little about what is going on in between.

4.5 Interpretation

Inputs and outputs can be described in a number of different ways. Sometimes, such as when one is typing commands into a computer, it is natural to think of the inputs and outputs as sentences, or, more precisely, as tokens of sentences, not just as sequences of keypunches or particular sorts of electrical activity. In fact, we often speak of computers as "information processors" and in doing so we imply that computers manipulate not only sentence-tokens, but the "information" those tokens encode. Calculators calculate mathematical functions on numbers, so it is natural to think that the inputs and outputs represent numbers. In fact people usually talk as if calculators manipulate numbers, and not just the "numerals" that represent numbers.

Although electrical engineers may describe the internal functions of computers in electronic terms, the rest of us usually describe them in terms of how they process information. We usually use flow-charts or programs to show how computers take certain sorts of information and produce other sorts of information: A disposition of a computer, described in terms of the numbers or information it has as inputs and output, is analyzed into other dispositions to manipulate numbers or information. For example, suppose we described the possible inputs for C1 - C4 as truth values of sentences: The top switch being down would "represent" the sentence A , the top switch being up would represent $\sim A$, and the possible positions of the bottom switch would similarly represent B and $\sim B$. In that case, if the red light being on represented $\sim(A \& B) \& (A \vee B)$, the red light being off represented $\sim(\sim(A \& B) \& (A \vee B))$, and the blue light being on and off similarly represented $A \& B$ and $\sim(A \& B)$, then C1 - C4 would be described by the input-output laws indicated in table 4.3 (page 75).

Table 4.3. The inputs and outputs of C1 - C4 interpreted as sentences and compound sentences.

Input		Output	
<u>Top switch</u>	<u>Bottom switch</u>	<u>Red light</u>	<u>Blue light</u>
$\sim A$	$\sim B$	$\sim(\sim(A \& B) \& (A \vee B))$	$\sim(A \& B)$
A	$\sim B$	$\sim(A \& B) \& (A \vee B)$	$\sim(A \& B)$
$\sim A$	B	$\sim(A \& B) \& (A \vee B)$	$\sim(A \& B)$
A	B	$\sim(\sim(A \& B) \& (A \vee B))$	$A \& B$

C1 - C4 might very well be designed to make these sorts of inferences. We can see why C1, C2 and C3 have this disposition by considering the dispositions of their components. We might point out that the sentences are represented by high and low charges which are input into the various components, etc., just as we did before. But it is easier to think of the components of the circuits as if they make simple inferences. The subcircuits 3, 5 and 6 can be thought of as calculating the truth value of $A \& B$ given the truth values for both A and B , and the subcircuits 1, 2, 4 and 7 can be thought of as calculating the truth-value of $\sim A$ given the truth value of A . By noting the dispositions of the subcircuits, and the connections between the subcircuits, it is easy to see why C1, C2 and C3 make the inferences that they make. Of course, we will not be able to explain the dispositions of the subcircuits by attributing information processing capacities to them. Instead we will have to point out that high and low charges of the wire running into and out of the subcircuits represent particular sorts of sentences, and then we would have to explain why the subcircuits have their particular dispositions to carry high and low charges.

C1 - C4 seems a little more useful when we interpret the inputs as single digit binary numbers 0 and 1, and the outputs as the digits of a two digit binary number: Either switch being down would represent the number 1, and either switch being up would represent the number 0. In that case, if the red light being on represented the second digit of a two digit binary number being 1, if the red light being off represented the second digit of a two digit binary number being 1, and if

the blue light being on and off similarly represented the values for the first digit, then C1 - C4 would be described by the input-output laws indicated in table 4.4.

Table 4.4. The inputs and outputs of C1 - C4 interpreted as binary numbers.

Input		Output
<u>First #</u>	<u>Second #</u>	<u>First # + Second #</u>
0	0	00
1	0	01
0	1	01
1	1	10

C1 - C4 might very well have been designed as a binary digit adders. We can explain why they are successful binary digit adders by referring to the relationships between their subcircuits and the dispositions of those subcircuits to perform elementary mathematical functions on the numbers 1 and 0.

4.5.1 Interpretive Functional Theories

In these two examples the dispositions of C1 - C3 were explained by interpreting the inputs and outputs and then breaking the dispositions down into other dispositions that are themselves interpreted. When input and outputs are associated with sentences, propositions or numbers, the functional relationships between the inputs, the unseen internal states, and the outputs are "mirrored" or "mapped" by the syntactic, semantic or mathematical relationships between the sentences, propositions, or numbers. This map would be a function taking sentences, propositions, or numbers to physical state-types of the systems in question. I will call functional theories of this sort 'interpretive functional theories.'

Suppose that we are considering only theories that map sentence-types to physical state-types, and that ' A_1, \dots, A_k ' are schematic letters. Then an interpretive functional theory that mapped sentences to physical state-types might look like this:

[8] For every s , if Fs , then there is an f such that $T(f(A_1)s, \dots, f(A_k)s)$.

However, it might not be that simple. If for each system, a single function mapped sentences to internal states, then for each system only one state-type could be associated with each sentence. But we might want to associate a state-type with one sort of functional role with a sentence, and also associate a different state-type with a different sort of role with that same sentence. For example, most belief/desire theories insist that people can believe and desire the same thing, but to believe something and to desire something are certainly different; one function might map sentences to "belief states" and another might map those same sentences to "desire states." Thus interpretive functional theories might quantify over several functions, two of which might map the same sentence to different physical state-types. I will represent them this way:

[9] For every s , if Fs , then there are f_1, \dots, f_m such that $T(f_1, \dots, f_m)$.⁷

4.5.2 A Short Defense

I have suggested how references to "semantic" entities might be used in explanations of complex dispositions. Interpretive functional theories exploit the similarities between the properties of sets of propositions or sentences and the properties of systems. For example, under an appropriate mapping, the causal relations between certain states of a particular sort of system might mirror the inferential relationships between sentences. The logical status of sentences can thus be used to pick out certain properties of the system in question. That is why there is nothing mysterious about explanations that refer to our relations with abstract objects like propositions and sentence-types--at least nothing more mysterious than the abstract objects themselves. In fact, explanations that interpretive functional theories provide are analogous to scientific explanations that employ numbers. Consider, for example, Robert Stalnaker's explanation of why referring to numbers helps us explain physical phenomena:

What is it about such physical properties as having a certain height or weight that makes it correct to represent them as relations between the

thing to which the property is ascribed and a number? The reason we can understand such properties--physical quantities--in this way is that they belong to families of properties which have a structure in common with the real numbers. Because the family of properties which are *weights* of physical objects has this structure, we can, (given a unit, fixed by a standard object) use a number to pick out a particular one of the properties out of the family. That, I think, is all there is to the fact that weights and other physical quantities are, or can be understood as, relations between physical objects and numbers.[1984, p. 9]

Interpretive functional theories are useful because they can demonstrate what systems have in common without making any claims about the physical properties of those systems. They also show why it is that we can think of purely material things as information processors. I have discussed interpretive functional theories with the aim of showing how "interpretation" might help us to explain the behavior of complex systems, not as a suggestion about how we might "analyze" the propositional attitudes. If we use an interpretive functional theory to explain the behavior of a system, we are not committed to "functionalism" or any other view about what "propositional attitudes really are." For example, suppose that there is an interpretive theory, T, that quantifies over functions from propositions to state-types. Let's suppose that one of those functions, \mathcal{B} , is specified as the "belief-relation." Consider this version of "propositional-functionalism":

- [10] For every subject A and every proposition P , being a belief that P = being a token of some state-type that has the functional role $\mathcal{B}(P)S$.⁸

According to [10] beliefs are *the very same things as* tokens of types that have particular functional roles. Thus, to say that A believes that P is to say the very same thing as that A has a property that has a particular functional role, or to say that A is in a particular functional state.⁹ But if our theory T is true, all we know is that whenever anyone stands in the "belief-relation" to a particular proposition, P , they are in a state that has a particular functional role. It is perfectly consistent with T that in each case where a person believes that P the very same neural-state performs the relevant functional role and that believing that P is identical to being

in a particular neural state; it is consistent with T that in each case where a person believes that P a certain type of "mental sentence" performs the relevant functional role, and that believing that P is identical to having a certain relation to that mental sentence; it is consistent with T that in each case where a person believes that P God intervenes in a certain way, and that believing that P is identical to having been divinely influenced in a particular way. Thus commitment to an interpretive functional theory does not entail commitment to a particular view about the nature of propositional attitudes.

Finally, the best evidence that semantic entities enter into explanations as they are exploited by interpretive functional theories--as indices of internal functional states--is that there is simply no other way to explain our success in providing "intentional" explanations. We are "black boxes," so we do not know much about the mechanics of what goes on inside our heads. If propositional attitudes are to be analyzed in terms of neural states, relations to "mental sentences," or in any other sort of manner where we cannot tell what psychological state people are in simply by looking, then we can refer only to the functional relations between those states and what we can observe. If using propositional attitude ascriptions did not allow us to do that, they would be irrelevant in explanations of behavior.

NOTES

- 1 Versions of this sort of argument can be found in [Skinner, 1953, pp. 27-35] and [Hempel, 1958].
- 2 Thus, the dispositional property to carry current that C1 and C2 both have is "emergent" relative to physical theory but not emergent relative to a theory stated in a vocabulary that allows references to electronic components. (See note 7, chapter 3.)
- 3 The generality of our theory will depend on how we individuate these state-types.
- 4 If quantifying over properties is not allowed, one can "name" the relevant states and then quantify over "things." See, for example, [Lewis, 1970] or [Loar, 1981].
- 5 It is important to distinguish functional theories from slightly different sorts of theories.

[6'] There are p_1, \dots, p_n , such that for every system s , if Fs , then
 $T(p_1s, \dots, p_ns)$

[6'] *does* entail that there are some state-types that perform the same functional role in each of the systems. Of course, theories of this sort can vary greatly about what the variables ' p_1, \dots, p_n ' range over. It would be appropriate to call them 'strong functional theories.'

- 6 [7'] *does not* guarantee that each of the state-types that fulfill particular functional roles for C1 are the same as the state-types that fulfill those roles for C2 or C3. The following strong functional theory (see note 5) *does* guarantee that each of the state-types that fulfill particular functional roles for C1 are the same as the state-types that fulfill those roles for C2 and C3:

[7'] There are $p_1 \dots p_{20}$ such that for all s , $s \in \{C1, C2, C3\}$,
 $T(p_1s, \dots, p_{20}s)$.

[7'] is not true because C3 is constructed out of semiconductor triodes instead of vacuum tube triodes, and so its subcircuits have different input and output voltages than those of C1 and C2. Thus C1, C2 and C3 are not "strongly

functionally equivalent." If the reference to C3 was removed, [7'] would be true--so C1 and C2 are strongly functionally equivalent.

7 As far as I know, Brian Loar [1981] was the first to explicitly characterize cognitive theories as what I call 'interpretive functional theories.'

8 See [Loar, 1981] and [Schiffer, 1987, pp. 24-48] for discussions of this type of theory.

9 It is worth noting that what seems to be one of the most common objections to functionalism--that it entails the "orthographic accident" view of propositional attitudes--does not apply to functionalist theories stated in terms of interpretive functional theories. See [Loar, 1981].

CHAPTER 5

THE CLASSICAL COMPUTATIONAL THEORY OF MIND

5.1 Cognitive Architecture

In the previous three chapters I sketched the rudiments of scientific explanation, discussed an argument for why physics cannot explain everything that can be explained, sketched how the dispositions of systems are typically explained, and suggested how semantic entities like sentences and propositions can be used to provide explanations that we might not be able to give in physics, electronics, neuroscience, etc. The examples I gave in chapter 4 were intended to illustrate how the dispositions of circuits could be explained by "interpreting" the dispositions of the circuits and subcircuits as dispositions to manipulate "information." Equivalent explanations could not always be given in the vocabulary of physics or electronics.

Does semantic interpretation have any essential role in explanations of human behavior? The answer seems to depend on how we individuate human behavioral events: If there are ways to individuate behavioral events such that there are "important" generalizations about them that cannot be stated in physical, chemical, biological, or neural terms, then we will have to explain them at a different "level." Cognitive psychology (and most of our everyday "folk" psychology) is based on the assumption that we cannot make sense of many or most of our dispositions without recognizing that there is a "representational" level of explanation. Cognitive theories often refer to human dispositions or capacities to process certain sorts of information, whether the information is "encoded" in "mental sentences," is contained in images, or is related to human functional states in some other way. The successes that cognitive psychologists have had explaining these dispositions is *prima facie* evidence that there is a representational level of explanation.

The task of cognitive psychology is to provide explanations of certain sorts of human behavior. But explanations of complex cognitive behavior cannot be given in a theoretical vacuum: Successful cognitive explanations both presuppose

and inform us about some of the possible answers to questions like 'What sorts of semantic entities are involved in cognitive processes?', 'What properties of those semantic entities involve them in cognitive processes?', and 'how are those properties exploited to produce complex cognitive capacities?' For example, there are significant differences between a view that presupposes the existence of only "mental images" and one that presupposes only "mental sentences." Our success in explaining our linguistic capacities may depend on what view we take. Thus answering questions about our "cognitive architecture" is part of the explanatory process in cognitive psychology.

5.2 The Computational Theory of Mind

Computational models of cognition are not cognitive theories in the sense that they are primarily intended to predict or explain specific sorts of human behavior. Instead they are theories about the sort of cognitive architecture humans have, and thus about what adequate cognitive theories will look like. They appear in many different guises with many different labels. Some are distinguished by the implicit or explicit use of an analogy between humans and (digital) computers, and all make use of the difficult notion of "computing." The view that Jerry Fodor has defended as *the* Computational Theory of Mind (CTM) is, perhaps, the "received view" about our cognitive architecture--among philosophers at any rate. In the next two sections I will sketch CTM and a stronger version of CTM, the Classical Computational Theory of Mind (CCTM), which has been vigorously defended by both Fodor and Zenon Pylyshyn.¹

5.2.1 Some Motivation

Fodor [1975, pp. 28-34] begins one argument for CTM by "assuming" that anyone reasonable will accept that filling in the details of the following (idealized) model will provide explanations of some human behavior:

8. The agent finds himself in a certain situation (S).
9. The agent believes that a certain set of behavioral options (B_1, B_2, \dots, B_n) are available to him in S ; i.e., given S , B_1 through B_n are the things that the agent believes that he can do.

10. The probable consequences of performing each of B_j through B_n are predicted; i.e., the agent computes a set of hypotheticals of roughly the form if B_j is performed in S , then, with a certain probability, C_j . Which such hypotheticals are computed and which probabilities are assigned will, of course, depend on what the organism knows or believes about situations like S . (It will also depend upon other variables which are, from the point of view of the present model, merely noisy: time pressure, the amount of computation space available to the organism, etc.)
11. A preference ordering is assigned to the consequences.
12. The organism's choice of behavior is determined as a function of the preferences and probabilities assigned. [1975, pp. 28-29]

Of course, behaviorists deny Fodor's assumption. That, according to Fodor, is a problem for the behaviorists: 'What everyone knows, but the behaviorist's methodology won't allow him to admit, is that at least some actions are choices from among a range of options contemplated by the agent' [1975, p. 33].

If our choices are correctly viewed as responses to (contemplated) possible outcomes, then, according to Fodor, we are committed to something like the model he suggests. That model entails that, in some sense, people "process information," "make inferences," or "perform computations." If we take the model as more than a mere heuristic device, then to explain the choices people make, it seems that we have to refer to specific relations in which people stand to "information bearing," "meaningful," "contentful" things like propositions, sentence tokens, sentence types, etc., and there must be certain computational or inferential relations between those things. Of course, this rather vague sort of claim does not significantly narrow down the competition, but those who would follow Fodor here are committed to a "relational" view of cognitive states that makes use of "propositional attitude" ascriptions such as 'A believes that φ ,' 'A fears that φ ,' and 'A desires that φ ,' in explanations of behavior.²

Fodor [1975, pp. 34-41] argues that several obvious facts about "concept learning" reinforce his claim. Concept learning is typically investigated by placing subjects in specific environments and then rewarding or punishing them for their responses to experimentally manipulated stimuli. The subjects are thus faced with determining environmental conditions under which particular

responses are appropriate. To do that the subjects have to determine both the "criterial properties" of certain stimuli and the appropriate responses to those stimuli; in effect, they have to learn the "concept" which characterizes the stimuli and the "concept" which characterizes the appropriate response. Since the rate of learning is influenced by the character of the appropriate response, the character of the reinforcement, the nature of the subject population, etc, a good deal of information about human learning has been obtained simply by changing the values of the variables in a variety of different experiments.³

According to Fodor,

What the organism has to do in order to perform successfully is to extrapolate a generalization (all the positive stimuli are P-stimuli) on the basis of some instances that conform to the generalization (the first n positive stimuli were P-stimuli). The game is, in short, inductive extrapolation, and inductive extrapolation presupposes (a) a source of inductive hypotheses (in the present case, a range of candidate values of P) and (b) a confirmation metric such that the probability that the organism will accept (e.g., act upon) a given value of P at t is some reasonable function of the distribution of entries in the data matrix for trials prior to t . [1975, p. 37]

So far as anyone knows, concept learning is essentially inductive extrapolation, so a theory of concept learning will have to exhibit the characteristic features of theories of induction. In particular, concept learning presupposes a format for representing the experimental data, a source of hypotheses for predicting future data, and a metric which determines the level of confirmation that a given body of data bestows upon a given hypothesis.[1975, p. 42]

According to Fodor, if we do indeed make inductive inferences in concept learning, then, in some sense, we "process information," or "perform computations" If we take this sort of language to be anything more than metaphorical, it seems that we have to refer to specific relations in which people stand to "information bearing," "meaningful," "contentful" things like propositions, sentences tokens, sentences types, etc., and there must be certain computational or inferential relations between those things.

5.2.2 The Theory

It seems that if attributions made using such locutions as 'A believes that φ ,' 'A fears that φ ,' and 'A desires that φ ' are true, then people must stand in believing, fearing and desiring relations to the sorts of things that can be believed, feared and desired.⁴ Propositions are leading candidates for members of the range of those relations, so ascriptions using the 'believes that,' 'fears that,' and 'desires that' locutions are usually called 'propositional attitude ascriptions.' Fodor accepts this principle as the obvious starting point for any theory about psychological explanation:

CTM 1: Cognitive states are identified via propositional attitude ascriptions.

Propositions are not the only candidates for the range of "propositional attitudes," so Fodor is faced with saying exactly what sorts of things we are related to when we are carrying out our cognitive processes. Unfortunately, it is here that Fodor's view begins to get difficult. Fodor [1978b] argues at some length that if we are to explain behavior we must refer to propositional attitudes--not just propositional attitude ascriptions, but attitudes about *propositions*. Presumably, then, we would predict and explain behavior by using those propositions as external indices for internal states with particular functional roles (as in an interpretive functional theory).

Usually, however, Fodor claims that cognitive states are relations to *representations*. In many places Fodor makes claims that either explicitly or implicitly entail that representations have physical properties, in which case they must be material things. This accords with the way we usually use the word 'representation': Pictures are representations of the way the world is or is not; strings of inscriptions on paper and utterances can also represent the way the world is or is not. Those are material things, and so it is natural to think of representations as material things, and not as abstract objects such as propositions. Finally there are the many places where Fodor seems to identify having a propositional attitude with standing in a certain relation to (or "storing") an

"internally inscribed formula." If they are inscribed anywhere, then they must be material things.

Fodor has tried to square up these two different accounts:

Now, this may suggest the following ontological picture: There are, as it were, *two* things--the organism's relation to propositions and the organism's relation to formulae--and these two things are so arranged that the latter is causally responsible for the former (e.g., the organism's being in a certain relation to formulae causes it to be in a certain relation to propositions.) I can imagine that someone might want to resist this picture on metaphysical grounds; viz., on the grounds that it takes propositions (or, anyhow, relations to propositions) as the bedrock on which psychology is founded. [1975, pp. 77]

At this point, I'm not particularly worried about the ontological status of propositions. But there being '*two* things--the organism's relation to propositions and the organism's relation to formulae' is a problem. Consider Fodor's comment that 'the organism's being in a certain relation to formulae causes it to be in a certain relation to propositions.' If that were true--if being-in-a-certain-relation-to-a-formula-events are different from being-in-a-certain-relation-to-a-proposition-events, and if the former events cause the latter--then which beliefs, fears and desires we have would not depend causally on our other beliefs, fears and desires, but just on our various relations to formulae. That sort of epiphenomenalism seems to make our attitudes to propositions irrelevant in psychological explanations, and I doubt that Fodor really means to support such a view.

Fodor suggests a slightly more palatable view:

The present point is that one *can* resist this picture while adhering to the account of psychological explanation that I have been proposing. In particular, one might take the basic explanatory formulae as expressing (not causal relations to formulae and relations to propositions but) contingent event identities. That is, one might think of cognitive theories as filling in explanation schema of, roughly, the form: *having attitude R to proposition P is contingently identical to being in computational relation C to the formula (or sequence of formulae) F.* [1975, p. 77]

First of all, and least important, this suggestion does nothing to relieve the ontological worry, since there are still relations to propositions on this view. Second, and more important, the notion of "contingent identity" is notoriously hard to understand. In some instances phrases of the form 'A is contingently identical to B' have been used in ways that seem to entail that an event (under one description) A is the same as an event (under another description) B, but that the one thing might have been two different events. I do not understand that. I suspect that the notion of contingent identity gains much of its appeal from a particularly insidious confusion: Events (or states of systems), as they are typically individuated in science, can be described in a number of ways. For example, the firing of a type-A nerve in my leg might be the very same event as a signal from my brain to my leg. Of course, the property *being a firing of a type-A nerve* is certainly different from the property *being a signal from my brain to my leg*. This observation might lead one to employ something like the following fallacious reasoning: Each firing of a type-A nerve has to be a firing of a type-A nerve, but firings of type-A nerves might have been something other than signals from brains to legs. So, though a firing of a type-A nerve might be the same event as a signal from my brain to my leg, it might not have been. In that case, the signal from my brain to my leg would have been a different event, thus the two events are "contingently identical." The correct conclusion, I think, is that if a particular type-A nerve firing were not a signal from my brain to my leg, then it simply would not have been a signal from my brain to my leg. That job might have been performed by some other event, but the event in question could not have been *two* different events.⁵

Some sort of similar confusion infects Fodor's comments. Instead of invoking "contingent identity," it would be more cogent to point out that, though it is not *necessarily* true that bearing-a-particular-relation-to-the-proposition-expressed-by-an-internal-formula-states are all bearing-a-particular-relation-to-that-internal-formula-states, all states of the first kind are also states of the second kind, and that fact is lawlike. If this sort of observation is supposed to solve the ontological worry about propositions, then, presumably, Fodor thinks that if there

is this close connection between relations to propositions and relations to formulae, the relations to propositions are irrelevant to psychological theorizing.

Unfortunately, even Fodor's most explicit writing on the role of propositions in psychological processes leaves his views uncertain. Here, for example, Fodor sometimes writes about the connection between relations to propositions and relations to formulae (without invoking the notion of contingent identity):

At the heart of the theory is the postulation of the language of thought: an infinite set of 'mental representations' which function both as the immediate objects of propositional attitudes and as the domains of mental processes. More precisely, [The Computational Theory of Mind] is the conjunction of the two following claims:

Claim 1 (the nature of propositional attitudes):

For any organism O , and any attitude A toward the proposition P , there is a ('computational'/'functional') relation R and a mental representation MP such that

MP means that P , and
 O has A iff O bears R to MP .

(We'll see presently that the biconditional needs to be watered down a little; but not in a way that much affects the spirit of the proposal.)

Its a thin line between clarity and pomposity. A cruder but more intelligible way of putting claim 1 would be this: To believe that such and such is to have a mental symbol tokened in your head in a certain way...[1987, p. 17]

Claim 2 (the nature of mental processes):

Mental processes are causal sequences of tokenings of mental representations.

A train of thoughts, for example, is a causal sequence of tokenings of mental representations which express the propositions that are the objects of thoughts. [1987, p. 17]

All of this may seem clear at first glance, but I find it very confusing. Fodor begins by telling us that mental representations function as 'the immediate objects of propositional attitudes.' Are propositions supposed to be more distant objects of propositional attitudes? He then tells us that to believe something is to 'have a mental symbol tokened in your head in a certain way...' Now, it may be true that whenever one believes P one has a mental symbol that means P tokened in one's head. But is believing that P the very same thing as having a mental symbol that means P tokened in one's head? Could someone believe that P without having a mental symbol that means P tokened in one's head?

Fodor then claims that mental processes are sequences of tokenings of mental representations. (He apparently means to say that mental processes are sequences of relations to tokenings of mental representations.) This seems to imply that mental states are involved in mental processes because they are relations to formulae, not because they are relations to propositions. At the end of the passage Fodor seems to commit himself to both views: a 'train of thoughts' is a sequence of relations to formulae, but the thoughts themselves have propositions as their objects.

This may seem like nit-picking, but surely it is important nit-picking. Cognitive states will be subsumed under laws in virtue of their properties. It seems to me that laws which subsume states because they are relations to propositions are significantly different from laws that subsume states because they are relations to representations. After all, if cognitive laws subsume states because they are relations to propositions, then things that lack such mental sentences don't seem to have cognitive processes--no matter how clever they are. I'm not sure how to read Fodor on this point, so I will do what all of his other commentators seem to do and ignore the problem.⁶ For the purposes of this dissertation I will assume Fodor holds that a sentence of the form '*A* believes that φ ' is true just in case *A* bears a particular (computational/functional) relation to a representation that "means" that φ .

CTM 2: Cognitive states are relations to representations.

The following claim is the third principle of CTM:

CTM 3. The Content Individuation Principle:

Types of representations are individuated according to 'that'-clauses on "opaque" readings.⁷

Let us say that an opaque reading of a 'that'-clause is a reading on which co-designative expressions cannot always be substituted without changing the truth-value of the "content" sentence. Suppose, for example, John sees a man with the brim of his hat pulled down over his face and, supposing that anyone who wears his hat that way is a spy, he correctly decides that the man is a spy. We could say that

[1] John believes that the man with his hat pulled down is a spy.

This would in some way characterize what John believes, and indicate to us how he might behave in certain situations. Since John has no other beliefs about the man in the hat, and since he feels no patriotic obligations, he does not worry about it. Now suppose that the man in the brown hat is actually John's father. If we were allowed to substitute co-designative expressions in 'that'-clauses of cognitive state attributions, we could say that

[2] John believes that his father is a spy.

On a transparent reading both [1] and [2] are true. However, John would never assent to the sentence 'My father is a spy,' and he would not treat his father with suspicion: On an opaque reading of the 'that'-clause [1] is true and [2] is false. On a transparent reading [2] does not tell us very much about John's dispositions. We have to know whether John "knows" that his father is the man with his hat pulled down before we can make any informed predictions about his behavior on the basis of [2].

Fodor defends the Content Individuation Principle in the following way:

...when we articulate the generalizations in virtue of which behavior is contingent upon mental states, it is typically an opaque construal of the mental state attributions that does the work; for example, it's a construal

under which believing that *a is F* is logically independent from believing that *b is F*, even in the case where $a = b$. [1980, p. 66]

Opaque ascriptions are true in virtue of the way that the agent represents the objects of his wants (intentions, beliefs, etc.) to himself. [1980, p. 66]

Fodor claims that this is 'roughly' right. It would be 'exactly right' if it were not for the imprecision of our "pretheoretic sense of opacity." The notion is notoriously difficult, and Fodor points out that in some cases it is difficult to tell how particular beliefs would be individuated according to the Content Individuation Principle. Many of these cases come from Putnam [1975] and the "New Theory of Reference." Suppose, for example, that there is another planet, called 'Twin Earth,' that is just like Earth in its composition and in the composition of its inhabitants. On Twin Earth I have a counterpart, Repro, that is just like me in his physical makeup. But there is, actually, one difference between Earth and Twin Earth: what they call 'water' on Twin Earth is made up of X, Y and Z, though it looks and behaves exactly like (real) water (to the non-chemist), which is, of course, H₂O. However, both Repro and I assent to 'water is wet.' According to Putnam we do not have the same beliefs: We both sincerely assent to 'water is wet', but I believe that water(H₂O) is wet--for that is what I am acquainted with or connected to in the appropriate way--and Repro believes that water(XYZ) is wet. I have never even seen water(XYZ), so it certainly seems as if I am not admitting to any beliefs about it when I assent to 'water is wet.' Since we have different beliefs, Putnam claims, Repro and I are in different types of cognitive states.

In this case I do not know whether or not the representations to which Repro and I stand in the belief relation are of different types according to the Content Individuation Principle; I, like Fodor, am afraid that the philosophical notion of opacity is just not clear enough. However, whether or not the Content Individuation Principle picks it out, in *some* sense Repro and I are in cognitive states of the same type. Whatever sense this is, Fodor says, is picked out by 'fully opaque type identification' [1980, p. 67]. Cognitive states, or rather, their constituent representations, are to be type-individuated by a "fully opaque" reading of 'that'-clauses.⁸

The 'computational' in 'computational model' is apparently supposed to convey the idea that our cognitive processes are somehow like the processes of computers. It seems to me, however, that most attempts to exploit the analogy result in a quagmire of imprecise and ambiguous jargon. Fodor is guilty of contributing to that quagmire, so I will try to be careful. First, what are cognitive processes? Fodor usually characterizes cognitive processes as "sequences of operations on representations" [Fodor, 1980]. Most of his commentators and critics follow him in using language of this sort. However, since the consequences of believing that φ are different from the consequences of fearing that φ , it is our relations to representations--our "cognitive states"--that are involved in cognitive processes. Fodor has been scolded for his sloppiness [Block and Bromberger, 1980], and he has admitted the error:

I have indeed played fast and loose with the question of whether mental operations apply to representations or to states... Probably the canonical formulation ought to be that mental operations apply only to states (it is states that are causally interrelated) but that mental states are relational (with mental representations figuring among the relata)... [Fodor, 1980 b]

However, in a later work, Fodor states that 'Mental processes are causal sequences of tokenings of mental representations'[1987, p. 17]. Assuming that he is committing the same oversimplification here, and that he means to say they are causal sequences of tokenings of cognitive states, it is probably safe to attribute this to Fodor:

CTM 4: Cognitive processes are causal sequences of cognitive states (and cognitive states have representations as constituents).

It is necessary to be careful here. Fodor wants his story about cognitive processes to make sense of our everyday propositional attitude ascriptions. We often point out that two people believe the same thing. But it would be perverse to say that Sally and John each believe that George Bush is the President of the United States just in case they stand in the same relation to a material object: the same *sort* of things might run through the heads of Sally and John when they think of

George Bush, but the very same thing does not. Fodor's story, then, has to be that two people are in the same type of cognitive state if and only if they stand in the same relation to representations of the same type. Given CTM, two people never have the same objects of belief, desire, etc., though the objects of their beliefs and desires can be of the same type.

We have the beginnings of a view here: Cognitive states are relations to representations. Representations are typically individuated according to 'that'-clauses on opaque readings. These representations are material things in our heads, and so they have causal properties that determine, in part, the causal roles of cognitive states. But if references to our cognitive states are to be of any explanatory value, when cognitive states of the same type are correctly attributed to two different people--when two people stand in the same relation to representations of the same type--then those cognitive states better not differ in their causal dispositions.

How are cognitive states involved in causal processes? Fodor (as well as other defenders of computational models) often make this sort of claim:

- [3] Cognitive processes are "formal" in that they apply to representations in virtue of their nonsemantic properties.

There are many instances, especially in [Fodor, 1980], where Fodor defends this sort of principle, which he calls 'the Formality Condition.' Of course, Fodor once again oversimplifies by characterizing processes as if they "apply to" representations. Presumably this is a slightly more precise formulation:

The Formality Condition

Cognitive processes are "formal" in that they apply to cognitive states in virtue of the nonsemantic properties of the representations that are the constituents of those cognitive states.

Though this claim may be appealing, it is not obvious what it says: to understand the Formality Condition we need, at least, to figure out what 'formal,' means here.

Fodor is not using the word 'formal' as it is used in symbolic logic, where a formal system is a set of rules governing the formation and manipulation of tokens

of types of strings of symbols. Syntactic operations, according to Fodor, are a species of formal operations because 'being syntactic is a way of *not* being semantic'[1980, p. 64]. But 'to say that an operation is formal is not the same as saying that it is syntactic since we could have formal processes defined over representations which don't, in any obvious sense, *have* a syntax. Rotating an image would be a timely example'[1980, p. 64]. Fodor uses 'formal' to make it clear that the properties that determine the relevant relations between cognitive states are *not* semantic properties, but, he says, that the notion of formality will 'have to remain vague and metaphoric'[1980, p. 64].

Whatever these nonsemantic properties are, they determine how cognitive states enter into cognitive processes. Thus representations of the same "formal" type will be indiscernible in cognitive processes: 'More generally: fix the subject and the relations, and then mental states can be (type) distinct only if the representations that constitute their objects are formally distinct'[Fodor, 1980, p. 64]. This, then, is fifth major claim of CTM:

CTM5:

Given two individuals, A and A', two mental representations, a and b, and a relation between individuals and representations, R, if a and b are of the same "formal" type, then ARa, ARb, A'Ra and A'Rb are all of the same cognitive type.⁹

Twin Earth cases provide some *prima facie* counterexamples to CTM 5. Remember that Twin Earth is like Earth in its composition and in the composition of its inhabitants--its inhabitants are molecule for molecule duplicates of the inhabitants of Earth. Now, suppose one day I see someone dressed in a silly coat wearing a fake nose, and I think to myself, *he's wearing a fake nose*. My counterpart, Repro, has the same sort of experience: he sees a twin earth person dressed in a silly coat, and thinks to himself, *he's wearing a fake nose*. By hypothesis our cognitive states must be of the same formal type. But, it seems, we are in different cognitive states. I believe of (Earth's) Mr. X that he's wearing a fake nose, and Repro believes of (Twin Earth's) Mr. X that he's wearing a fake nose. Since we have these beliefs about different people, they must have different "truth-

conditions." If these beliefs have different truth-conditions, they must be different types of beliefs.

But CTM already has an answer to this sort of problem: Representations are correctly individuated according to the Content Individuation Principle. Someone could point to one of the men and truly say that

[5] Randy believes that he is wearing a fake nose,

and he could point to the other one and truly say that

[6] Repro believes that he is wearing a fake nose.

But under an opaque reading of the 'that'-clauses, we cannot substitute '(Earth's) Mr. X' for 'he' in [5], or '(Twin Earth's) Mr. X' for 'he' in [6]. So, according to the Content Individuation Principle, Repro and I are in the belief relation to representations of the same type.

In fact the motivation for the Content Individuation Principle is part of the motivation for CTM 5. The Content Individuation Principle is supposed to be plausible because individuation of representational states according to "opaque"-readings of 'that'-clauses indicate how 'the agent represents the objects of his wants (intentions, beliefs, etc.) to himself.' Agents 'have no access to the semantic properties of such representations, including the property of being true, of having referents, or indeed, the property of being representations *of the environment*.' Instead, it is the nonsemantic properties of representations--i.e., the "formal" properties of representations--that involves those representations in cognitive processes.

What one would like to say, in particular, is that if two people are identically related to formally identical mental representations, then they are in opaquely type-identical mental states. This would be convenient because it yields a succinct and gratifying characterization of what a computational cognitive psychology is about: such a psychology studies propositional attitudes opaquely taxonomized. [1980, p. 66]

Computational processes, according to Fodor, are 'both *symbolic* and *formal*' [1980, p. 64]. They are "symbolic" because they involve representations and they are "formal" because they involve representations in virtue of the nonsemantic properties of the representations. Thus, a theory that honors CTM 1- CTM 5 is a computational theory.

5.2.3 Some Consequences

Some people seem to think that the Computational Theory of Mind is obviously correct; others seem to think that it is obviously wrong. There are two main areas of contention: (1) CTM entails that types of representations are individuated according to 'that'-clauses on "opaque" readings. If CTM is correct about our cognitive architecture, the taxonomy of cognitive states that everyday propositional attitude ascriptions provide better be explanatorily valuable. However, several people, notably Stephen Stich [1983] and Patricia Churchland [1986], have argued that everyday propositional attitude ascriptions yield a taxonomy of cognitive states that is either too fine-grained or too coarse-grained to be of any explanatory value.

(2) CTM entails that if two people are identically related to formally identical mental representations, they are in the same cognitive state. Thus CTM is supposed to conform to what Putnam calls 'methodological solipsism.' Methodological solipsism is the view that the only psychological states that should be recognized in psychological theories are those the ascription of which does not 'presuppose the existence of any individual other than the subject to whom the state is ascribed.' Putnam [1975], Tyler Burge [1986], Lynne Baker [1987], and many others, have criticized CTM on this count. The issues involved are difficult and the debate is a mess. I will try to avoid becoming part of it.

5.3 The Classical Computational Theory of Mind

There is still a lot of room for disagreement among those who accept CTM: Two computational theories might agree that cognitive states should be "solipsistically" or "opaquely" taxonomized, yet still disagree about how representations are individuated according to "nonsemantic" type--and thus about

which "nonsemantic" properties of representations are involved in determining the relevant causal roles of cognitive states. There are a variety of views that qualify as computational theories.

In cognitive science the "received" view about what properties are relevant to the individuation of representations by "nonsemantic" type seems to be what is sometimes called 'The Information Processing View' and what Fodor and Pylyshyn have defended as 'the Classical view.'¹⁰ Since there are a lot of ways to "process information" the first label may be misleading. I will adopt 'The Classical Computational Theory of Mind' (CCTM) as a label for this view.¹¹

5.3.1 The Theory

CCTM is distinguished by the following claim:

CCTM 1: Representations have a combinatorial syntax and semantics.

According to Fodor and Pylyshyn, this entails that

...(a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or are structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure. For purposes of convenience, we'll sometimes abbreviate (a)-(c) by speaking of Classical Theories as committed to 'complex' mental representations or to "symbol structures". [1987, p. 7]

Since these desiderata for representations are true of sentences, CCTM 1 is often expressed by saying that the constituent representations of cognitive states constitute a "language of thought." Tokens of those representations are "inscribed" or "encoded" in our brains. Of course, the details are missing. In any case, CCTM 1 entails that the "semantics" of representations are recursively describable in a way similar to a Tarski style interpretation of predicate logic. First there are some "syntactic" requirements: (1) Things of particular types count as tokens of "primitive symbols," just as tokens of the same types as 'V', 'x', 'P1' and '&' do in

monadic predicate logic. Apparently certain states of our heads (individuated according to their physical, chemical, neural or functional properties) are tokens of primitive symbols. (2) When certain sorts of relations hold between tokens of certain types of primitive symbols they form tokens of "atomic expressions." In monadic predicate logic any tokens of the same type as 'V', 'x', and 'P' form an atomic expression when they are concatenated in this way: V-x-P¹-x. In our heads tokens of primitive symbols form atomic expressions just in case they stand in particular (physical, chemical, neural or functional) relations to each other. (3) When certain sorts of relations hold between tokens of atomic expressions they form tokens of "compound expressions." (4) Two expression-tokens are of the same "syntactic" type if they are formed from primitive symbols of the same type and if the constituent primitive symbol-tokens in each of the expression-tokens stand in the same relevant relations to each other. Since expression tokens are all built up out of primitive symbol-tokens of certain sorts that stand in specific kinds of relationships to each other, it is possible to recursively define both what counts as an expression-token and what counts as an expression-type, even if there are infinitely many expression-tokens and types. (5) If the "semantics" of the "language" is also "combinatorial," then it must be possible to recursively specify the semantic "content" of expression-tokens given the semantic content of their constituents and the relevant relationships between those constituents.

Of course, it is senseless to defend CCTM 1 if there is no connection between the combinatorial syntax and semantics of representations and the role of representations in cognitive processes. The (physical, chemical, neural or functional) properties that determine the syntactic types of representations also determine the causal roles of the representations in cognitive processes:

Because Classical mental *representations* have combinatorial structure, it is possible for Classical mental *operations* to apply to them by reference to their *form*. The result is that a paradigmatic Classical mental process operates upon any mental representation that satisfies a given structural description, and transforms it into a mental representation that satisfies another structural description. (So, for example, in a model of inference one might have an operation that applies to any representation of the form 'P & Q' and transforms it into a representation of the form 'P'.)[1987, p. 8]

...The Classical theory is committed not only to there being a system of physically instantiated symbols, but also to the claim that the physical properties onto which the structure of the symbols is mapped *are the very properties that cause the system to behave as it does.* [1987, p. 8]

Once we have picked out the relevant "structural" properties of the primitive symbols, the properties which involve representations in cognitive processes will depend on how those primitive symbols are combined. It appears that these structural or syntactic properties of representations are the nonsemantic or formal properties of Fodor's Formality Condition, so I think that we can safely attribute this stronger version of The Formality Condition to CCTM:

The Syntacticity Condition:

Cognitive processes are "syntactic" in that they apply to cognitive states in virtue of the "syntactic" properties of the representations that are the constituents of those cognitive states.

According to CTM 5, cognitive states are of the same type if their constituent representations are of the same "formal" type. If CTM 5 is true, and if The Syntacticity Condition is true, then the following is true:

CCTM 2: Given two individuals, A and A', two mental representations, a and b, a relation between individuals and representations, R, if a and b are of the same "syntactic" type, then ARa, ARb, A'Ra and A'Rb are all of the same cognitive type.

5.3.2 Some Evidence

Fodor and Pylyshyn have recently recapitulated a well known family of arguments for why representations must have combinatorial structure. I will review two of them: (1) Thought is Systematic. Suppose that the sentences of natural languages did not have combinatorial structure, so that all expressions were atomic. Then sentences like 'John loves Sally' and 'Sally loves John' would be no more similar to each other than they are to 'President Bush enjoyed a long honeymoon'--the apparent structural similarities of the first two sentences would be an "orthographic accident." If that were true, it would be difficult to figure out

how we learn to use language, since each sentence would have to be learned independently of the rest. But natural languages do have a combinatorial structure; not all expressions are atomic. Competent users of English apparently recognize that 'John loves Sally' and 'Sally loves John' each have the form 'N Vt N' and that any expressions with that form are sentences, and they know the function of the nouns and verbs in the sentence. So competent speakers know what 'John loves Sally' means if they know what 'Sally loves John' means.

The same considerations that lead us to assume that language has combinatorial structure is supposed to provide a good reason to think that the "language of thought" has combinatorial structure. Suppose, for example, that cognitive representations did not have any combinatorial structure. Then thinking that *John loves Sally* would have nothing in common with thinking that *Sally loves John*.

What does it mean to say that thought is systematic? Well, just as you don't find people who can understand the sentence 'John loves the girl' but not the sentence 'the girl loves John,' so too you don't find people who can *think the thought* that John loves the girl but can't think the thought that the girl Loves John.

But now if the ability to think that John loves the girl is intrinsically connected to the ability to think that the girl loves John, that fact will somehow have to be explained. For a Representationalist..., the explanation is obvious: Entertaining thoughts requires mentally representing them. And, just as there are structural relations between the sentence 'John loves the girl' and 'the girl loves John,' so too there must be structural relations between the mental representation of the thought that John loves the girl and the mental representation of the thought that the girl loves John, namely, the two mental representations, like the two sentences, *are made of the same parts*. But if this explanation is right (and there don't seem to be any others to offer), then mental representations of thought have internal structure and there is a language of thought....[1987, p. 25]

(2) Inference is systematic. In logic the occurrence of '&' in 'P & Q' indicates that we may write down a token of either of the types of tokens that flank

the occurrence of '&'. But if 'P & Q' did not have any combinatorial structure, we could not have any such rule--instead it would be like making an inference from 'A' to 'B'. But we do make those sorts of inferences all the time, so that is evidence that we do exploit the combinatorial structure of language. The same sort of considerations apply to cognitive representations. If cognitive representations had no combinatorial structure, then inferring that *John loves Sally* from *John loves Sally and Sally loves John* would have nothing in common with inferring that *Sally loves John* from *John loves Sally and Sally loves John*. In that case there would be no reason to believe that someone could make the second inference if they could make the first. According to Fodor and Pylyshyn '...you don't find cognitive capacities that have these sorts of gaps.... Given a notion of logical syntax--the very notion that the Classical theory of mentation requires to get its theory of mental processes off the ground--it is a *truism* that you don't get such minds. Lacking a notion of logical syntax, it is a *mystery* that you don't'[1987, p. 32].

NOTES

- 1 The version of CTM that I present is intended to represent the position Fodor has developed over the last 15 years, but particularly the position as it was presented in [Fodor, 1975] and [Fodor, 1980].

Fodor and Pylyshyn defend CCTM in various places. I will concentrate on [Fodor, 1975], [Pylyshyn, 1984] and [Fodor and Pylyshyn, 1987].
- 2 Fodor actually make much stronger claims on the basis of his assumption about 8 through 12. His inferences strikes me as *non sequiturs* so I will introduce those stronger claims later.
- 3 It is important to note that learning a concept is not just learning an appropriate response to a type of stimulus: in many cases concept learning seems to occur in the absence of any specific designated appropriate response [Fodor 1975, p. 36 fn.]. In experiments designated responses help determine to what degree the subject has learned the concept in question.
- 4 If this relational view is not accepted, it is hard to see how sentences that use these sorts of locutions could have compositional semantics.
- 5 [Boyd, 1980] and [Feldman, 1980] provide valuable discussions of contingent identity and event individuation.
- 6 Perhaps the best way to account for Fodor's inconsistencies is to suppose that he is continually confusing the properties in virtue of which mental states are subsumed under psychological laws--i.e., the fact that they are bearing-a-particular-relation-to-proposition-states--with the properties of the mental states which instantiate the former properties--i.e., the fact that they are bearing-a-particular-relation-to-an-internal-formula-states. I am sure that he would disagree.
- 7 Fodor presses this claim most vigorously in [Fodor, 1980].
- 8 What Fodor refers to as 'fully opaque type identification' is apparently the same as individuation according to "narrow content." [Fodor, 1987] is devoted to saying what narrow content is.
- 9 This needs a proviso. Whether strings of symbols are of the same type will depend on the "formal system" that governs them. Some formal systems have

rules that are ignore differences between strings that the rules of other systems would not ignore. Some machines will be sensitive to differences between representations that others would not. Thus CTM 5 will hold only if A and A' use the same sort of 'rules.' [Fodor, 1980 p. 106]

10 See [Fodor and Pylyshyn, 1987].

11 CCTM is a theory about the nature of only those states that are involved in "cognitive" processes. I am not sure exactly what counts as a cognitive process, or whether CCTM is supposed to be a theory about all cognitive processes. Fodor and Pylyshyn often seem to claim that CCTM governs all cognitive processes. Of course, then being governed by CCTM might just turn out to be their necessary condition for being a cognitive process. However, Pylyshyn [1984, chapter 7] has discussed whether certain sorts of cognitive processes might be analogue, and so not combinatorial as CCTM demands.

CHAPTER 6

THE CLASSICAL COMPUTATIONAL THEORY OF MIND AND STATE-STATE LAWS

6.1 Introduction

The Classical Computational Theory of Mind has been criticized for a variety of reasons. Some people, such as Stich [1983] and Churchland [1986], argue that individuating cognitive states according to "content" will yield few helpful generalizations about our cognitive processes. Both the Formality Condition and the Syntacticity Condition seems to require cognitive theories to be "solipsistic" or "individualistic" in a way that is unpalatable to others, such as Baker [1987] and Burge [1986]. Finally, "connectionists," such as McClelland, Rumelhart and Hinton [1986a], argue that "representations" do not have anything like a combinatorial syntax.

Curiously, the Classical Computational Theory still seems to be quite popular. If it were not, it probably would not be attacked so often. In this chapter my project is less ambitious than those attacks: All that I want to question is whether, given CCTM, it is necessary to appeal to the semantic properties of "representations" in order to capture the generalizations about relations between cognitive states that are supposed to be captured when we do refer to those semantic properties.

6.2 The Syntactic Theory of Mind

Back in Chapter 3 I discussed an argument that I called the "loss of generalizations argument." The gist of the argument is that since many types of events have a variety of different physical instantiations, we are not able to state in the vocabulary of physics many important generalizations about the relationships between those types of events. That sort of argument is often offered as a defense of "representational" psychological theories. Consider, for example, the following passages from Pylyshyn:

A pedestrian is walking along a sidewalk. Suddenly the pedestrian turns and starts across the street. At the same time, a car is traveling rapidly down the street toward the pedestrian. The driver of the car applies the brakes. The

car skids and swerves over to the side of the road, hitting a pole. The pedestrian hesitates, then goes over and looks inside the car on the driver's side. He runs to a telephone booth and dials the numbers 9 and 1.[1984, p. 3]

What will he do next? The answer is obvious when the situation is described using the particular terms I used. Because the pedestrian knows the emergency number is 911, and because he perceives the situation to be an emergency, his next behavioral act is overwhelmingly likely to be: dial 1. The way I describe the situation, no account stated solely in behavioral terms (that is, in terms that do not incorporate the person's knowledge or goals) can make such a prediction. The reason is, if a systematic account is to connect the prediction that person's next act will be to dial 1 with the "stimulus" conditions, such an account must at the least mention: that the pedestrian interpreted the scene as an accident; that the pedestrian knows or remembers the phone number; and that the pedestrian's behavior is an instance of the category "phone for help." That this is true can be seen by considering the other possibilities, that is, by considering certain counterfactual possibilities. There are physically identical situations in which the prediction does not hold (for example, telling the person in advance that there is to be a rehearsal of a television show on this street). And there are situations in which the prediction does hold--for exactly the same reason (that is, the explanation is the same) but the physical situation is vastly different (for example, instead of the person seeing the injured driver, he hears the driver's cries for help or is told by a passer by that someone has been injured). Or the physical conditions may be the same but the detailed behavior different (the person's arm is in a cast, so he asks someone else to dial 911, thereby realizing the prediction with highly different behavior). The point is, there are innumerable many physically distinct ways in which the same generalization can be realized; yet they remain cases of the same generalization. If that generalization were not recognized, each instance would count as a different sequence, and we would miss an important regularity. [1984, pp. 7-8]

There is a lot going on here. First let's consider Pylyshyn's claim that we cannot give an explanation of the pedestrian's behavior in purely physical terms. He admits elsewhere that we could give a description of the situation in purely physical terms.[1984, p.3] I think that what Pylyshyn is claiming here is something like this: We could explain why certain of the pedestrian's neuron's fired in a certain way when he saw the accident, and we could explain why that caused some

of his nerve muscles to twitch in a certain way (that might be described as going to and looking in the window of the car), and we could explain why his finger finally twitched in a certain way (that might also be described as dialing a '1'), all in the vocabulary of physics or biology. But that does not explain the dialing of '1' described as 'dialing a '1''. This is not merely because 'dialing a '1'' is not part of the vocabulary of physics or biology, but because it picks out a type of event that can be caused in indefinitely many ways at the level of physical description. Any satisfactory account of the pedestrian's dialing a '1' will have to provide an explanation of the occurrence of that type of event in similar situations.

Consider how Pylyshyn would explain the pedestrian's dialing the '1', and still preserve the appropriate generalizations. He would presumably describe the pedestrian's behavior in something like these terms: The pedestrian saw an accident. He interpreted what he saw as evidence of an accident. He already had a desire to help in cases where accidents occur, so the pedestrian came to desire to help. He believed that the best way to help was by phoning the police. He remembered that the number of the police was '911', so the pedestrian dialed '911'. Most of this account holds also for cases where the pedestrian heard a cry for help, and then interpreted that as evidence of an accident, or where he was told about the accident by a passerby, and then interpreted that as evidence of an accident. These different situations are very different physically, but they are all cases where the pedestrian acted in a particular way because of his particular beliefs, desires, memories, etc.

Since this sort of account is supposed to preserve the appropriate generalizations, we should consider whether we might be able to give a similar account without losing those generalizations, and without appealing to "representational" states. One candidate alternative is Stephen Stich's Syntactic Theory of Mind (STM).[Stich, 1983, chapter 8] STM is closely related to CCTM. According to STM, cognitive states are best construed as relations to syntactic objects, and cognitive states are individuated according to the "syntactic" properties (such as in CCTM 2) of their constituent "representations." But, Stich claims, everyday propositional attitude ascriptions, or ascriptions that conform to the Content Individuation Principle, do not individuate cognitive states in a way

adequate for psychological explanation. As evidence he provides several Twin Earth-style examples, and a few of his own, to illustrate that though we might know all about the behavior of two different people, their physical characteristics and their environment, we would still not be sure whether to ascribe the same beliefs to each of them. Thus, Stich claims, individuating cognitive states according to "semantic content" is inadequate for cognitive explanation. STM includes no notion of "semantic content": cognitive states are not relations to "representations" but relations to (uninterpreted) syntactic objects, and cognitive states stand in certain causal relationships with each other in virtue of the syntactic properties of those syntactic objects.

Stich argues that STM is sufficient for most cognitive psychology and that many cognitive theories are best construed as syntactic theories. For example, Pylyshyn captures the generalizations that he wants to capture by ascribing particular cognitive states to the pedestrian. The pedestrian would be in the same types of cognitive states in relevantly similar situations. For example, seeing the injured driver, hearing cries for help, and being told 'that's an accident' by a passer by, all cause the pedestrian to have a cognitive state of believing that there has been an accident. But if cognitive processes apply to cognitive states in virtue of the syntactic properties of the representations that are the constituents of the cognitive states, we should be able to provide an explanation of the pedestrian's behavior while referring to only the syntactic properties of the cognitive states when we provide the explanation of the pedestrian's behavior.

A purely syntactic explanation might be produced in the following way: According to CCTM each representation, φ , is a string of symbols with certain sorts of syntactic properties. We stand in "believing", "desiring", "fearing", and "remembering" relations to those representations. Instead of talking in terms of the pedestrian's relations to representations, we can talk in terms of his relations to purely syntactic entities. Remember, according to CCTM it is the syntactic properties of representations that determine the causal relations between representations, so, it seems, nothing will be lost by adopting this view.

Now consider the following explanation: The pedestrian, *A*, saw an accident. Seeing the accident caused him to go into state $AB\varphi$ (where 'B' expresses the

relation that Pylyshyn would call 'the believing relation,' and ' φ ' is the syntactic object that Pylyshyn would interpret as *that an accident occurred*). A had a D-state, $AD(\varphi \rightarrow \psi)$ (where 'D' expresses the relation that Pylyshyn would call 'the desiring relation,' and ' $(\varphi \rightarrow \psi)$ ' is the syntactic object that Pylyshyn would interpret as *that he helps if an accident occurs*). By referring to only the syntactic properties of φ and $\varphi \rightarrow \psi$, we can show that A then had to come to be in D-state $AD\psi$ (where ' ψ ' is the syntactic object that Pylyshyn would interpret as being *that he helps*). A was in a B-state $AB\theta$ (where ' θ ' is the syntactic object that Pylyshyn would interpret as *that the best way to help in case of an accident is by phoning the police*). A was also in an R-state $AR\zeta$ (where 'R' expresses the relation that Pylyshyn would call 'the remembering relation,' and ' ζ ' is the syntactic object that Pylyshyn would interpret as *that the phone number of the police is '911'*). So the pedestrian dialed '911'. The same sort of explanations could be given in the case where the pedestrian was told of the accident, or heard cries for help. Those events would also cause the pedestrian to go into state $AB\varphi$. Of course actually providing this sort of explanation would require a good deal of ingenuity, but that is for the psychologists to worry about.

6.3 A Problem

Stich argues that purely syntactic theories can provide most of the explanations that cognitive psychologists want to discover, and he claims that most of the explanations that cognitive psychologists actually give do conform to STM. Curiously enough, he accuses Fodor of agreeing with him. Stich cites as evidence for the accusation several instances where Fodor makes claims such as 'Mental representations have their causal roles in virtue of their formal [or syntactic] properties...' [1981, p. 26]. If Fodor believes that, Stich argues, then he must be a defender of STM.

However, Fodor also claims that cognitive generalizations apply to cognitive states in virtue of their content and, as Fodor states emphatically, 'YOU CAN'T SAVE THESE GENERALIZATIONS WITHOUT APPEALING TO THE NOTION OF THE CONTENT OF A MENTAL STATE...' [1981, p. 26]. Stich thinks that the mystery can be cleared up by ascribing the "correlation thesis" to Fodor. The correlations thesis is that the

'semantic features [of mental state tokens] are correlated with the syntactic type of the token' and that 'if a pair of mental state tokens are of the same syntactic type, then they must have the same content or truth conditions as well'[Stich, 1983, p. 186]. Thus the correlation thesis seems to be the same as the Syntacticity Condition (CCTM 2). Stich then offers this 'benign' interpretation of Fodor:

How is it possible for Fodor to have it both ways...? One way to take the bite out of this apparent contradiction would be to endorse the correlation thesis which holds that differences in content are mirrored by differences in syntax. If this were true, then generalizations couched in terms of content would, so to speak, be coextensive with generalizations couched in terms of syntax. And although strictly speaking it might be their syntactic properties which account for causal interactions among mental state tokens, there would be no harm in talking as though semantic properties were causally relevant, since if they were, the system would behave in the same way.[1983, p. 188]¹

Stich thinks that if his "benign" interpretation is right, then Fodor agrees that the relations between cognitive states can be explained without referring to the semantic properties of representations, and so the matter is closed. However, as it stands, Stich's argument is a *non-sequiter*: the correlation thesis--which seems to be equivalent to the Syntacticity Condition (CCTM 2)--does not guarantee that generalizations couched in terms of "content" are "coextensive" with generalizations couched in terms of syntax. It is perfectly compatible with the Syntacticity Condition that two cognitive states of the same type can be relations to representations of different syntactic types. In other words, the problem is that the truth of the following statement is not guaranteed by the Syntacticity Condition:

- [1] Given two individuals, A and A', two mental representations, a and b, and a relation between individuals and representations, R, if ARa and ARb, (and A'Ra and A'Rb) are of the same cognitive type, then a and b are of the same "syntactic" type.

If Fodor does indeed believe that representations with the same "content," "meaning," etc., might be of different syntactic types, then Stich should, at least,

supply an argument to show why this sort of multi-instantiation does not make it impossible to state in syntactic terms the sort of generalizations that Fodor wants to state in terms of "content."

I think Fodor would admit to holding [1]. Just suppose that [1] is not true, but that the rest of CCTM is. Now consider a case where Repro--my molecular duplicate--and I stand in the belief relation to representations of the same type according to type individuation by "fully opaque" readings of 'that'-clauses. [5.5] and [5.6] are both true, so, according to the Content Individuation Principle, we stand in the same relation to representations of the same type. However, suppose that Repro and I are not quite molecular duplicates: our mental representations are of different syntactic types. There are two possibilities: (1) The syntactic differences between our representations could be significant in the sense that, because of those differences, Repro's cognitive state and my cognitive state do not have the same "computational," "causal," "functional" properties, even though they are cognitive states of the same type. But then Repro and I might share the same types of cognitive states and yet do completely different things. It is unreasonable to suppose, then, that representations of the same syntactic type would always be of the same (semantic) type according to the Content Individuation Principle. I doubt, that Fodor (or Pylyshyn) would find this possibility attractive. (2) The syntactic differences between our representations could be insignificant in the sense that those differences are irrelevant to the "computational," "causal," "functional" properties of the cognitive states of which they are constituents. In this case there would not seem to be any reason to claim that there were syntactic differences between the representations in the first place. Such fine-grained individuation of syntactic-types would be irrelevant in explanations of cognitive processes, so I doubt that anyone would argue for this possibility.²

If both CCTM 2 and [1] are true, then so is the following:

The Strong Correlation Thesis:

Given two individuals, A and A', two mental representations, a and b,
and a relation between representations, R, ARa and ARb, (and A'Ra

and A'Rb) are of the same cognitive type if and only if a and b are of the same "syntactic" type.

Does Fodor, and the other defenders of CCTM, have to admit that generalizations about cognitive processes need refer to only the syntactic features of representations? Fodor obviously does not think so. He argues vigorously that attributions of "content" are necessary in cognitive explanations, so he seems to have some reason to believe that references to the semantic properties of representations are necessary in some cases. I suspect that Fodor's apparently inconsistent claims stem from his belief that the semantic properties of representations in some sense "supervene" on their syntactic properties: the syntactic properties might "do the work," even though cognitive generalizations still must appeal to the semantic content of the representations. But given the Strong Correlation Thesis, whenever we appeal to a law that subsumes cognitive states in virtue of their "content"--i.e., the semantic properties of their representations--there will be a lawlike statement that subsumes those very states in virtue of the syntactic properties of their representations.

6.4 Pylyshyn's Defense

Pylyshyn has responded directly to Stich's suggestion:

I don't believe we could get away with it and still have explanatory theories. It simply will not do as an explanation of, say, why Mary came running out of the smoke-filled building, to say that there was a certain sequence of expressions computed in her mind according to certain expression-transforming rules. However true that might be, it fails on a number of counts to provide an explanation of Mary's behavior. It does not show how or why this behavior is related to very similar behavior she would exhibit as a consequence of receiving a phone call in which she heard the utterance "the building is on fire!", or the consequence of her hearing the fire alarm or smelling smoke, or in fact following any event interpretable (given the appropriate beliefs) as generally entailing that the building was on fire. The only way to both capture the important underlying generalizations (which hold across certain specific nonverbal inputs as well as certain classes of verbal ones, but only when the latter are in a language that Mary

understands) and to see her behavior as being rationally related to certain conditions, is to take the bold but highly motivated step of interpreting the expressions of the theory as goals and beliefs....

In the above example, simply leaving them as uninterpreted formal symbols begs the question of why these particular expressions should arise under what would surely seem (in the absence of interpretation) like a very strange collection of diverse circumstances, as well as the question of why these symbols should lead to building evacuation behavior as opposed to something else. Of course, the reason the same symbols occur under such diverse circumstances is precisely that they represent a common feature of the circumstances--a feature, moreover, that is not to be found solely by inspecting properties of the physical environments. (E.g., what physical features do telephone calls warning of fire share with the smell of smoke?) What is common to all these situations is that a common interpretation of the events occurs--an interpretation that depends on what beliefs Mary has about alarms, smoke, and so on.... But what in the theory corresponds to this common interpretation? Surely one cannot answer by pointing to some formal symbols. The right answer has to be something like the claim that the symbols represent the belief that the building is on fire--i.e., it is a semantic interpretation of the symbols as representations of something.[1980 p. 161]

Judging from the number of times this passage has been cited in the literature, it must provide a standard defense of why we have to appeal to the semantic properties of representations. I am not sure what the defense is.

Pylyshyn's first point is that an explanation of Mary's behavior that appeals to her relation to uninterpreted formulae '...does not show how or why this behavior is related to very similar behavior she would exhibit as a consequence of receiving a phone call in which she heard the utterance "the building is on fire!", or the consequence of her hearing the fire alarm or smelling smoke, or in fact following any event interpretable (given the appropriate beliefs) as generally entailing that the building was on fire.' But the obvious answer seems to be that in each case Mary would come to stand in the same relation to the same formula.

Pylyshyn then says that this sort of answer '...begs the question of why these particular expressions should arise under what would surely seem (in the absence of interpretation) like a very strange collection of diverse circumstances,

as well as the question of why these symbols should lead to building evacuation behavior as opposed to something else.' Pylyshyn is surely right when he claims that appealing to uninterpreted formulae does not explain why they, in particular, would be implicated in bringing about Mary's behavior. But we do not expect that our reference to Mary's relation to uninterpreted formulae will explain why she has that relation to those formulae. Surely the explanation has to do with Mary's perceptual capacities and her other cognitive states. How would Pylyshyn explain why Mary came to believe that the building is on fire in each of the different situations? Surely he would not explain it by claiming that Mary would "interpret" the situations in the same way.³

Finally, Pylyshyn says that 'What is common to all these situations is that a common interpretation of the events occurs--an interpretation that depends on what beliefs Mary has about alarms, smoke, and so on.... But what in the theory corresponds to this common interpretation? Surely one cannot answer by pointing to some formal symbols.' It *seems to me* that Pylyshyn is saying something like this: 'To really capture a generalization about Mary's behavior in this sort of situation we have to relate it to the sorts of circumstances that bring it about. But there is little physical similarity between the different circumstances, so we cannot relate her behavior to some physical properties of the circumstances. However, those circumstances do have something in common: Mary interprets them as cases where the building is on fire. In this way a representational theory relates Mary's behavior to a common property of the precipitating conditions.' I said that it seems to me that this is what Pylyshyn is saying because it is a little hard to believe that he would actually say this. Of course it is true that all circumstances interpreted by Mary to be cases where the building is on fire are circumstances interpreted by Mary to be cases where the building is on fire. But surely we cannot just point out that each of those circumstances has the property *being interpreted by Mary to be cases where the building is on fire* in an explanation of her belief or of her behavior. After all, every circumstance that brings about a certain relation between Mary and a particular formula is a circumstance that brings about a certain relation between Mary and a particular formula. But we cannot appeal to the fact that the circumstances each have the property *bringing about such and*

such relation between Mary and such and such formula to explain either Mary's relation with the formula or her behavior.

It might seem as if I am missing an obvious point. Referring to that passage, Stillings, et al. [1987] claim that

The point of Pylyshyn's argument is fairly straightforward. A good cognitive explanation of behavior that is motivated by beliefs ought to explain how those beliefs are related to the behavior and to the circumstances that give rise to them. If the beliefs are characterized by the theory as uninterpreted symbols, and if believing is characterized as an uninterpreted process in the believer, then the theory cannot explain their connection either to behavior or to stimulation--or for that matter, to other beliefs. In any real explanation, this objection goes, the content of the belief plays a role. The Symbols in Mary's head cause her behavior *because* they represent the fact that there is a fire, and any symbols that did not represent that fact would by themselves not explain their behavior. The conclusion that a naturalistic individualist draws is that in a cognitive theory internal information-processing states must be identified by their content, and in order for this to happen, one must of course examine their connections not only to other cognitive states and processes but also to the organism's environment. [1987, p. 339]

I don't understand this any better than I understand Pylyshyn's passage: the role that the semantic properties of the representations play in the explanations remains a mystery to me. However, the interpretation seems to point out correctly that Pylyshyn--in the previous passage, at least--thinks that references to semantic properties are necessary to explain the connection between cognitive states and both behavior and stimulation. This is interesting, because, as we will see in the next chapter, both Pylyshyn and Fodor think that, in one important sense, there is no lawlike connection between our cognitive states and either behavior or stimulation.

NOTES

1 Stich hesitates to ascribe the correlation thesis to Fodor because in some places Fodor seems to reject the thesis. Stich accounts for Fodor's inconsistency by attributing two different notions of content to Fodor. Those two different notions are apparently the notions of "narrow content" and "wide content" that Fodor tries to explicate in [Fodor, 1987]. Individuation according to narrow content should be the same as individuation according to the Content Individuation Principle, "plus or minus a bit."

2 Pylyshyn does write this:

Typically, we only get an explanatory advantage by appealing to representational content, when the semantic interpretation *S* we give to certain states *I* happens to have such characteristics as the following:

(a) According to our method of assigning content there will [be] more than one distinct *I* in our model that corresponds to some actual or possible *S*; that is, there will, in general be more than one way of representing something in our scheme....[1984, p. 46]

This passage seems to be at odds with [1]. Elsewhere Pylyshyn claims that the semantic properties of representations will be determined by their functional roles, which are in turn determined by their syntactic properties. Given that, representations of the same semantic type could only differ in syntactic type if those differences were causally irrelevant or (presumably what Pylyshyn has in mind) if those representations were involved in completely different cognitive processes--i.e. if they were involved in processes in different "modules."

3 Stich [1983, p. 176] gives similar replies to Pylyshyn's argument.

CHAPTER 7

STIMULUS-STATE AND STATE-BEHAVIOR LAWS

7.1 Introduction

If the Classical Computational Theory of Mind is correct, references to semantic properties do not seem to play an essential role in explaining the relationships between cognitive states. However, references to semantic properties may have essential roles in explaining how cognitive states are "connected" to stimuli and to behavior. Of course, determining what roles they play will turn on how types of stimuli and behavior are individuated. The task of those defending the necessity of representational theories is to show that at least one type of behavior, however it is individuated, can be explained only by referring to semantic properties.

There is one way to individuate types of stimuli and behavior that deserves immediate comment. We often individuate types of stimuli and behavior according to their relationships with our "representational states." For example, we might refer to some events as the kind that cause people to believe that they are witnessing an accident or we might refer to some events as attempts to get help. If there have been events of the first type, then there have been people who have been caused to believe that they are witnessing accidents; if there have been events of the second type, then someone must have acted with the intention of getting help. Individuating types of stimuli and behavior in this way seems to automatically require explanation in terms of how the subject "represents" the world. Is this a good reason to think that we need references to representational states to explain some types of behavior?

I think that there are at least three reasons to deny that it is: (1) If types of stimuli and behavior were individuated according to their relationships to our representational states, we could only know what sort of behavior people exhibit by first knowing what their representational states are. But how are we going to figure that out? Psychology would be *really* hard if those were the only types of events that are lawfully connected to our cognitive states. (2) If types of stimuli and behavior were individuated according to their relationships to our

representational states, it's not clear what explanatory value representational theories could have. Suppose, for example, we were trying to explain why we sometimes come to believe that we are witnessing accidents. We would not succeed by pointing out that the beliefs were caused by causes-people-to-believe-that-they-are-witnessing-an-accident-events. Suppose we were trying to explain why someone tried to get help in a certain situation. Since trying to get help is no more than acting with the intention of getting help, we would not explain anything by pointing out that the behavior was caused by intending-to-get-help-events. When we individuate types of stimuli and behavior this way, references to stimulus-events seem to be irrelevant in explanations of why we come to believe what we believe, and references to our intentions seem to be irrelevant in explanations of why we do what we do. (3) Finally, when we individuate types of stimuli and behavior according to their relationships to our representational states, we presuppose that we actually do have representational states. But that begs the question. If we do not have representational states, then there are no types of behavior that are individuated according to their relationships with our representational states. We need an argument for the necessity of interpretive explanation *before* we can consider any sort of taxonomy of stimulus conditions or behavior that smuggles in references to the representational states of the subject.

Given this constraint, there are two diametrically opposed strategies for arguing that references to semantic properties do have essential roles in cognitive explanations: The first and most obvious strategy is to argue that there are lawlike generalizations about the relationships between types of stimuli and our cognitive states, and between our cognitive states and our behavior, that cannot be stated without referring to representational states. The second strategy, pursued by both Fodor and Pylyshyn, is to argue that some generalizations must refer to our representational states because there can be no such lawlike generalizations about the relationships between certain sorts of stimuli to which we respond and our cognitive states.

7.2 Churchland's Argument

Suppose that there are lawlike generalizations about the relationships between stimulus conditions and our cognitive states and that there are lawlike generalizations about the relationships between our cognitive states and our behavior. If the "representations" to which those laws refer are uninterpreted, then the generalizations will only state connections between stimuli and behavior and certain relationships we bear to uninterpreted formulae. As Devil's advocate, Patricia Churchland has suggested that much more gratifying generalizations can be made if the formulae really are representational.¹ Stich [1983] developed the suggestion in the following way. We might offer a generalization of this sort:

- [1] For all subjects A and all noun phrases N , if a sentence of the form
- n comes into A 's view
- is true, then typically A will acquire a belief ascribable by a sentence of the form
- A believes that n is in front of him,
- where 'n' throughout is replaced by N .

This is an instance of the generalization:

- [2] If a computer comes into A 's view, then A believes that a computer is in front of him.

This sort of generalization can be given only if the words of the noun phrases that characterize the stimuli can also be used to characterize the belief. But of course, there is no way to provide these sorts of generalizations if we refuse to provide any semantic interpretation for "representations": Instead we could only get a "law" to the effect that if someone sees a computer then they bear a certain relation to such and such uninterpreted formula.

Generalizations like [2] look attractive only until we begin to think about them. For example, someone who did not know what a computer was would not come to believe *that there is a computer in front of me* if he saw one; someone would not necessarily come to believe *that there is a spy in front of me* if one came into view.

If the noun phrases referred to in [1] are restricted to "observational terms", i.e., ones such that a subject acquires beliefs attributable by using it whenever an object denoted by the terms come into view, then [1] will be true. But, of course, we need some *independent* way to identify those noun phrases if [1] is going to be interesting.

Stich has argued that the prospects are not very hopeful. Whether we can appropriately attribute a particular "belief" to a person depends what other "beliefs" the person has: there is no stimulus that would cause every subject to have a particular belief. Stich has a whole battery of examples and makes a convincing case for his claim. Most of his examples involve people who have beliefs that are radically different from our own. For example, he mentions the Nuer, a "primitive" tribe, who have been credited by anthropologists with the belief that a particular cucumber used in certain rituals is really an ox. 'Cucumber' then seems to be suspect as an observational term. Of course we might object to this on the grounds that the translation is "indeterminate"--perhaps we should never attribute such views to people. But we do not have to travel to distant cultures with Stich to see convincing cases. Consider devout Catholics who believe the doctrine of transubstantiation. If they really believe it, then they believe that during communion they eat flesh and drink blood. Most non-Catholics would swear that it is bread and wine. If you are Catholic, this is an example that shows that 'flesh' and 'blood' are not appropriately observational; if you are not, then this example shows that 'bread' and 'wine' are not observational terms. Philosophical "idealists" also provide some good examples: When material objects come into view, they do not acquire beliefs that material objects are in front of them, so noun phrases denoting material objects must not be observational.

But what about terms such as 'red' and 'blue'? Stich suggests that we can see that they are not observational by considering the case of the Dani, an aboriginal tribe of New Guinea, who have only two color terms: one for dark colors and another for bright ones. In this case 'red', 'blue' and other color terms that are common to us could not be used to satisfy [1]. Of course one might argue that the Dani simply do not have terms for the colors, and that [1] holds in these cases because they still have beliefs about red things, and we can attribute them to the

Dani. However, the Argentine gauchos provide a less exotic example. I understand that they use more than forty different color-terms to describe their local horses. They must be sensitive to distinctions that I am unable to make .

If, for any term that we might consider, we could always come up with a set of beliefs that *A* (in [1]) might have that would make us question whether the term in question was observational, then [1] will be of value only for the psychology of "normal, average, nonreligious, nonphilosophical and uninteresting" people. Stich claims that those who believe that generalizations such as [1] are important in psychological explanation have the burden of proof here: They have to provide us with enough "observational terms" to make that sort of generalization useful.

There are also glaring problems with generalizations that tie beliefs to behavior. Consider the following generalization:

- [3] For all subjects *A* and all declarative sentences *P*, if *A* has a desire ascribable with a sentence of the form

A desires that *p*

where 'p' is replaced by *P*, and if *A* has no stronger incompatible desires, then *P* will come to be true.

The obvious problem here is that people usually cannot bring about the things that they desire most. We need, at least, to limit the generalization to "obtainable desires," or something of that sort. But what, Stich asks, is an obtainable desire besides one that will come true if we desire it to come true more than any other desire? The burden of coming up with a suitable notion of what an obtainable desire is, Stich adds, on the psychologist or philosopher who would like to employ a generalization like [3].

Stich may be rushing too quickly to condemn Churchland's suggestion. After all, we do recognize computers, cucumbers, wine, bread and colors most of the time, and we seem to depend on that sort of fact when we predict the behavior of those around us. It seems that there must be useful statistical generalizations. But, though *we* recognize computers most of the time, *people* usually do not: most people, living or dead, never recognized a computer *as a computer*; nor would they have had they been shown one. Statistical laws, like lawlike universal

generalizations, are not just statements of actual correlations between certain types of events, but are statements about what would happen if such and such *were* to happen. As such they are expected to hold for cases similar in certain "relevant" ways. But most humans would not recognize computers, cucumbers, wine, or bread (and possibly certain colors) unless they lived in the sort of society that we live in.

The obvious remedy is to restrict the laws to people who are like us in particular ways. Here the danger is that if the group of subjects is too homogeneous then psychological "laws" might be guaranteed, since people who are "just like us" are going to react to the environment the same way we do. As we shrink the population of subjects to include only those who are most like us we will discover more and more generalizations about the members of the population. Those generalizations might be useful tools for predicting behavior, but they would require explanations, not provide them.

The sorts of generalizations suggested by [1] are supposed to be stimulus-state laws. Ideally, *laws* connecting stimuli to our cognitive states will not presuppose a tremendous amount of information about the content of our other cognitive states. They will be most useful where the perception of the stimulus that is described will not alter depending on one's other beliefs and cognitive states. In other words, the processes that the laws will govern will not be "cognitively penetrable." Research ([Marr, 1982] for example) and common sense tell us that there are many such perceptual processes. As Pylyshyn says,

An organism's contact with the environment must, at some level, be decoupled from its cognitive processes; otherwise the organism has no stable base of causal interactions with the world from which to develop veridical perceptions. If it were not for this decoupling the organism would be in a position of despotic leader whose only contact with the world is through compliant servants who, in their eagerness to please, tell their leader whatever he wants to hear, thus ensuring that he remains confined to a world of his wishes. [1984, p. 155]

But as we have seen, what we say when we report 'I see _____' usually depends on what we already believe. What sorts of things we recognise and how many shades of brown we can distinguish depends on our history--just as our

ability to distinguish between different musical notes will depend on our musical training. Thus explanatory laws linking stimuli and cognitive states will probably subsume only early perceptual processes. Those cognitive states may not resemble the familiar "belief" states of everyday explanation, and we may not be able to state the content of those cognitive states in English. Whether there are useful laws of that sort is an empirical matter. Whether appeals to content are necessary to explain those processes will still be open to debate.

It is worth noting that the theory of vision that currently seems to enjoy the most popularity [Marr, 1982] describes early visual processes in terms of representations--"raw primal sketches" and "primal sketches"--which do not seem to be combinatorial in the sense required by CCTM 1.² Also, Burge [1986] argues that the "content" of those representations is not "individualistic" and so does not conform to Fodor's Formality Condition. As we will see in section 7.4, Fodor and Pylyshyn do not rest their claims that we must refer to the "content" of cognitive states in explanations of behavior on anything like Churchland's argument.

7.3 The Argument from Stimulus Independence

It is not unusual to come across the claim that we must refer to our beliefs or our other propositional attitudes to explain our behavior because our behavior is "stimulus-independent." Presumably, to say that something's behavior is stimulus-independent is just to say that it does not necessarily react in the same way in the same type of situations. Unfortunately, it is hard to make this more explicit. Consider the case where Fred saw a car careen off of a road at high speed. In that situation Fred hurried to the nearest phone and called an ambulance. However, it is possible that in another case where Fred sees a car veer off of a road at high speed he will laugh and mutter 'natural selection, hah.' Even though Fred's seeing the accident might be a crucial part of the explanation of why he called the police, Fred might very well do something different in the same sort of situation. This then is my suggestion:

- [4] *A* exhibits stimulus-independent behavior, if and only if
 - (i) *A* perceives something, *S*,
 - (ii) *S* has properties p_1, \dots, p_n ,

- (iii) (i) and (ii) enter into an explanation of A 's behavior coming to be C
- (iv) It is (physically) possible that
 - (iva) A perceives something S^* , where $S \neq S^*$,
 - (ivb) S^* has properties p_1, \dots, p_n , and
 - (ivc) (iva) and (ivb) do not enter into an explanation of A 's behavior coming to be C .

'Perceives' must be read in a broad sense, to avoid begging any questions. In this broad sense, some microorganisms can perceive things because they can (somehow) sense lightwaves reflected off of things, and respond (somehow) to those lightwaves. Thus A 's perceiving something, S , does not entail that A "recognizes" S as any particular sort of thing. "Behavior" must also be read in a broad sense, where reflex reactions to a doctor's hammer count as behavior. Thus A 's exhibiting a type of behavior, C , does not entail that A "intended" to do C . Unfortunately 'explanation' is also a problem, since, it is probably safe to say, at this point no one has produced a completely adequate account of *any* sort of explanation. Because of the difficulty of some of the notions involved, this account of stimulus-independence has some shortcomings. But it should be clear enough that we can evaluate the claim that stimulus independent behavior can only be explained by interpretive theories.

When Fred saw the car careen off of the road, he saw something that had the property of being a bad accident, and he called the ambulance *because* what he saw was an accident. He might see another car run off of the road in the same way, and it might be a bad accident too, but Fred might not call an ambulance. (And so, obviously, in that case Fred's seeing an accident would not enter into an explanation of his calling an ambulance.) If this is all true, then Fred's behavior is stimulus-independent. Even the most compulsive people exhibit stimulus-independent behavior. How can we explain this simple fact about people? Very easily: people have beliefs and desires; sometimes they believe that it would be in their best interest to help other people, sometimes they do not. Sometimes they want to help other people, sometimes they do not. We have to appeal to what is going on inside of people in order to explain their behavior.

Of course, this simple little argument is also a terrible little argument. The main idea is that the differences in a person's behavior in the same sorts of situations must result from changes in what's going on inside of him or her. But even if we concede that people must have different internal states at different times, there is still no reason to insist that the states are representational. We can see this clearly by considering the cheapest sort of chess playing computer. It will not always respond the same way to its opponent's moves because sometimes it will be on level 1 and other times it will be on level 2. Certainly, we do not want to ascribe to it the belief that it should play more quickly or better when it is set on level 2. All that we can suppose is that there was some change in the state of the machine that changed the way it played.

7.4 Nonnomic Properties of Stimuli and Representation

Jerry Fodor and Zenon Pylyshyn have presented a related, but considerably more interesting argument for the necessity of representational explanation. Fodor [1986] claims that it is our ability to respond to the "nonnomic" properties of stimuli that requires references to representational states. Zenon Pylyshyn [1984] holds the apparently equivalent position that references to representational states are necessary to account for our ability to respond to properties of stimuli that are not "projectable." Since a precise account of what makes representational explanations necessary in some cases also indicates what sorts of cases do not require representational explanations, Fodor suggests that the ability to respond to nonnomic properties of stimuli is a 'criterion of intentional ascription.' In that way Fodor hopes to draw a nice distinction between representational systems, like us, and nonrepresentational systems, like paramecia.

7.4.1 The Argument

These two claims seem perfectly innocent:

- (a) We sometimes see crumpled shirts.
- (b) The fact that we sometimes see crumpled shirts and the fact that they are crumpled sometimes enters into explanations of our behavior.

Fodor calls these two sorts of claims "truisms." However, they are especially interesting truisms in conjunction with another sort of claim that Fodor has made:

- (c) *Being a crumpled shirt* is a nonnomic property.

According to Fodor, a property is nonnomic just in case nothing is subsumed under a law in virtue of having that property. For example, Fodor suggests that nothing is subsumed under a law in virtue of having properties like *being such and such distance from the Eiffel Tower*, *being a crumpled shirt*, or *being a left shoe*. Of course, there are many laws that hold for left shoes, and for things within one hundred feet of the Eiffel Tower. But that does not entail that laws hold for those things because they are left shoes or within one hundred feet of the Eiffel Tower: left shoes and the Eiffel Tower are subsumed under laws because of their weight, their chemical composition, etc.

If *being a crumpled shirt* is nonnomic, then there cannot be a law such as

- [5] If any person sees a crumpled shirt, then that person will (try to) iron it,

or else something would be subsumed under a law because it is a crumpled shirt. There cannot be any laws of this sort either:

- [6] If any person sees a crumpled shirt, then that person will go into state s_1 .

This is not very surprising, since there are a lot of different ways that crumpled shirts can be put together. Each crumpled shirt will have a different set of physical and chemical properties, so even though two things might both be crumpled shirts, their causal properties will differ widely. It would be surprising if we could build a crumpled shirt "detector" which would give a particular sort of "output" when and only when it was within such and such proximity to a crumpled shirt. Of course, if we could, then *being a crumpled shirt* would be nomic, since the existence of such a machine would show that there is some lawlike relation between *being a crumpled shirt* and the machine's output.³

If people do see crumpled shirts, and if they do behave in certain ways *because* what they see are crumpled shirts--i.e., if the property of *being a crumpled shirt* enters into explanations of their behavior--and if *being a crumpled shirt* is a nonnomic property, then people "respond selectively" to nonnomic properties:

- [7] *A* responds selectively to a nonnomic property if and only if
- (i) *A* perceives something, *S*,
 - (ii) *S* has a property *O*, where *O* is nonnomic,
 - (iii) (i) and (ii) enter into an explanation of *A*'s behavior, *C*.⁴

Having followed Fodor this far, we will surely want to ask the question, 'How can we respond selectively to nonnomic properties of stimuli?' How, for example, can the fact that we see crumpled shirts enter into explanations of our behavior if *being a crumpled shirt* is nonnomic? According to Fodor, it is in answering this question that we recognize the need for representational states:

...there appears to be a puzzle: how does a property of a stimulus come to be implicated in the explanation of a property of a behavioral response if there is no law by which the two properties are connected? *It is largely this puzzle that motivates the representational theory of the mind.* Or, to put it less in the formal mode: selective response to nonnomic properties is, on the present view, the great evolutionary problem that mental representation was invented to solve. [1986, p. 14]

Fodor suggests this solution:

When a stimulus property ... is nonnomic, what connects *S*'s being *O* with *A*'s response coming to be *C* is that *O is a property that A represents S as having*, and the relation between *A*'s representing *S* as *O* and *A*'s behavior coming to be *C* is lawlike (given mediation by other psychological states of *A*'s). [1986, p. 14]

The story, then, is something like this: We know that there cannot be laws like [1] and [2] because *being a crumpled shirt* is a nonnomic property--nothing can be subsumed under a law in virtue of being a crumpled shirt. It would be

surprising if our sensory mechanisms were good enough always to put us in the same state (or equivalence class of states) each time we saw a crumpled shirt. In fact, our sensory mechanisms do not always put us in the same state. We might express this more naturally by saying that sometimes people do not "realize" that what they see are crumpled shirts. Instead, someone might see a crumpled shirt and think that he or she sees a crumpled pair of pants. However, there might be a law of this sort:

- [8] If any person *thinks* he or she sees a crumpled shirt, then that person will (try to) iron it.

Zenon Pylyshyn gives a similar account of why 'it is ... the environment or the antecedent event *as seen or interpreted by the subject*, rather than as described by physics, that is the systematic determiner of actions... '[1984, p. 9]. According to Pylyshyn, we need representational explanations because generalizations that appeal to representational states are often the only generalizations available:

This is not to deny that *some* causal chain connects stimulation and perception. The point is simply that there exist regularities stateable over perceptual (or cognitive) categories that are not stateable over properties of the stimulation itself. This, in turn, is true because it appears that virtually no physical properties (including arbitrarily complex combinations and abstractions over such properties) are necessary and sufficient for the occurrence of certain perceptions; yet it is these perceptions that determine psychological regularities in behavior. In fact, there could be no such properties in general, since ... the way something is perceived can vary radically with physically identical stimuli and can be the same with physically very different stimuli.

Another way to put the matter is to say that organisms can respond selectively to properties of the environment that are not specifiable physically, such properties as being beautiful, being a sentence of English, or being a chair or a shoe. These properties are not properties involved in physical laws; they are not *projectable properties*. ...it is not surprising that an organism reacts to nonphysical or nonprojectable properties of the environment inasmuch as "reacting to an environment" (at least for humans, and probably for many animals) typically involves such processes

as drawing inferences to the best available hypothesis about what, in fact, is out there in the distal world. [1984, p. 15]

The claim that Pylyshyn makes here is certainly not transparent, since it is not obvious what he means by 'projectable properties' or the (apparently synonymous) terms 'properties involved in physical laws,' 'physical properties' and 'properties specifiable physically.' Fortunately, there are several other instances where Pylyshyn uses the word 'projectable' in the same way that Fodor uses 'nomic.'⁵ If that is the case, and if Pylyshyn is using 'projectable property' and 'property specifiable physically' in the same way, then Pylyshyn's claim that some 'organism's can respond selectively to properties of the environment that are not specifiable physically...' is the same as Fodor's claim that some organisms can respond selectively to the nonnomic properties of stimuli.

Pylyshyn provides an example of a person who has witnessed an accident. If *being an accident* is a nonprojectable property, then there cannot be a law such as

[9] If any person sees an accident, then that person will (try to) get help.

or else something would be subsumed under a law in virtue of being an accident. There cannot be a law of this sort either:

[10] If any person person sees an accident, then that person will go into state s_j .

But there might be a law of this sort:

[11] If any person *thinks* he or she sees an accident, then that person will (try to) get help.

Fodor and Pylyshyn claim that if something responds selectively to nonnomic (or nonprojectable) properties of stimuli then it has representational states. Another way to put the claim is to say that the involvement of nonnomic properties of stimuli in the explanation of an organism's behavior guarantees or is

sufficient for the possession of representational states by that organism. Fodor claims that we are distinguished from paramecia, thermostats and other organisms that do not have representational states because nonnomic properties of stimuli can enter into explanations of our behavior but cannot enter into explanations of their behavior. Though he never explicitly says that the ability to respond to nonnomic properties of stimuli is a *necessary* condition for the possession of representational states, there are several passages where Fodor seems to commit himself to that position. For example, Fodor engages in some "empirical speculation" and suggests that all of the "behaviors" of both phototropic organisms and thermostats are lawfully related to what they "see" or "detect". He claims that "...it is *because* this speculation is plausible that it seems clear that paramecium, thermostat, and the rest don't have representational states"[1986, p. 9]. Since Fodor is willing to conclude that paramecia and thermostats do not have representational states simply because he believes that only nomic properties of stimuli are involved in explanations of their "behaviors," it is probably safe to attribute both the necessary and sufficient conditions, and thus the following criterion, to Fodor:

[8] Fodor's Criterion of Intentionality:

An organism has representational states if and only if
it can respond selectively to nonnomic properties of stimuli.

6.4.2 Why Nonnomic Properties?

There may be several plausible objections to both Fodor's and Pylyshyn's claims about nonnomic properties and their relation to representational states. Fodor anticipates several counterexample-style objections in his presentation. One which he does not anticipate in that paper is that for any property, including *being a crumpled shirt* and *being a left shoe*, there is some true statistical statement that subsumes something under it because that thing has that property. Of course, not every true statistical generalization is a statistical law. But we surely could build crumpled shirt detectors and left shoe detectors that work *most of the time*, just like we do. That is a reason to wonder whether *being a crumpled shirt* and *being a left shoe* are nonnomic properties. I think, though, that Fodor and Pylyshyn have an

available and appropriate answer: Suppose there were crumpled shirt detectors that worked reasonably well. They would probably work by detecting typical properties of shirts. In the cases where they did not detect those properties, they would fail to detect the shirt. The detector would also falsely indicate that some things were shirts if they had those properties. *Being a crumpled shirt* is not the statistically relevant property, and even though there is a high probability that the detector will correctly detect shirts, the fact that they are shirts does not explain its success.

I think that there is firmer ground on which to question the position of Fodor and Pylyshyn. I think that even if we accept the claims about nomic and nonnomic properties, it is difficult to understand how nonnomic properties of stimuli do, or even can, enter into explanations of behavior. How can a nonnomic property be implicated in any explanation at all, considering the dominant view in the philosophy of science that the occurrence of events are explained by subsuming those events under causal (or statistical) laws? When Fred saw the crumpled shirt, he saw something that had the property *being a crumpled shirt*, as well as many other properties. Since *being a crumpled shirt* is nonnomic, we cannot, it seems, depend on the fact that a crumpled shirt was what Fred saw when we explain why he ironed the shirt. It appears as if only the *nomic* properties of what Fred saw can be implicated in the explanation.

Fodor offers this account of how a shirt being crumpled can be involved in the explanation of why Fred ironed the shirt--in letters, how *S*'s being *O* can be involved in the explanation of *A*'s behavior coming to be *C*:

...*S* has psychophysical (hence nomic) properties p_1, \dots, p_n as well as (nonnomic) property *O*. In virtue of psychophysical law, causal interaction between *S* and organism *A* eventuates in (nonbehavioral) psychological states s_1, \dots, s_m . In effect, states s_1, \dots, s_m carry the information that *S* has p_1, \dots, p_n , and this information serves as the 'premise' of a perceptual inference of which the 'conclusion' is an attribution of *O* to *S*. Drawing this inference leads to behavioral consequences.... [1986, p. 15]⁶

Presumably, the "perceptual inference" is also governed by some, as yet unknown, causal laws, so nothing mysterious is going on there.

According to Fodor, organisms can respond to nonnomic properties of stimuli because of their causal relations with the nomic properties of the stimuli. But this leaves us with an obvious question: Why couldn't we give explanations that refer only to nomic properties of the stimuli? Nonnomic properties of stimuli enter into explanations of our behavior, it seems, because we "infer" that what we perceive has a particular nonnomic property via the presence of nomic properties that our sensory mechanisms actually do "detect." But consider the "inference" that leads A to the "belief that S is O " (which I will call " s_c "). s_c is a causal consequence of (nonbehavioral) states s_1, \dots, s_m (along with other states of A). s_1, \dots, s_m are in turn caused by the stimulation A received when A saw S . According to Fodor, we can explain how s_1, \dots, s_m came about by referring to S 's being p_1, \dots, p_n , without referring to S 's being O . Since we have been offered a way of explaining A 's behavior without referring to any nonnomic properties, there is no reason to talk about representational states at all.⁷

There might appear to be an obvious reason why we still need references to nonnomic properties of stimuli: We might be able to explain any *token* of A 's behavior by referring to the nomic properties p_1, \dots, p_n , but we might not be able to explain important *types* of behavior that A displays by referring to those same nomic properties. Suppose, for example, that Fred always irons his crumpled shirts. Each of those crumpled shirts has a unique set of nomic properties, so it is unlikely that the same nomic properties are involved in every explanation in all the different cases of shirt ironing. However, the nonnomic property *being a crumpled shirt* is present in each case where Fred irons one of his crumpled shirts, so we might explain that type of behavior by referring to that property. Thus it seems that though we cannot explain certain types of behavior by referring to nomic properties of stimuli we might be able to explain them by referring to nonnomic properties of stimuli. However, if there were a genuine causal (or lawlike statistical) regularity between shirts being crumpled and Fred's subsequent ironings, and if it that regularity held because each of those shirts had the

property of *being a crumpled shirt*, then the supposition that *being a crumpled shirt* is nonnomic would seem to be wrong.

Fodor expected this sort of objection:

... one can imagine [the objection] uttered in the tone of voice: "Show me why I should suppose we ever see shirts at all; show me why I shouldn't suppose instead that what we always do is just detect light structures." I don't, however, propose to take the objection seriously in that form; we do see shirts, and it is loopy to deny that we do, and there's an end of it. [1986, p. 17]

Fodor's claim that it is obvious that "we do see shirts" might be interpreted three different ways: (i) He might be claiming that it is obvious that we sense the light that shirts reflect, though we might not realize that what we see are shirts. I doubt that he means to claim only that much, since it would be consistent to admit that we see crumpled shirts in that way and still deny that some things enter into explanations of our behavior because they are crumpled shirts. (ii) Fodor might be claiming that it is obvious we "recognize" crumpled shirts as crumpled shirts. I doubt he means to claim so much, since he is trying to show that we need to appeal to representational states to explain how organisms can respond to nonnomic properties of stimuli. It would be perverse to begin the argument by simply claiming that we do have certain sorts of representational states. (iii) The only remaining interpretation is that Fodor is simply claiming that it is obvious we sometimes respond to nonnomic properties like *being a crumpled shirt* and *being an accident*, or, in other words, that it is obvious properties like *being a crumpled shirt* and *being an accident* enter into explanations of our behavior. Perhaps he thinks it is obvious because there is no apparent reason to doubt those sorts of properties do enter into explanations of our behavior. But if we can always explain our behavior by referring to nomic properties of stimuli, then references to nonnomic properties of stimuli are simply gratuitous.

7.4.3 Circularity

Though Fodor might think it is obvious that the nonnomic properties of stimuli enter into explanations of our behavior, we are entitled to ask at least this

much: If organisms actually do respond to nonnomic properties of stimuli, how can we tell when they are responding to the nonnomic properties rather than the nomic ones? According to Fodor, whenever a nonnomic property of a stimulus is involved in the explanation of an organism's behavior, there are also causal connections between the crucial "representational" states and nomic properties of the stimulus. In fact, the causal connections are sufficient to bring about those states. Because of this it is difficult to figure out when nonnomic properties of stimuli are involved in bringing about some particular behavior and when merely coextensive nomic properties are involved.

We can appreciate the problem if we consider how Fodor, or anyone else, could ever apply his Criterion of Intentionality. According to that criterion, those things and only those things that can respond selectively to nonnomic properties of stimuli have representational states. Fodor believes that it rules out the behavior of paramecia as suitable material for representational explanation. According to the example, paramecia are negatively phototropic: they (almost) always move away from areas of intense light toward areas of less intense light. Since *being of such and such intensity* seems to be a nomic property of light, we should be able to explain the behavior of paramecia by referring to nomic properties. But how do we know that they do not respond to, say, the sun? Paramecia usually retreat from the sun, since the sun is usually the source of the most intense light. Why shouldn't we say that paramecia "recognize" the source of the light as the sun? When they retreat from a different light source we could say that they "misinterpreted" it as the sun. Presumably, *being the sun* is an example of a nonnomic property. (Certainly the sun is subsumed under laws, but only because it has the property of *having such and such mass, having such and such position, producing such and such heat*, not because it is that particular star. If the sun were replaced by a different star of the same mass, producing the same amount of radiation, etc., the change would not effect the rest of the solar system.) If paramecia do respond to the sun because it is the sun, rather than simply because of its nomic properties, then, according to Fodor, the behavior of paramecia is suitable for representational explanation.

It might seem that we can rule out the possibility that the nonnomic property *being the sun* is involved in bringing about the behavior of the paramecia simply because we can explain their behavior without referring to the sun--all we have to refer to is light intensity. But, according to Fodor's own account of how organisms respond to nonnomic properties, it appears as if it is not necessary to refer to nonnomic properties at all in order to explain our own behavior. How, then, could we know when nonnomic properties of stimuli are involved in explanations of behavior and when they are not? This, according to Fodor, is a reasonable way to put the "loopy" skeptical objection:

Supposing...that selective response to nonnomic properties is a real, not to say commonplace, phenomenon; and supposing too that, when it occurs, it is mediated by inference and mental representation, it is nevertheless reasonable to ask how bona fide occasions of selective response to nonnomic properties are to be distinguished from occasions on which *all* that happens is that some locally coextensive psychophysical property is detected.

I haven't any *criteria* for drawing this distinction, on account of there not being any. But I'll suggest some indexes: sorts of considerations that tend to tip the balance in favor of explanations that advert to *S*'s being *O* (and not just to *S*'s being p_1, \dots, p_n) in accounting for *A*'s behavior becoming *C*. [1986, p.17]

Fodor suggests three indices: (i) The involvement of a nonnomic property *O* (and not just p_1, \dots, p_n) in the etiology of *A*'s behavior coming to be *C* is indicated if *A* says '*S* is *O*' (rather than '*S* is p_1, \dots, p_n '). [1986, pp. 17-18]

(ii) The involvement of a nonnomic property *O* (and not just p_1, \dots, p_n) in the etiology of *A*'s behavior coming to be *C* is indicated 'whenever it turns out that p_1, \dots, p_n are *sufficient but not necessary* for the behavior coming to be *C*. This is, in fact, the *normal* relation between psychophysical properties and perceptual categories' [1986, p. 19]. There are many different ways that crumpled shirts can be put together, and so there are a lot of different nomic properties that crumpled shirts can have. However, the *behavior* of the perceiver bent upon crumpled shirt

identification will be largely insensitive to these sorts of variations; it will correlate *better* with the presence of O than with the presence of p_1, \dots, p_n ' [1986, p. 19].

(iii) The involvement of a nonnomic property O (and not just p_1, \dots, p_n) in the etiology of A 's behavior coming to be C is indicated if that type of behavior is not exhibited by other organisms in 'circumstances normally sufficient for seeing S ' [1986, p. 19]. What we come to "interpret" certain stimuli as depends on what we know, and how we behave depends on what we interpret the stimuli as. So we would expect that organisms with different "information" might act differently in the presence of the same p_1, \dots, p_n .

The first index gives us no reason to think that paramecia ever respond to nonnomic properties of stimuli. Since the behavior of paramecia correlates better with light intensity than with the presence of sunlight, the second index does not provide a reason to think that the property *being the sun* is involved in the explanation of that behavior. All paramecia, according to the assumptions of the example, behave the same way when they are in the presence of sunlight, so the third index does not give us any reason to believe that the property *being the sun* is involved in the explanation of their behavior.

(i) - (iii) do draw a line between paramecia and humans, but, since they are only indices and not criteria, they do not rule out the possibility that paramecia do respond to nonnomic properties such as *being the sun*. I think that with a little thought it is easy to see why Fodor does not claim that indices are actually criteria for when organisms respond to nonnomic properties of stimuli. I think, in fact, that a little imagination and a healthy dose of skepticism can provide some good reasons to doubt whether the indices provide either necessary or sufficient conditions for when organisms respond to nonnomic properties of stimuli.

Of course, Fodor's indices may still succeed as indices. They are intended only to suggest when nonnomic properties of stimuli are involved in bringing about particular behaviors. But how could anyone be *sure* that they are good indices? How could we discover whether they are reliable? Consider the two sorts of explanations that we might get for a particular organism's behavior: The behavior might be explained as a response to nonnomic properties of stimuli, or the behavior might be explained as a response to merely nomic properties of stimuli.

In both cases some nomic properties of the stimuli, p_1, \dots, p_n , bring about (nonbehavioral) states s_1, \dots, s_m . These in turn bring about a behavioral state, s_c . The only difference between a case where a nonnomic property of stimuli is involved in determining behavior and a case where coextensive nomic properties of stimuli are involved appears to be that when a nonnomic property is involved s_c is a representational state, but where coextensive nomic properties are involved s_c may not be a representational state. On Fodor's account, we can determine cases where organisms respond to nonnomic properties of stimuli from cases where they do not only if we already know whether s_c is a representational state (and even then we would often go wrong when organisms make mistakes).

If we are using response to nonnomic properties of stimuli as our criterion of intentionality, then we have things backwards if we try to point to an organism's representational states as evidence that nonnomic stimulus properties are involved in the explanation of their behavior. However, Fodor appears to have done just that. For example, Fodor's most plausible index is that if A produces a verbal representation, ' S is O ', then O is probably involved in A 's behavior being whatever it is. Fodor also suggests that we will be able to provide more indices after we have learned more about mental representation and the phenomena mental representation can be invoked to explain.

None of this shows that Fodor's Criterion of Intentionality is wrong. But it does show that the criterion will not get us anywhere. Suppose, once again, that we try to apply it to paramecia. It tells us they have representational states if and only if they respond selectively to nonnomic properties of stimuli. Unfortunately, the only solid evidence we could have that paramecia respond to nonnomic properties of stimuli is evidence that they have certain sorts of representational states. We need some other criterion to determine that.

Now consider the argument for representational explanation that both Fodor and Pylyshyn have suggested. Given Fodor's account of how nonnomic properties of stimuli are involved in the explanation of behavior, that argument has a similar problem. The initial premise of the argument is that we do see things like crumpled shirts and accidents and the fact that what we see are crumpled shirts and accidents

can enter into explanations of our behavior. But what evidence could we have for the claim that nonnomic properties like *being a crumpled shirt* and *being an accident* enter into explanations of our behavior besides previous evidence that we really do have certain representational states? What could make Fodor's claim that we respond to nonnomic properties so obvious except for the conviction that we really do see crumpled shirts *as* crumpled shirts? If that conviction is the premise for the argument for representational explanation, then the argument just begs the question.

7.4.4 More Mystery

So far my criticisms have centered on Fodor's account of how organisms respond to nonnomic properties of stimuli. I have argued that, given his account, we could replace any explanation of behavior that adverts to a nonnomic property of stimuli with one that adverts only to coextensive nomic properties of stimuli. I have also argued that, given his account, the only evidence we could have that nonnomic properties are involved in bringing about a particular piece of behavior would be evidence that (at least one of) the relevant behavioral state(s) was a representational state. Both of these claims rest on this observation: Given Fodor's account of how organisms respond to nonnomic properties of stimuli, there is *no difference* between a case where a nonnomic property is involved in determining behavior and a case where coextensive nomic properties are involved, except that when a nonnomic property is involved a representational state is also involved.

My next observation may be obvious by now: despite his claims to the contrary, Fodor has not actually provided an account of how organisms respond to nonnomic properties of stimuli. According to Fodor, a nonnomic property O of a stimulus S can enter into the explanation of A 's behavior because A can "represent" S as being O . But this does not explain how S 's *being* O enters into the explanation. For example, we are supposed to be able to explain why Fred called for help after witnessing an accident by pointing out that the nomic properties of the accident caused him to come to have states s_1, \dots, s_m and that he then "inferred" that the thing he saw was an accident. Because he figured out that the thing he saw was an accident, he called the police. Fred's "representing" what he saw as an accident

enters into the explanation of Fred's behavior, but it does not seem necessary to point out that what Fred saw was in fact an accident. Suppose what Fred saw had instead been a Hollywood stunt. We might be able to give *exactly the same* account of why Fred called the police: The nomic properties of the stunt caused him to come to have states s_1, \dots, s_m and he then "inferred" that the thing he witnessed was an accident. Fodor's story about how nonnomic properties of stimuli are involved in explanations of behavior provides a good reason to believe that they are *not* involved in explanations of behavior.

One of Fodor's statements may help here:

S's being O thus enters into the story twice; once in rerum natura, and once as represented. Very roughly, the first occurrence is required in order that the truth conditions of A's perceptual belief should be satisfied... . And the second occurrence is required in order that S's being O should have consequences for A's behavior, and in order that those consequences should be specific to the property O (as opposed, for example, to merely coextensive properties). [1986, pp.14 - 15]

Fodor is suggesting something like this: If *A* behaves in a certain way because what *A* sees, *S*, has a nonnomic property *O* (and not merely because *S* has some other, nomic, properties), then two things must happen: (i) *A* must "represent" *S* as being *O*, and (ii) *if A is actually responding to S's being O, then when A represents S as being O, S had better be O.*

We do have to refer to *S's being O* to explain how *A* responds to *S's being O*. This, however, does not tell us why we have to refer to *S's being O* in order to explain *A's* behavior. Instead, we apparently only have to refer to *S's being O* in order to distinguish *A's correctly* coming to believe that *S* is *O* (with the consequence that *A's* behavior comes to be *C*) from *A's mistakenly* coming to believe that *S* is *O* (with the consequence that *A's* behavior comes to be *C*). We do not have to refer to *S's being O* in an explanation of *A's* behavior coming to be *C*. Of course, if nonnomic properties of stimuli do not figure in the explanation of our behavior, then any arguments premised on that claim fail immediately, and Fodor's Criterion of Intentionality rules that none of us is a representational system.⁸

None of this should be terribly surprising--after all, nonnomic properties are nonnomic. The dominant view in the philosophy of science is that events are explained by subsuming them under scientific laws, and, since no events are subsumed under scientific laws because of their nonnomic properties, it would be very surprising if nonnomic properties of stimuli entered into the explanation of any sort of behavior. Appealing to representational states does not provide a link between nonnomic properties and behavior. However, if all of this is obvious, I may be guilty of misrepresenting Fodor's position. Perhaps Fodor did not intend to be taken literally when he claimed that nonnomic properties of stimuli are involved in explanations of behavior, but instead meant to claim that nonnomic properties of stimuli are involved in the etiology of a behavioral response *only* as they are represented by organisms. But, there are two reasons to reject this sort interpretation: it would mean ignoring many instances where Fodor seems to explicitly make the stronger claim, and, under this interpretation Fodor's argument for representational explanation would be circular and his Criterion of Intentionality would be trivial.

7.4.5 Another Try

There is another strategy for providing a link between the nonnomic properties of stimuli and representational states, and it has the advantages of being somewhat plausible, resembling some of what Fodor has written, and of not resting on the claim that we actually do respond to nonnomic properties of stimuli. (Of course, at the same time it has the defect of ignoring the many instances where Fodor appears to explicitly claim that we do respond to nonnomic properties of stimuli.) According to Fodor's account of how we respond to stimuli, it seems as if we can explain each instance of a particular type of behavior by appealing only to the nomic properties of stimuli. But, since different sets of nomic properties would probably be involved in the different cases, the various instances of a particular type of behavior would probably receive very different explanations. For example, there might not be any set of nomic properties that is both sufficient to cause Fred's call for help and present in every case where Fred sees an accident. As Fodor pointed out, Fred's behavior might correlate better with the presence of a

nonnomic property, such as *being an accident*, than with any set of nomic properties. That does not mean that a nonnomic stimulus property is involved in bringing about Fred's behavior, but the fact that the nonnomic stimulus property correlates so well with the behavior appears to require some explanation. We may have to appeal to representational states because Fred and the rest of us *seem* (at least pretheoretically) to respond to nonnomic properties of stimuli: the fact that it *seems* Fred calls for help because what he sees are accidents may be evidence that he represents what he sees as accidents.

Of course, that sort of correlation is not enough to support the claim that human behavior is best understood in representational terms. Anyone who believes that correlations between nonnomic stimulus properties and behavior can be taken seriously as evidence that representational states are involved in bringing about behavior has (at least) two related tasks. The first task is to pin down instances where particular nonnomic stimulus properties *seem* to be involved. That is not easy to do, since there are many different correlations between any given type of behavior and the various nonnomic properties of the stimuli that are causally implicated in bringing about that behavior. For example, suppose that Fred always calls for help when he witnesses accidents and that he is always "fooled" by Hollywood stunts and calls for help in those cases too. It would be natural to treat the correlation between the accidents and Fred's behavior as the important one and claim that he *seems* to be responding to the fact that what he sees are accidents. But, of course, there is also a good correlation between Fred's behavior and the presence of accidents or Hollywood stunts, so why not say that he *seems* to be responding to accidents or Hollywood stunts? Given any type of behavior, there will always be a perfect correlation between instances of that type of behavior and some nonnomic property of the stimuli whose nomic properties are involved in bringing about the behavior. For example, an organism *A* performs a behavior of type *C* as a result of seeing something *S*, when and only when *S* has the nonnomic property *being such that its nomic properties are involved in bringing about A's performance of a behavior of type C*. In fact, there are an infinite number of trivial examples of nonnomic properties that will yield perfect correlations.⁹

If the existence of an (imperfect) correlation between the presence of a nonnomic stimulus property and an organism's behavior is evidence for the claim that representational states do play a part in explaining the organism's behavior, then there must be some principled way to dismiss the (infinitely many) perfect and (infinitely many) imperfect competing correlations. Since nonnomic properties of stimuli cannot be involved in bringing about any behavior, we cannot single out correlations on the grounds that the nonnomic property in question is involved in bringing about the relevant sort of behavior. I do not see any obvious way to single out one correlation from among the many.

If one correlation can be picked out, the second task is to show why, exactly, we need to explain it by referring to representational states. If Fred *seems* to respond to accidents, even though the property *being an accident* is nonnomic, then, presumably, the instances of Fred's behavior have something in common. Since different sets of nomic properties will probably be involved in bringing about the different instances of Fred's behavior, we might suppose that what is common to them is that the different sets of nomic properties, with the help of some of Fred's "mental" states, bring about the same type of state in Fred. It may be natural to characterize that state as representing an accident, and to characterize the causal processes as inferences, but why, exactly, must we characterize them that way? The hard questions remain unanswered.

NOTES

- 1 Churchland made this suggestion to Stich, who considers it in [Stich 1983 pp. 171-181]. [1] and [3] are taken almost directly from his presentation of Churchland's suggestion.
- 2 Pylyshyn recognizes this. See the footnote [1984, p. 164].
- 3 It might appear as if *we* are crumpled shirt detectors, since we locate crumpled shirts all the time. But we apparently make mistakes, so we don't recognize crumpled shirts *when and only when* we are within a certain proximity to them. If we could, then *being a crumpled shirt* would seem to be a nomic property.
Of course, we "recognize" crumpled shirts most of the time, so that might be reason to suppose that there is a lawlike connection between crumpled shirts and our "psychological states"--a statistical connection (see the beginning of section 7.4.2).
- 4 This is my interpretation what Fodor means by 'responds selectively.' This interpretation is suggested by his comments in his [1986], especially, pp. 6-9.
- 5 For example, in a paper which they coauthored, Pylyshyn and Fodor [1981, p. 146] (informally) define the word 'projectible' in the same way that Fodor [1986] (informally) defines 'nomic.' (I assume that the orthographic difference between 'projectable' and 'projectible' is unimportant.) Also, Pylyshyn claims that Fodor (in an early, unpublished draft of Fodor [1986]) 'views a system's capacity to respond selectively to nonprojectable properties as precisely what differentiates inferencing systems from systems that merely react causally to environmental stimulation' [1986, p. 15]. Since Fodor argues that the capacity to respond selectively to *nonnomic* properties is what differentiates inferencing systems from systems that merely react causally to environmental stimulation, it is probably safe to conclude that Pylyshyn uses 'projectable' in the same way that Fodor uses 'nomic.'
- 6 This quotation is from Fodor's remarks on his Figure 1.
- 7 Dan Lloyd [1986] presents a similar objection.
- 8 Since we have to refer to *S*'s having a nonnomic property *O* in order to distinguish *A*'s *correctly* coming to believe that *S* is *O* from *A*'s *mistakenly*

coming to believe that S is O , one might consider that a good reason to believe we do respond to nonnomic properties. But, of course, that would presuppose that representational states play a role in cognitive explanation.

- 9 For example, *being such that its nomic properties are involved in bringing about A 's performance of a behavior of type C , and such that n is a number, for any natural number n .*

CHAPTER 8

ANOTHER COMPUTER EXAMPLE

8.1 Introduction

In defense of the Classical Computational Theory of Mind, Fodor and Pylyshyn have argued that explanations of cognitive capacities and behavior must be explained in terms of our relations to material representations, and they have argued that those representations have syntactic structure. They also claim that those representations have semantic "content," and that cognitive generalizations subsume cognitive states in virtue of those semantic properties. But (according to the Strong Correlation Thesis) the individuation of representations by semantic type corresponds to individuation by syntactic type. Thus it seems that for every lawlike generalization that subsumes a pair of cognitive states in virtue of the semantic properties of their constituent representations there would be a lawlike generalization subsuming that same pair in virtue of the syntactic properties of those representations. Those latter laws would enable us to predict just as accurately as the former laws, and they would allow us to discover the causes of whatever we could explain using the former laws. The two sorts of laws seem to "capture the same generalizations."

That, it seems to me, is a good reason to maintain that, given CCTM, references to semantic properties are not essential to explaining human cognitive processes or the behavior that results from those processes. But, of course, I have not convinced everyone. Neil Stillings probably spoke for many when he wrote this note:

Take an old-saw computer example: an accounting program written in Pascal. We can give a purely syntactic account of why it does what it does on a variety of different computers. But we seem to learn something additional when we read the documentation and discover that the thing balances books by carrying out algorithms that have been proven correct in the accounting literature. (personal correspondence)

I will not deny any of this. Nor will I deny that we would learn something more were we informed about the content of cognitive states and not just the

syntactic type of their constituent representations. The hard part is determining what exactly it is that we would learn and whether that information is essential to explaining the relevant behavior, or whether it just makes the explanation a little easier to understand.

In this chapter I will take an old-saw computer example and try to figure out just what we would miss if we gave purely syntactic explanations of computer behavior. Presumably, whatever we would miss in an explanation of the computer's behavior we would also miss in an explanation of human behavior.

8.2 An Old-Saw Computer Example

Suppose we have several computers that (1) run the same (accounting) program, and (2) are similar enough functionally so that their computational states will be of the same "semantic" type (under the "accounting interpretation") if and only if they are of the same functional/syntactic type. This second requirement puts the computers in the position we are in, according to CCTM.

I will make two suggestions about what it is that we miss unless the states and processes of the computer are given a semantic interpretation:

(1) Referring to computer states by their functional or syntactic properties may explain why the computers display symbols on the screen when others are typed in on the keyboard, but it does not explain why the computers carry out accounting functions or calculate the correct answers. The computers compute information, and until we supply an interpretation we cannot understand why they give us the appropriate answer instead of merely displaying some symbols on the screen.

This sounds a lot like things that Fodor and Pylyshyn, and others, have written. There is a *prima facie* plausible reply. Suppose a computer displays a string of symbols, A , and we interpret that "behavior" as coming to the conclusion that φ . Events can be described in different ways, and it seems that displaying- A is the same event as concluding-that- φ . Surely if we were to explain why the computer displayed A (by referring only to the functional/syntactic properties of

the computer states), and if displaying- A is the same event as concluding-that- φ , then we would have explained why the computer concluded that φ .

There are three possible objections: (1) One might argue that displaying- A is not the same event as concluding-that- φ . I am not sure how exactly one would attack or defend this claim, since it brings up a great number of very difficult questions. But if it were successful we would be left with the problem of explaining how the computer concluded that φ . I see only two possibilities: (i) The concluding-that- φ event was caused by the displaying- A event (which can be explained by referring only to the functional/syntactic properties of the computer states). But if this were true, the concluding-that- φ event would not be caused by the other "contentful" states of the computer, and we would really have to wonder what semantic interpretation is for. (ii) The concluding-that- φ event was caused by other contentful states of the computer (which are different from the internal functional states of the computer). On this account the computer's actual computations/inferences would parallel, but not interact with, its functional processes. This strikes me as very mysterious.

(2) One might plausibly claim that the displaying- A event and the concluding-that- φ event are identical, but that events are explained "under descriptions," and the explanation of an event under one description is not a explanation of the same event under another description.¹ Pylyshyn puts the point a little differently:

There are general reasons why one account of a sequence of events might qualify as an explanation while another *true* account of the same sequence does not. These reasons have to do with the fact that such claims as "The occurrence of X (together with...) *explains* the occurrence of Y " are not, in general, equivalent (that is, they need not preserve truth values) when we replace the X or the Y by phrases that refer to the same event or to the same objects (in general, "explains" provides what philosophers refer to as an "opaque context"). [Pylyshyn, 1984, p. 4]²

In one sense it is a truism that "explains" provides an opaque context and that events are explained under descriptions. Laws that govern transitions between

events subsume those events because of their properties. Conventional wisdom tells us that explanations (at least of the causal variety) consist of a law, a statement of initial conditions (which is a substitution instance of law's antecedent) and a statement of the event to be explained (which is a substitution instance of law's consequent). If in an explanation the event-to-be-explained were characterized by a property different from the property expressed in the consequent of the transition law (say if it were described as an F instead of as a G), the explanatory inference would simply be invalid. But this is not a good reason to suggest that if we were to provide a proper explanation of the event as a G we would not have explained that same event were it was described as an F.

Fodor and Pylyshyn have used the claim that events are explained under descriptions to argue for the "autonomy" of psychology. The argument sounds familiar: Consider a science, S , that is intended to explain events in systems like s , and that includes among its typical predicates two predicates, ' S^1 ' and ' S^2 ', that express the properties in virtue of which the two events are subsumed under the relevant law. Events that are typically described by ' S^1 ' in S might have multiple instantiations in physics. Though we might truly say that ' $S^1 s$ ' in the vocabulary of S , an "equivalent" statement in physics might look like ' $(P^l a \text{ and } \dots \text{ and } P^d d) \text{ or } \dots \text{ or } (P^e e \text{ and } \dots \text{ and } P^h h)$.' Thus particular transitions that are similar when stated in the vocabulary of S will seem to have nothing in common when stated in the language of physics: a generalization is lost when the transition is "described" in the vocabulary of physics.

However, our computer example is not analogous to this case. By hypothesis, for every lawlike generalization that subsumes a pair of computer states in virtue of their semantic properties there is a lawlike generalization subsuming that same pair in virtue of their functional/syntactic properties. I do not see how the loss of generalizations argument would apply in this case.

There might be some tendency to feel that referring to the computer's functionally/syntactically described processes in an attempt to explain why the computer concludes that φ is simply a "category mistake": the functional/syntactic level and the semantic level are not compatible. But it seems obvious that in some cases we can explain the semantically interpreted outputs of computers without

referring to their other semantically interpreted states. Suppose, for instance, that instead of giving me the correct output, the computer displayed B on the screen--a series of symbols that I interpret as a conclusion that ψ . Presumably, if the computer suddenly made an uncharacteristic mistake, we would look to its electrical properties to find the glitch that caused B to be displayed on the screen. There we would have a case where a semantically interpreted output was explained by referring to the electrical components of the machine. Going to a "lower-level" for an explanation would not be a rare event. As Fodor pointed out (back in chapter 3),

We allow the generalizations of the non-physical sciences to *have* exceptions, thus preserving the kinds to which the generalizations apply. But since we know that the *physical* descriptions of the members of these kinds may be quite heterogeneous, and since we know that the physical mechanisms which connect the satisfaction of the antecedents of such generalizations to the satisfaction of their consequents may be equally diverse, we expect both that there will be exceptions to the generalizations and that they will be 'explained away' at the level of the reducing science. [1974, p. 112]

Either we must admit that the computer's conclusion that ψ is explained at some other level than the level of semantic interpretation, or we must admit that it cannot be explained at all. It seems clear, then, that referring to the functional/syntactic level can explain the "semantically interpreted behavior" of the computer in the exceptional cases. I do not see any principled way of arguing that referring to the functional/syntactic level cannot explain the normal behavior of the computer.

(3) One might maintain that without referring to the semantic content of the computer's states one cannot even explain why the computer displayed A , much less explain why it concluded that φ . But this seems much too extreme. Suppose a mad electrical engineer built a computer and programmed it so that it had processes of the same functional/syntactic type as the computers in our example. Suppose also that he did not have an intended use for the computer, and so did not provide and interpretation of its processes. Surely he would be able to explain its behavior in the absence of such an interpretation.

This is my second suggestion about what it is that we miss unless the states and processes of computers are given semantic interpretations:

(II) Computers are useful because we can interpret their inputs and outputs. Though we might be able to explain why computers behave as they do without providing an interpretation of their internal processes, we cannot explain why it is appropriate to attribute content to the inputs and outputs without providing an interpretation of the processes in between. For example, though we might explain what *caused* a computer to conclude that φ without providing a semantic interpretation of its internal processes, we do not know what makes the computer's output a conclusion-that- φ . What has not been explained is how *being a conclusion that φ* is *instantiated* by the output, and that cannot be done without interpreting the other states of the computer.

I want to avoid the whole issue of how semantic properties are instantiated in states of computers and people (mostly because I do not know what to say). I suppose the first requirement is that it be possible to map meaningful things like sentences and propositions to those states in such a way that the transitions between the states mirror the formal or semantic relations between the semantic objects (in a way that an interpretive functional theory could exploit).

In any case, (II) may be a non-starter. We attribute content to the inputs and outputs of many systems without ever supposing that they have contentful internal states. For example, by measuring the change in air pressure (with a bellows) and compensating for the temperature (with a coil thermometer), a sophisticated altimeter will provide fairly accurate reports of your altitude if you tell it the original altitude (by setting the altitude dial). But people do not seem to have any inclination to say that altimeters "process information." The reason that we can attribute content to the inputs and outputs is that under the correct interpretation the altimeter provides the right outputs for those inputs. That, it seems to me, is why it is appropriate to attribute content to the inputs and outputs of

a computer: under an appropriate interpretation it provides the right outputs for those inputs--the nature of its internal processes is irrelevant.

8.3 Why Interpretation?

Why, then, do we typically attribute content to computational processes and states when we program computers and when we try to figure out how computers work? The answer is easy: we attribute content to those states because it is easier that way--interpretation allows us to exploit what we already understand. But just because interpretation makes it easier to understand a computer's processes does not mean that that interpretation is essential to an explanation of the computer's behavior.

Suppose, for (my very last) example, we discovered a Martian computer that carried out complex computations that we found very useful, but had no idea how to calculate ourselves. By investigating the computer and how it functioned--by investigating the functional processes of that computer--we could make other computers that carried out the same computations. Of course we still might not understand how the internal functional processes were supposed to be interpreted. There are only two choices: (1) We know why the computers provide the right output, even in the absence of the intended interpretation, or (2) we do not know why the computers provide the right output, because we lack an interpretation of their internal functional processes. Those who would choose the second alternative seem to put very strict requirements on explanations.

- 1 The plausibility of the claim that events are explained under descriptions will depend a great deal on how events are individuated. Is pendulum p swinging in an arc of 30 degrees at t the same event as pendulum p swinging from east to west at t? Certainly pendulums can swing in 30 degree arcs and from east to west at the same time. Presumably, though, an explanation of why one swung in an arc of 30 degrees at t would not be an explanation of why it swung from east to west at t. I am not, however, going to discuss how events are individuated--that is another dissertation. I will proceed without any exact criterion of event identity.

- 2 Pylyshyn presents his case in [Pylyshyn, 1984, chapter 1]. He attributes the idea to an unpublished paper by Ned Block and Fodor.

BIBLIOGRAPHY

- Anderson, A. (ed.). 1960. *Minds and Machines*. New Jersey: Prentice Hall
- Baker, Lynne Rudder. 1987. *Saving Belief: A Critique of Physicalism*. Princeton, New Jersey: Princeton University Press.
- Berney, Ronald. 1964. *Understanding Digital Computers*. New York: John F. Rider Publisher, Inc.
- Block, Ned (ed.). 1980. *Readings in Philosophy of Psychology*, vol. I. Cambridge, Massachusetts: Harvard University Press.
- Block, Ned (ed.). 1981. *Readings in Philosophy of Psychology*, vol. II. Cambridge, Massachusetts: Harvard University Press.
- Boyd, Richard. 1980. 'Materialism Without Reductionism: What Materialism does not Entail.' In [Block, 1980].
- Burge, Tyler. 1986. 'Individualism and Psychology.' *The Philosophical Review*, XCV No. 1.: 3-46 (January, 1986).
- Carnap, Rudolph. 1966. *An Introduction to The Philosophy of Science*. (edited by Martin Gardner.) New York: Basic Books, 1974. (Originally published as *An Introduction to The Philosophy of Physics* in 1966.)
- Churchland, Patricia. 1986. *Neurophilosophy, Toward a Unified Science of the Mind/Brain*. Cambridge, Massachusetts: MIT Press.
- Churchland, Paul. 1984. *Matter and Consciousness*. Cambridge, Massachusetts: MIT Press.
- Cummins, Robert. 1975. 'Functional Analysis.' *Journal of Philosophy* 72:741-760, (November, 1975). Reprinted in [Block, 1980].
- Cummins, Robert. 1983. *The Nature of Psychological Explanation*. Cambridge, Massachusetts: MIT Press.

- Cummins, Robert. 1988. 'Critical Notice: *Computation and Cognition*.' *Canadian Journal of Philosophy* 18 vol. 1:147-162.
- Davidson, Donald. 1970. 'Mental Events.' In [Foster and Swanson, 1970]. Reprinted in [Block, 1980].
- Dennett, Daniel. 1978a. *Brainstorms*. Cambridge Massachusetts: MIT Press.
- Dennett, Daniel. 1978b. 'Toward a Cognitive Theory of Consciousness.' In [Savage, 1978], reprinted in [Dennett, 1978a].
- Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge, Massachusetts: MIT Press.
- Dreyfus, H., and Dreyfus, S. (forthcoming). *Making a Mind vs Modeling the Brain*. Deadelus Press.
- Feigl, Herbert, and Maxwell, Grover (eds.). 1958. *Minnesota Studies in the Philosophy of Science* 2. Minneapolis: University of Minnesota Press.
- Feigl, Herbert, and Maxwell, Grover (eds.). 1962. *Minnesota Studies in the Philosophy of Science* 3. Minneapolis: University of Minnesota Press.
- Feldman, Fred. 1980. 'Identity, Necessity and Events.' In [Block, 1980].
- Fetzer, James. 1981. 'Probability and Explanation.' *Synthese* 48:371-408.
- Fodor, Jerry. 1974. 'Special Sciences (or: The Disunity of Science as a Working Hypothesis).' *Synthese* 28:97-115. Reprinted as 'Special Sciences' in [Fodor, 1981].
- Fodor, Jerry. 1975. *The Language of Thought*. New York: Crowell Company.
- Fodor, Jerry. 1978a. 'Computation and Reduction.' In [Savage, 1978], reprinted in [Fodor, 1981].
- Fodor, Jerry. 1978b. 'Propositional Attitudes.' *The Monist* 61,4. Reprinted in [Fodor, 1981].

- Fodor, Jerry. 1980. 'Methodological Solipsism Considered as a Research Strategy in Psychology.' *The Behavioral and Brain Sciences* 3:63-109. Reprinted in [Fodor, 1980].
- Fodor, Jerry. 1981. *Representations*. Cambridge Massachusetts: MIT Press.
- Fodor, Jerry. 1986. 'Why Paramecia don't have Mental Representations.' In [French, P., Uehling, T., and Wettstein, H., 1986].
- Fodor, Jerry. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry, and Pylyshyn, Zenon. 1981. 'How Direct is Visual Perception?: Some Reflections on Gibson's "Ecological Approach."' *Cognition*. 9:139-196.
- Fodor, Jerry, and Pylyshyn, Zenon. 1987. 'Connectionism and Cognitive Architecture: A Critical Analysis.' Manuscript copy.
- Foster, L. and Swanson, J. W. (eds.). 1970. *Experience and Theory*. Amherst, Massachusetts: University of Massachusetts Press.
- French, Peter, Uehling, T., and Wettstein, H. (eds.). 1986. *Midwest Studies in Philosophy Volume X. Studies in the Philosophy of Mind*. Minneapolis: University of Minnesota Press.
- Garfield, Jay. 1988. *Belief in Psychology: A Study in the Ontology of Mind*. Cambridge, Massachusetts: MIT Press.
- Haber, R. N. (ed.). 1969. *Contemporary Theory and Research in Visual Perception*. New York: Holt, Rinehart & Winston.
- Hempel, Carl. 1958. 'The Theoritician's Dilemma.' In [Feigl and Maxwell, 1958].
- Hempel, Carl. 1962. 'Deductive-Nomological vs. Statistical Explanation.' In [Feigl, Herbert and Maxwell, 1962].
- Hempel, Carl. 1965a. *Aspects of Science and Other Essays in the Philosophy of Science*. New York: Free Press.
- Hempel, Carl. 1965b. 'Aspects of Scientific Explanation.' In [Hempel, 1965].

- Hempel, Carl, and Oppenheim, Paul. 1948. 'Studies in the Logic of Explanation' *Philosophy of Science* 15:135-175. Reprinted in [Hempel, 1965].
- Hochberg, Julian. 1969. 'In the Mind's Eye.' In [Haber, 1969].
- Johnson-Laird, Philip N. 1983. *Mental Models*. Cambridge, Massachusetts: Harvard University Press.
- Julesz, B. 1971. *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press.
- Julesz, B. 1975. 'Experiments in the Visual Perception of Texture.' *Scientific American*, 232:34-43.
- Kim, Jaegwon. 1982. 'Psychophysical Supervenience.' *Philosophical Studies*, 41:51-70.
- Kim, Jaegwon. 1989. 'The Myth of Nonreductive Materialism.' *Proceedings and Addresses of The American Philosophical Association*, 63, 3:31-47.
- Kripke, Saul. 1972. *Naming and Necessity*. Cambridge, Massachusetts: Harvard University Press.
- Lewis, David. 1970. 'How to Define Theoretical Terms.' *Journal of Philosophy*, 67:427-46.
- Lloyd, Dan. 1986. 'The Limits of Cognitive Liberalism.' *Behaviorism*, 14:1-14.
- Loar, Brian. 1981. *Minds and Machines*. Cambridge: Cambridge University Press.
- Mackie, J. L. 1974. *The Cement of the Universe*. Oxford: Clarendon Press.
- Marr, David. 1982. *Vision*. New York: W. H. Freeman and Company.
- McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. 1986. 'The Appeal of Parallel Distributed Processing.' In [Rumelhart and McClelland, 1986].
- Nagel, Ernest. 1961. *The Structure of Science*. Indianapolis, Illinois: Hackett.

- Nelson, Alan. 1985. 'Physical Properties.' *Pacific Philosophical Quarterly*, 66:268-282.
- Parks, T. 1965. "Post-Retinal Visual Storage." *American Journal of Psychology*, 78:145-147.
- Putnam, Hilary. 1960. 'Minds and Machines.' In [Anderson, 1960].
- Putnam, Hilary. 1973. 'Reductionism and the Nature of Psychology.' *Cognition*, 2:131-146.
- Putnam, Hilary. 1975. 'The Meaning of Meaning.' In [Putnam, 1987].
- Putnam, Hilary. 1987. *Mind Language and Reality*, Cambridge: Cambridge University Press.
- Pylyshyn, Zenon. 1980. 'Cognitive Representation and the Process-Architecture Distinction.' *Behavioral and Brain Sciences*, 3, 1.
- Pylyshyn, Zenon. 1984. *Computation and Cognition*. Cambridge, Massachusetts: MIT Press.
- Reicher, C. M. 1969. 'Perception Recognition as a Function of Meaningfulness of Stimulus Material.' *Journal of Experimental Psychology*, 81:275-80.
- Rumelhart, D. E. and McClelland, J. L.(eds.). 1986. *Parallel Distributed Processing, Vol. 1*. Cambridge: MIT Press.
- Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble, 1949.
- Salmon, Wesley. 1965. 'The Status of Prior Possibilities in Statistical Explanation.' *Philosophy of Science*, 32:137-146.
- Salmon, Wesley. 1967 *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.
- Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, New Jersey: Princeton University Press.

- Savage, C. W. (ed.). 1978. *Perception and Cognition: Issues in the Foundations of Psychology*. Minnesota Studies in the Philosophy of Science, vol. 9. Minneapolis, Minnesota: University of Minnesota Press.
- Schiffer, Steven. 1987. *Remnants of Meaning*. Cambridge, Massachusetts: MIT Press.
- Scriven, Michael. 1959. 'Explanation and Prediction in Evolutionary Theory.' *Science*, 130:477-482.
- Searle, John. 1980. 'Minds Brains and Programs.' *The Behavioral and Brain Sciences*, vol. III, no. 3.
- Skinner, B.F. 1953. *Science and Human Behavior*. New York: MacMillan.
- Stalnaker, Robert. 1984. *Inquiry*. Massachusetts: MIT Press.
- Stich, Steven. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Massachusetts: MIT Press.
- Stillings, Neil, et al. 1987. *Cognitive Science, An Introduction*. Cambridge, Massachusetts: MIT Press.
- Wilson, Mark. 1985. 'What is this Thing Called 'Pain'--The Philosophy of Science Behind the Contemporary Debate.' *Pacific Philosophical Quarterly*, 66:227-267.

