1-1-1996

Traditionalism and parallel distributed processing as qualitatively distinct models of the mind.

Mary M. Litch
University of Massachusetts Amherst

TRADITIONALISM AND PARALLEL DISTRIBUTED PROCESSING AS QUALITATIVELY DISTINCT MODELS OF THE MIND

A Dissertation Presented

by

MARY M. LITCH

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

DOCTOR OF PHILOSOPHY

February 1996

Philosophy

TRADITIONALISM AND PARALLEL DISTRIBUTED PROCESSING AS QUALITATIVELY DISTINCT MODELS OF THE MIND

A Dissertation Presented

by

MARY M. LITCH

Approved as to style and content by:

Lynne Rudder Baker

Lynne Rudder Baker, Chair

Bruce Aune

Bruce Aune, Member

Vere Chappell

Vere Chappell, Member

Andrew Barto

Andrew Barto, Member

John G. Robison

John Robison, Department Head
Department of Philosophy

ACKNOWLEDGEMENTS

There are several individuals and organizations whose assistance has helped make this work possible. First, part of the research for and most of the writing of this dissertation occurred while I was a visiting scholar at the Institute for Medical Psychology at Ludwig-Maximilians University in Munich, Germany under a grant (number A/94/08938) from the German Academic Exchange Program. I would like to express my appreciation to both organizations for their support.

I also owe a debt of gratitude to the members of my dissertation committee, Bruce Aune, Andrew Barto, and Vere Chappell, for their many suggestions on ways to improve this manuscript.

Finally, I would like to thank my chair, Lynne Rudder Baker, for her help and constant encouragement. I aspire to be as good an educator to my future students as she has been to me.

ABSTRACT

TRADITIONALISM AND PARALLEL DISTRIBUTED PROCESSING AS QUALITATIVELY DISTINCT MODELS OF THE MIND

FEBRUARY 1996

MARY M. LITCH, B.S., OLD DOMINION UNIVERSITY
M.S., UNIVERSITY OF MASSACHUSETTS AMHERST
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST
Directed by: Professor Lynne Rudder Baker

My main concern in this work is answering the question: does parallel distributed processing (PDP) as a model of the mind offer a genuine alternative to traditionalism? There has been vigorous debate within the last eight years on the subject of the relative merits of the one model over the other; however, a detailed examination of the nature of their respective differences has not been attempted.

The mental realm is that realm in which causal interaction is governed by laws quantifying over *representational* states. Traditionalism is the thesis that the law-governed transitions between mental states are transitions between computational states. PDP is the thesis that the transitions between mental states are transitions between distributed representational states in a PDP-type system. The representational content of a distributed state is determined by the causal history of the system as a whole, and results from the changing of system parameters via learning so as to insert this state in the causal chain between the perception of some external state-of-affairs and behavior.

Traditionalism and PDP are best considered not as providing a detailed picture of the causal processes involved in mental activity, but rather as providing a general framework that sets broad constraints on how such law-governed transitions proceed. I describe two aspects of qualitative distinctness that can be used even when comparing such non-specific models. The first involves examining the ontological commitment of each: assuming a realist interpretation, what must exist if traditionalism (or PDP) is a true model of the mind? If the two models make the same commitments, one may ask the further question: do the constraints imposed on the form that mental causal transitions take allow the possibility of an isomorphism between causal sequences permitted by the one model with those permitted by the other? An examination of the manner in which representational content is determined within PDP systems shows that there is no possible isomorphism. Therefore, the two models are qualitatively distinct.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENT.......... iv

ABSTRACT.......... v

LIST OF FIGURES.......... ix

Chapter

1. INTRODUCTION.......... 1

1.1 Overview of the Issue.......... 1
1.2 Relation of Issue to Other Areas of Philosophy.......... 14
1.3 Relation of Issue to Other Disciplines.......... 17
1.4 Personalities and Their Positions.......... 20
1.5 Outline of Rest of Dissertation.......... 21

2. WHAT IS MENTAL CAUSATION, ANYWAY?.......... 26

2.1 What is Causation in General?.......... 29
2.2 What is the Mental?.......... 50
2.3 Problems of Applying Notion of Causation to the Mental Realm.......... 58
2.4 Analysis of Mental Causation as Providing Model of the Mental.......... 74

3. TRADITIONALISM AS A MODEL OF THE MENTAL.......... 80

3.1 What is Traditionalism?.......... 81
3.2 Traditionalism as a Model of the Mental.......... 100

4. PDP AS A MODEL OF THE MENTAL.......... 118

4.1 History of PDP.......... 120
4.2 A Syntactic Description of PDP.......... 130
4.3 PDP as Currently Practiced.......... 157
4.4 PDP as a Model of the Mind.......... 195

5. ARE THE TRADITIONALIST AND PDP MODELS QUALITATIVELY DISTINCT?.. 210

5.1 What is Qualitative Distinctness?.. 212
5.2 Kuhn´s Theory and the Relationship between Traditionalism and PDP... 223
5.3 Some Answers Given by Others.. 236
5.4 Issues Concerning Computability and Computation (Plus, the LOT Argument).. 250
5.5 Are Traditionalism and PDP Qualitatively Distinct?..... 257

BIBLIOGRAPHY.. 265

LIST OF FIGURES

Figure Page

1. Portion of folk psychological web........................... 77

2. Relationship between physical and computational states... 89

3. Relationship of monolithic and sub-state types.. 108

4. The unit... 130

5. Types of unit output functions.................................. 132

6. 3-layered feed-forward network............................. 133

7. 3 unit network capable of learning........................ 139

8. Learning example in 3-unit network.................... 149

9. Network, post-learning... 150

10. PDP network in system capable of supporting representational states.................................... 173

11. Hypothetical pleasurableness values returned as a function of temperature........................... 179

CHAPTER 1
INTRODUCTION

1.1 Overview of the Issue

My main concern in this work is to argue that two widely held views of how the mind works present genuine alternatives to one another: the model of the mind associated with the one view is qualitatively distinct from the model associated with the other. The two models being compared are traditionalism and parallel distributed processing (PDP). "Traditionalism" is the name I have chosen to designate that model of the mind commonly held within post-behaviorist psychology and mainstream artificial intelligence. It is also known as "classicism", "symbolism", and "computationalism". Among philosophers, its most oft-cited proponent is Jerry Fodor. According to traditionalism, the causally efficacious mental states are structured, and the manipulation of these states is governed by formalizable rules. Parallel distributed processing (also known as "connectionism" and "the neural networks approach") is a newer model of the mind gaining in popularity within the cognitive science community. According to PDP, cognitive processes are implemented in networks of many interconnected, simple processing units.

The title of this dissertation, "Traditionalism and Parallel Distributed Processing as Qualitatively Distinct Models of the Mind", provides a key into the structure of the remaining four chapters.

Obviously, my argument that the two models are indeed qualitatively distinct is premised on a particular interpretation of the four key phrases found within the title. Each of Chapters 2-5 includes an explication of one of the four phrases. Thus, Chapter 2 poses and then answers the question: what is a model of the mind? Chapters 3 and 4 provide an analysis of traditionalism and PDP as models of the mind, respectively. Chapter 5 begins with a discussion of what it means for two models to be qualitatively distinct. Once the task of explication is complete, my actual argument for the qualitative distinctness of the two models follows rather easily.

In the rest of this section, I provide an overview of the line of argument found in the body of the dissertation. This overview should aid the reader in obtaining a feel for the philosophical landscape of the work, so that my motivation for including the particular topics and the particular arguments in subsequent pages becomes transparent. Stated in more poetic terms, the remainder of this section describes the layout of the forest and the examination of individual trees begins in Chapter 2.

The first topic to consider involves the meaning of "model of the mind" when applied to either traditionalism or PDP. A clear understanding of causation is integral to the interpretation of one system as a model of another, so I begin my examination of the concept of modelhood with an analysis of causation.[1] I assume a

[1] My greatest complaint against those who write in the field of cognitive science is that they so often employ words understandable only in the context of a particular interpretation of causation, yet fail to make their assumed view of causation explicit. A case in point is provided by Smolensky´s "On the Proper Treatment of Connectionism", which is the most widely read work by a PDP researcher within the philosophical literature on PDP. Although he repeatedly refers to representational states within PDP systems, and to a

modified Lewisian view of causation: the causal laws are determined by simplicity and strength criteria. These laws are analogous to axioms, which, in combination with the initial conditions, yield the set of facts. I must make several amendations to Lewis´ original proposal in order to produce a theory that allows for the existence of psychological laws.[2] I prefer Lewis´ approach to causation, because, with the accompanying possible worlds interpretation of counterfactuals, I am able to explain how representational content can be causally efficacious within the framework of a physicalist metaphysics. This is an important feature, given that this whole work presupposes that there are mental causal laws, and that these laws pick out the states based on their representational content.

I must here differentiate between two senses of "model", as that word appears within the phrase "model of the mind". According to one sense, a model is the supplier of a general, abstract framework: it is not a concrete instantiation, hence, it cannot be said

relationship between the dynamics of PDP systems and cognitive processes, he describes neither how representational states could possibly arise within such systems nor how the relationship between PDP system states and cognitive states is to be understood. As I shall argue in later chapters, the resolution of both of these questions involves an explication of causation and its role in determining meaning. The result of this failure to make the underlying theory of causation explicit is sometimes arguments with ambiguous premises, and a tendency for the potentially philosophically interesting exchanges between proponents of traditionalism and proponents of PDP to degrade into the two camps "talking past" one another, because they bring to the discussion differing understandings of what causation (and, in particular, causation as applied to the mental realm) is.

[2] One of several key assumptions to be found in this work is that there are mental causal laws. I nowhere attempt to justify this claim against either of the two groups opposed to the existence of mental causal laws. As a result, convinced eliminativists and non-eliminativist philosophers and psychologists who maintain that mental causal laws as such do not exist will right from the start want to reject my analysis of mental modelhood.

to itself instantiate causal laws. Rather, a model in this sense supplies a set of guidelines according to which a concrete instantiation of the model would be constructed. “Model” in the other sense refers to the concrete instantiation itself. Thus, when one refers to a model as a concrete physical object, subject to causal laws, one is using “model” in this second sense. Strictly speaking, traditionalism and PDP each are the supplier of a model of the mind in the first sense. However, I sometimes refer to, for example, a PDP system (i.e., a physical implementation of a network along PDP lines) as a model of the mind. This ambiguity in the word “model” allows me to avoid use of the more cumbersome, but technically more correct descriptor “concrete implementation of a system subject to the constraints supplied by the PDP model of the mind”, when refering to such a PDP system. I hope that in each case the context in which “model of the mind” appears makes its clear which sense of “model” is meant.

When we use one system to model another, we imply that the modelling system reproduces certain relevant features of the modelled system. Clearly, relevancy is relative to our purposes. If our purpose in using the model is to *explain* the behavior of the modelled system, one relevant feature of the modelled system is the set of causal laws operative in producing the behavior (or, more precisely, the set of causal laws operative in producing the subset of behavior of the modelled system of interest to us). Thus, an *explanatory* model's behavior will not only mirror the behavior of the modelled system, but will do so by virtue of instantiating the same causal laws responsible for the behavior in the modelled system. If our purpose in using one system as a model of another is merely prediction, all that is required is that the modelling system reproduce the sequence of states of the modelled system, without regard to whether or not that sequence is reproduced by instantiating the same causal processes. When proponents of traditionalism and PDP put their respective models forward, they do so with the understanding that the models can be used not merely to predict, but also to explain mental phenomena -- that the mental causal laws operative in (biologically-based) entities are likewise operative in bringing about the state transitions associated with their respective models. (A related thesis endorsed by both camps is that mentation is nothing above and beyond the instantiation of mental causal laws. In particular, issues relating to consciousness and its role in mentation are not a concern.)

One sees then, why, within the domain of cognitive science, it is so important to make explicit the interpretation of causation being presupposed. In order for a system to be an explanatory model of the mind, it must at a minimum be the sort of system that can possibly instantiate mental causal laws. Since these laws quantify over representational states, the system must likewise be capable of supporting representational states. To say (for example) that traditionalism is a model of the mind is then to say that a system structured according to the traditionalist guidelines can support tokens of the very same state types mentioned by the mental causal laws, and that the transitions between these states are governed by the mental causal laws.

What, then, distinguishes a *traditionalist* model: what constraints does traditionalism place on the hypothesized structure of the mind? (This is the second of the four key phrases in the dissertation title.) According to this view, the mind is a computer, not in the concrete sense in which "computer" is most often used (namely, as a particular piece of hardware), but in the abstract sense in which the term appears in the theory of computation. A computer is defined as something that engages in computation, which is in turn defined as something with readily identifiable states whose transition function is described by a formal, explicit algorithm. How these states (usually called "computational states") are realized is irrelevant: they could equally well be "realized" in the total states of an abstract Turing machine (the system state + tape contents + position of the read/write head) or in the states of a garden-variety silicon-based computer or in the states of a human nervous system.[3]

The theory of computation has advanced to the point where the border of the class of computable functions is well-known. Traditionalism assumes not only that mental state transitions constitute a computable function, but also that mental processing is computational processing, whereby the formalizable rules that force the mental state transitions are the mental causal laws. This hypothesis places clear constraints on the nature of mental processing. Two that bear directly to my main argument in this work are that: (1) any possible transition sequence of mental states

[3] This of course glosses over the fact that all real computers are resource-bounded, whereas the standard Turing machine is not.

consistent with traditionalism must be computable, and (2) the representational content of a particular mental state is inherited from the representational content of the computational state of which it is an instance.

A view that goes hand-in-hand with traditionalism is the language of thought (LOT) hypothesis. Within the theory of computation, it is the total computational state that, strictly speaking, must be rule-governed. According to the LOT hypothesis, those monolithic computational states that are also mental states have structure to which the mental causal laws are sensitive: in particular, they have a combinatorial semantic structure that mirrors a combinatorial syntactic structure. This structure explains certain attributes of mental phenomena (e.g., its systematicity). An implicit assumption of the argument for the LOT is that the semantic structural parts of these mental states are the same as (or, at least, very similar to) the semantic structural parts of our natural language. Thus, a further feature of traditionalism (via its association with the LOT argument) is that the level of reality represented by the mental states is that of word-concepts and propositions (i.e., entities easily representable in natural language).

A final point to note on traditionalism is its (sometimes love/hate) relationship with folk psychology (i.e., the "folk" theory explaining the behavior of mind-possessing beings by reference to beliefs, desires, etc. held by the entity and a set of generalizations linking the having of certain beliefs and desires with certain types of behavior). I explicitly distinguish the two: traditionalism is coherent on the view that mental states are something other than beliefs, desires, and all the rest. However, what traditionalism would be in that case is unclear. As a result, within this text, whenever I want to illustrate a traditionalist principle with a concrete example, I pull one from the domain of folk psychology. This does not, however, demonstrate an equation of the two. Indeed, folk psychology is best viewed as a specification of traditionalism: traditionalism provides the broad framework and folk psychology provides the particular details (i.e., the set of efficacious mental states and mental causal laws).

Parallel distributed processing, on the other hand, is not so widely familiar (and, may I dare to say, not so intuitive) as a model of the mind. It has gained in popularity over the last decade or so to the point where many within the cognitive science community view it as a rival to the continued hegemony of traditionalism. Its earliest roots are in neuroscience -- initially PDP networks were constructed as models of neural processing within the brain. Because of its relative youth, there is not even a consensus within the field of PDP researchers about what their current systems are modelling. In this work, I explicitly assume that PDP is interpreted as a model of the mind. This means that I identify states within PDP networks that are capable of supporting representational content, and construe the PDP model of the mind as the view that the way that these meaningful states follow upon one another is the very same way in which the mentally causally efficacious states in biologically-based agents follow upon one another. One consequence of my siding with the mind-modelling (as opposed to the brain-modelling) contingent among PDP researchers is that I downplay the physiological plausibility angle that one often meets in the PDP literature.

The most striking feature of PDP networks is that they consist of many simple processing units that pass signals to one another. The connections along which these signals are passed have a (perhaps alterable) number encoding the strength of connection between the unit sending the signal and the unit receiving it. This number is called the "weight". Each unit instantiates a simple function of the sum of the *weight x output value* (one product per unit to which this unit is connected); this output will in turn serve as input for the other units with which this unit connects. The pattern of connections for a network is one of the features determining its architecture. Some networks have bidirectional connections (i.e., if unit-a is connected to unit-b, then unit-b is connected to unit-a). Other networks are segmentable into layers, such that information passes (via the connections) in only one direction. (This net-type is called "feed forward".) Still other networks are predominately feed forward, but allow some connections to go "in the other direction".

The network architecture determines the class of functions that a PDP system can instantiate, and, hence, the types of tasks that it can perform; some network architectures are extremely limited in their task-solving capabilities, whereas others are powerful enough to instantiate any Turing-computable function. It is important for the reader to keep in mind that instantiating a function is *not* the same as computing it. In particular, PDP systems do not engage in computation. (Many contributors to the PDP literature, both PDP researchers and philosophers, fail to note this point. However, as even a cursory examination of PDP system dynamics shows, PDP networks fail to satisfy the conditions for a computational process, as understood within the theory of computation.)

This said, how then does a PDP state possess representational content, if it is in theory disbarred from inheriting it from the corresponding computational state? This question leads naturally into a discussion of how representation is in general explained. I adopt Dretske´s approach to the naturalization of content: a state (in this case, a state in a PDP system) comes to mean x when the state comes to play a causal role mediating the presence-of-x and behavior appropriate to the presence-of-x (for example, avoidance behavior when x´s are dangerous to the continued survival of the system as a whole). Dretske´s theory constitutes, I think, the best hope for the naturalization of content, and it applies just as well to artificial mental agents as to biological ones.

His theory does, however, make *learning* a necessary feature for any such agent, as the causal role that a particular state takes on is a result of learning. I mentioned above that, within PDP systems, each unit´s connections has associated with it a weight. Learning within PDP is accomplished by the changes of these weights over time, as the system adapts itself to its environment: as learning progresses, the system becomes more and more likely to produce the "correct" behavior, given its immediate environmental conditions. Various methods for achieving this directed changing of weights have been developed for use with PDP networks; the most popular is called "back-propagation" (or, usually, just "back-prop"). Using this technique, networks can be effectively (albeit slowly) trained to correlate certain inputs (e.g., detection of the presence of an x) with certain outputs (e.g., guidance of behavior appropriate to the presence of an x). In particular, the states mediating this correlation meet all of Dretske´s criteria for the attainment of intentional state status.

There is a good deal of debate, even within the mind-modelling contingent amongst PDP researchers, as to which states within PDP systems are the bearers of content. This issue is all the more contentious because there are two dimensions to consider: (1) Is it the unit level or the patterns over units level that provides the correct level of analysis of a PDP system as a model of the mind? (Are the intentional states, the contents of which are quantified over in mental causal laws, to be found at the unit or at the pattern level?) (2) Is it the unit output state that is the sought after representational state, or the weight state, or perhaps both together? In my analysis of PDP as a model of the mind, I identify the output[4] plus weight state over patterns of units as the states that, by virtue of their content, participate in mental causal laws. This content, like that associated with the causally efficacious states within traditionalism, is at the level of word-concepts and propositions (i.e., again, as with traditionalism, those objects and statements that are easily representable in natural language).

PDP is useful as a model of the mind because the transitions between these states can be studied (both within the framework of particular experiments and theoretically) in isolation from much of

[4] Technically speaking, I adopt the activation value as one component of the meaning state, not the output. This difference, though, is not of any theoretical importance.

the psychologically irrelevant details that co-occur in humans. The mathematical basis of the syntax of PDP systems is fairly well understood, and can be tapped to provide information on the constraints governing transitions amongst these states. PDP as model of the mind holds that these constraints are operative in any mind-possessing being, and result from the causal governedness of mental state transitions.

The final of the four key phrases within the statement "traditionalism and PDP are qualitatively distinct models of the mind" relates to qualitative distinctness. What is it? What criteria must be satisfied when two models are qualitatively distinct? My choice of words indicate that what I am after is a general framework for deciding whether two scientific theories (irrespective of their domain) are the same theory with a difference in terminology, distinct theories differing only in quantitative respects, or really two theories with differences that allow neither an easy intertranslation nor an easy shifting from one to the other by a change in the value of some constant appearing in both theories. My initial reaction to this need was to use Kuhn´s "incommensurability" to try to accomplish this task, but I quickly abandoned that concept as inappropriate. I therefore developed the notion of qualitative distinctness, which is, perhaps, best explained by giving the algorithm that tests for it. The algorithm consists of two stages. In the first, one asks whether the two theories differ with respect to their ontological commitments. This is accomplished by giving a realist interpretation to each of the two theories, and asking, for each one: what must exist if this theory is true? Among the ontological commitments of a theory are its very broad metaphysical assumptions, plus a commitment to the existence of the objects and their states quantified over in the causal laws forming that theory. If the two theories differ with respect to their ontological commitments, then they are qualitatively distinct. If not, one continues to the second stage of the test for qualitative distinctness.

This second stage involves sameness of posited causal processes. Given that both theories have the same ontological commitments with respect to the causally efficacious entities, one first matches up the corresponding entities across the two theories. The two theories are qualitatively distinct when the causal relation within the one theory is not isomorphic to the causal relation within the other, using the correspondence mentioned above as mapping. When such an isomorphism exists, the two theories are qualitatively *in*distinct.[5]

Once this (extended) bit of stage-setting is complete in Chapter 5, I can give my answer: traditionalism and PDP are qualitatively distinct models of the mind. Although the two theories make the same ontological commitments, the constraints on the possible sets of causal laws imposed by the two implies that there will be no possible isomorphism.

[5] Actually, this is too simple, for this way of putting it distinguishes theories differing only in quantitative ways.

1.2 Relation of Issue to Other Areas of Philosophy

Even the above sketch is sufficient to show that this dissertation is not easily pigeon-holed into one of the traditional areas of philosophy. While my top-level concern is in the philosophy of mind, I also deal with issues more properly part of the philosophy of science.

The most obvious classic philosophical issue that I address is the nature of the mental. For both traditionalism and PDP, to have a mind is to engage in mentation; to engage in mentation is to possess representational states, the transitions between which are governed by the mental causal laws; the mental causal laws are those laws of nature that advert to content. The main difference between traditionalism and PDP relates to the transition function between mental states. It is just as important to note what is omitted in traditionalism´s and PDP´s account of the mind. We see no mention whatsoever of an aspect of human mentation that some philosophers take as a defining characteristic: namely, consciousness.

In both traditionalism and PDP, the overarching goal (so overarching, that most researchers within both camps are probably unaware of it) is to explain mentation, and, in particular, to make room for the causal efficacy of mental states within a broad physicalist framework. In this regard, they find company with those philosophers who reject both eliminativism and dualistic-based attempts to argue for the reality of mind. Even traditionalism, with its Cartesianesque rationalist assumptions with respect to the innateness of the mental conceptual framework, thoroughly rejects Cartesian metaphysics.

In addition to the above issues, this dissertation also deals in depth with some topics most often associated with the philosophy of science. The most prominent of these surrounds the nature of causation: what is it in general, and how is the causal efficacy of mental states in particular to be understood? As already mentioned, I adopt a Lewisian construal of causation, but modify it so as to make it more realist in general, and more amenable to the existence of causal processes at levels other than that of basic physics. The appropriateness of the attribution of realism to this modified view is achieved by (1) assuming that a rationalist explication of the (true) simplicity and strength criteria can (at least, in theory) be given, and (2) giving a realist interpretation to the closeness ordering on the possible worlds. My motivation for bringing in possible worlds (which are, I think, best to be avoided if at all possible, given their metaphysical suspectness) is to help explain how content can be causally relevant. I emphasize repeatedly (especially in Chapter 2) that philosophers have been too quick to discard content. Thus we see Stich´s "syntactic theory of the mind" and Fodor´s "methodological solipsism" as rather misguided attempts to justify the continued use of terminology referring to the mind and to mental states. Their respective attempts are counterproductive, in my view, for the theory of the "mind" that remains after meaning as a causally relevant property has been removed is no theory of the mind at all.[6] I trace this misguided rejection of content back to a misconstrual of the counterfactual testing for the causal relevance of content. The need for a worked-out interpretation schema for the analysis of counterfactual statements thus drives me to Lewis.

A second topic that I take up in this dissertation that is also labelled as a part of the philosophy of science involves the autonomy of scientific disciplines. If both models of the mind are through and through physicalist, then there must be either reductive or at least supervenient relations linking mental states and physical states. Doesn´t the presumed existence of these relations make all non-physical states (in particular, mental states) causally inert? The generally assumed ceteris paribus nature of all causal laws outside of basic physics lends additional weight to the argument that causal processes, properly understood, occur only at the level of basic physics. In the course of arguing against this limitation of the scope of "causation", I consider evidence both pro and con relevant to the topic.

A third area of concern (also a quintessential part of the philosophy of science) within this work is the relationship between two models attempting to explain the same level of reality. If traditionalism and PDP both are models of the mind, must they necessarily be understood as competing (in the sense that consistency requires that the acceptance of one implies the rejection of the other)? If it is possible for them to be non-competing, what would that mean for their relationship to one another and for the

[6] Fodor is less than consistent in his rejection of meaning as causally relevant; thus, I isolate the "solipsistic" tendency within his writings as the less representative of his view as reconstructed by me.

nature of the mind? While I examine this issue using traditionalism and PDP as examples, the same concerns and questions apply to other scientific domains.

There are a few other subtopics within the philosophy of science that I touch on (e.g., the sociology of scientific practice a la Kuhn), but they are best viewed as side-issues, not directly relevant to the line of argument that I develop in the following four chapters.

1.3 Relation of Issue to Other Disciplines

The issues being considered in this dissertation span not only multiple areas within philosophy, but also multiple disciplines. In particular, the other disciplines with interests in cognitive science (i.e., psychology, artificial intelligence, and neuroscience) are the suppliers of many of the concepts and theories that appear throughout the rest of the work.

Artificial intelligence[7] is the provider of the two models being compared. Normal scientific practice within AI is not concerned with the implications, whether conceptual or psychological, of its research. Rather, the usual methodology is to isolate some interesting task and to then try to build a particular system that can solve it. While AI is thoroughly empirical, the issue of whether the constructed system solves the task in the same way as a human would is irrelevant. Perhaps a human task-solver can serve as a

[7] I include under this rubric all attempts at producing intelligence via non-natural systems. Thus, PDP is just as much a part of AI as traditionalism. When I mean to refer only to the traditionalist wing of AI, I use the phrase "mainstream AI".

source of ideas for strategies to use in the construction of the artifical system, but the question of whether the machine is doing the same thing as the human is a non-issue. To this extent, AI has remained faithful to Turing´s original advice vis-a-vis testing for intelligence: roughly, if a system´s behavior leads you to think it is intelligent, then it is.[8] The traditionalist and PDP models, as tools of AI, help to constrain the search for a system that can solve the task. So, for example, a PDP researcher identifies a target task, and sets about answering the question: can an artificial system with a PDP architecture solve this task?

It is only in the hands of the cognitive psychologist that the two AI frameworks take on the role of genuine mental models. A reconstruction of the process by which psychologists have come to accept either of the two models might go something like this. AI has produced artificial systems that can solve some cognitively interesting tasks. Perhaps the abstract architecture implemented within (either traditionalist or PDP) AI systems is the same as that implemented in the mind. Let´s work on that assumption and see if the data from psychological experiments fits the model. The theoretical advantage of the AI models over some other potential candidate model is that the former are consonant with physicalism, and the vast majority of psychologists assume a physicalist metaphysics. As with my above portrayal of AI, this portrayal of cognitive psychology is also an oversimplification. Of course, some

[8] Clearly, this portrayal of "normal science" is a generalization of what goes on in the process of research within AI labs. Many AI researchers *are* interested in reproducing not only I/O behavior related to human task-solving, but also the intermediate steps involved. In doing so, however, they are entering the domain properly belonging to psychology.

cognitive psychologists have research interests only tangentially related to that above (e.g., those psychologists interested in the relationship between mind and brain are a case in point). And, of course, people outside of cognitive psychology are interested in empirical support for one or the other model. Indeed, Fodor´s language of thought argument is, I think, best viewed as belonging to the domain of psychology.[9]

It is clear that my concern here is philosophical rather than psychological, for I explicitly state my lack of interest in empirically-based arguments of any kind. (I expend effort in examining the LOT argument only to help elucidate the traditionalist position vis-a-vis the level of reality represented by mental states and to consider but then reject the transcendentalist interpretation of it.) I state here and will from time to time reiterate this lack of interest. One result is the neglect of the question: which of the two models is the best? While the traditionalism versus PDP debate revolves around this question, I disregard it as outside of the proper domain of philosophy.

The last discipline to consider within cognitive science is neuroscience; in particular, that area of neuroscience concerned with the relationship between neural and mental level phenomena. This dissertation contains very little of interest to the neuroscientist. In fact, I only touch on neuroscientific issues in providing a brief history of the development of PDP.

[9] I shall interpret the LOT argument, not as a transcendental argument, but as an inference to the best explanation.

1.4 Personalities and Their Positions

Many writers have expressed their view on the issue of the superiority of either traditionalism or PDP as a model of the mind. Within these writings, one can often tease out assumptions relating directly to the topic of this dissertation: namely, are the two models qualitatively distinct?

The philosopher most often cited in the literature on this topic is Jerry Fodor. He clearly enunciates his view that the two models *are* qualitatively distinct. His reasoning is that empirical evidence supports the interpretation of traditionalism as a model of the mind and PDP as a model of the implementation level of the mind. Because the two models are models of different things, they must be qualitatively distinct.[10]

Another author whose views are often cited is Paul Smolensky (an AI researcher and supporter of PDP as the correct model of the mind). He is likewise of the opinion that the two models are qualitatively distinct, although, as one might guess, his reasons differ from those of Fodor and Pylyshyn. For him, PDP at the unit level of description is the correct model of the mind, and traditionalism is an approximation to the gross characteristics of pattern level activity. Thus he, like Fodor and Pylyshyn, understands traditionalism and PDP as modelling two distinct levels of reality. He describes the relationship between traditionalism and PDP as analogous to that

[10] Because Fodor co-wrote with Zenon Pylyshyn (a computer scientist and mainstream AI researcher) the first work in which he explicitly mentions PDP, I use both names whenever I refer to the LOT argument as applied specifically to the issue of the adequacy of PDP as a model of the mind.

between Newtonian and quantum mechanics: the laws of quantum physics are the true, counterfactual supporting laws governing all transitions between physical states. The laws of Newtonian physics, while offering accurate predictions over a limited range of physical phenomena, are only an approximation of the underlying, genuine physical laws. Thus, in a sense, quantum physics implements Newtonian physics (at least to the extent that the laws of the latter can be derived from an averaging over a very large number of individual quantum mechanical processes). Just so, PDP provides the true mental causal laws, and traditionalism approximates the mental microprocesses by averaging over a large number of these causally determined microprocesses. On this view of their relationship, traditionalism and PDP are qualitatively distinct.

Some philosophers who have of late contributed to the traditionalism versus PDP debate on the side of the latter fall outside the scope of this work, for their main motivation in supporting PDP involves interpretting it as eliminativist. An example in this group is Patricia Churchland.[11] Stich can perhaps also be put into this group, although he denies being an eliminativist.[12]

1.5 Outline of Rest of Dissertation

Before embarking on the body of this work, I would like to give the reader a general idea of where various topics are taken up, and of which views and arguments are original, and which are re-

[11] See her *Neurophilosophy*.

[12] See Ramsey, Stich, and Garon´s "Connectionism, Eliminativism, and the Future of Folk Psychology".

hashes of the ideas of others. Each of the remaining four chapters includes a detailed analysis of one of the four key phrases in the thesis title.

In Chapter 2, I provide the description of mental causation that will be presupposed in the other chapters. In the first section, I describe how I will be understanding causation as a general relation. As already mentioned, I use Lewis´ theory as a starting point. However, given the large number of amendations that I make to it, it is not at all clear that it is correctly described as "Lewis´ view". Indeed, I would assume that, if asked, Lewis would openly reject it. To my knowledge, no one else has written on how a Lewis-style interpretation of causation would need to be changed to make it applicable to the special sciences in general, and to psychology in particular. In the second section, I try to delimit the mental realm from the rest of reality: what properties do mental phenomena possess that set them off as mental? While identifying mental phenomena as those governed by laws adverting to content is not new, I do produce several arguments working out some of the ramifications of this equation, both in general and as relevant to traditionalism and PDP as models of the mind. Most of the third section is taken up with an (original) argument that content *is* causally relevant, even supposing a physicalist metaphysics. The so-called problem of mental causation is one among the several classic problems in applying the notion of causation to the mental realm. In this third section, I also consider and suggest solutions for some of the others. The fourth section gives my view of the relationship between a model and the scientific domain being modelled. Along the way, I make a distinction between two types of models (explanatory versus merely predicting) that seems to me to be important to understanding what role traditionalism and PDP are playing within psychology. As far as I know, this distinction and the working out of its implications are original. In the fourth section, I try to tie together the various views put forward in the first three sections to produce a coherent picture of the relationship between a model of the mind and mental causation.

Chapter 3 deals with an explication of the traditionalist model. In the first section, I describe traditionalism, both its implicit and its explicit assumptions. I have tried to make this section as unoriginal as possible, lest I be accused of presenting a false picture of traditionalism. In the second section, I work through the implications of what I have written in the first section in the light of my view of the relationship between a model and its domain. Along the way, I isolate the ontological commitments of traditionalism and the constraints that it places on the form of the mental causal laws. Another subject taken up in Section 2 is an analysis of computational statehood as that concept is used within traditionalism. Again, to my knowedge, this is original.

Because PDP is perhaps new to some readers, I give a slightly different treatment to the topic of PDP as a model of the mind than the one I used in Chapter 3. I start off the fourth chapter with a brief history of PDP, and provide some sample quotations from the literature showing the diversity that fits under the PDP banner. This section is mostly summary and direct quotation. In the second section, I give a syntactic description of PDP systems, again, on the assumption that this is all new to the reader. I describe the building blocks of PDP networks, and try to give a feel to the reader for the dynamics of such systems. Sections 3 and 4 cover the same ground for PDP as Chapter 3 covered for traditionalism. I distinguish between two differing views of the model of the mind being offered by PDP (namely, the local and distributed interpretation schema), and argue for the superiority of the latter as offering the most coherent model of the mind. Along the way, I need to explain how PDP states come to have content. The first stage of the naturalization of content for PDP systems is the enunciation of a theory of representation. I adopt Dretske´s, wholecloth. The remainder of this project is wholly original. I give general principles for explaining how the PDP states come to have content and illustrate it with an example of a particular state´s coming to have a particular content. Once all of the pieces are in place for interpretting PDP as a model of the mind, I identify the ontological commitments made and the constraints on the causal laws offered by it.

The final chapter begins by providing an explication of the fourth key phrase: qualitative distinctness. The concept as such is new, but the parts out of which it is constructed are borrowed. The idea of comparing ontological commitments comes from Kuhn, and the idea of checking for an isomorphism between items originates with Putnam (although, his functional isomorphism needed some re-working to make it fit an inter-model comparison). I try to illustrate and make clear what I mean by qualitative distinctness with several examples. The second section is not directly relevant to the main line of argument in this work, but it was fun to think about, so I included it anyway. In this section, I examine Kuhn´s theory of scientific evolution (with particular emphasis on the role played by incommensurability within that theory) and apply it to the current state of cognitive science with respect to the traditionalism versus PDP debate. I include some criticisms of Kuhn´s theory, and distinguish his incommensurability from my qualitative distinctness. The title of the third section ("Some answers given by others") might lead one to mistakenly believe that the section is nothing but summarization. However, many of the writings in this area are so ambiguous and require so much "reading between the lines" that the arguments found in this section are more original than not. My way of approaching that topic is to isolate a clearly-stated view from one or another person who wrote on this topic, and, using that as a premise, try to construct an argument either that the two models are or are not qualitatively distinct. In the fourth section, I consider and then discard several arguments within the traditionalism versus PDP debate relating to computability. I then consider and reject the interpretation of Fodor´s LOT argument as supplying necessary conditions for something´s being a mind. The argument is old, this form of its rejection is new. The final section gives my preferred answer to the question: are the two models qualitatively distinct? My aim in the first 200-odd pages of this dissertation is to set the stage, so that my conclusion "yes, they are" would follow in a straightforward manner. I hope that I have succeeded.

CHAPTER 2
WHAT IS MENTAL CAUSATION, ANYWAY?

The logical place to begin a work dealing with a comparison of cognitive models is with a description of what it means for something to be a model. After all, not any old system counts as a model of a domain, so the question arises: what features of a system make it a possible candidate as a model of another system? I maintain that, at a minimum, there must be a correspondence between the constituent parts of the modelled system (at a suitable level of description) and the modelling system. For example, were I to model our solar system (say, with a desktop reproduction of it), the relevant parts that I would need to include in order for the desktop system to be a genuine model of the solar system would be the sun and planets. In order to be a genuine model, the desktop version need not reproduce every detail of the actual solar system. A second necessary condition for modelhood is that the important interrelationships between the parts of the modelled system are reproduced in the modelling system. What is *important* is relative to the use to be made of the model. When it is to function in a non-explanatory mode -- merely keeping track of places -- the causal relationship amongst the parts of the modelled system need not be reproduced in the modelling system. When, however, the model is intended as providing an *explanation* of the modelled system, it must reproduce the relevant causal relations.

A simple thought experiment should convince the reader of this assertion:

> Imagine I set a pendulum in motion and a child approaches me and asks: "Explain to me the motion of the pendulum, especially its going from the one extreme to the other." Suppose further that my response is as follows: "You see, there exists (abstractly) this Turing Machine with the two states, S1 and S2, such that S1 corresponds to the pendulum´s being at its left-most extreme position, and S2 corresponds to the pendulum´s right-most extreme position. The look-up table of the Turing Machine consists of two items:
>
> (<S1 ! * (don´t care)><S2 ! 0 ! no move>)
>
> (<S2 ! * (don´t care)><S1 ! 0 ! no move>)[1]
>
> So you see that the Turing Machine goes from S1 to S2 and back again forever, and that *explains* the motion of the pendulum."

My intuitions tell me that the above usage is improper: merely displaying a correspondence between the states of an entity and the states of some abstract machine (being put forward as a potential model) does not *explain* the former. A more appropriate word to use in this context is "describe": the abstract machine *describes* the behavior. Explanation requires something more than just regular correspondence -- in particular, it requires subsumption under causal laws. Were I to retell the above thought experiment, replacing my "explanation" with a description of the pendulum´s behavior as resulting from the effect of gravity and tension in the

[1] The formalism I adopt for the items in the look up table of the Turing Machine is :
(<machine state at time T!char under read/write head at T>
<machine state at T+1!write at current head position!move head>)

pendulum´s string, my intuitions would immediately change, and I would submit that, in that case, "explain" was properly used. It is for this reason that I must begin this work with an explication of causation -- both in general and in the mental domain. Returning to our desktop solar system as place-holding example, the relevant relations would include relative distances among the sun and planets and relative diameters. If the desktop system is to model not the solar system at a particular time, but rather the solar system as a dynamic system, then an additional interrelationship that the desktop system must capture is the relative rotational velocities of the planets. If, on the other hand, we wish to use the desktop solar system as an *explanatory* model of the real solar system, we will somehow have to capture the causal relations which underlie the behavior of the real solar system in our desktop model. Because of the difficulty of overcoming the interfering causal relations to which the desktop (but not the real) solar system is subjected, this is nearly impossible on earth. Hence, such a desktop model could not serve as an explanatory model of the real solar system. While the solar system example is rather simple, it serves to illustrate the sorts of considerations that go into construing one system as a model of another.

Returning to the task at hand, what are the parts and their relevant interrelationships in a cognitive system? Unlike the solar system case, the "parts" are not physical parts (this must be the case, even if one believes that all mental states are reducible to physical states, for, as a description of a *cognitive* system, it is states *as mental* that are relevant, irrespective of how those states are realized). Rather, the parts of a cognitive system include intentional states.[2] A cognitive system may also include non-intentional states (for example, pain states and other qualia) as parts. The relevant interrelationships among these states that any adequate model must capture are the causal interrelationships. It is for this reason that I begin an explication of modelhood with a discussion of causation.

2.1 What is Causation in General?

It is seldom that a philosopher of mind begins a paper with an explicit account of the theory of causation being assumed whenever she uses terminology adverting to causal interaction. Rather, the general rule is to use causal terminology without making it clear what is meant thereby. This can result in misunderstandings when the reader assumes one theory while the author assumes another. Particularly prevalent are situations in which one party assumes a realist understanding and the other party an irrealist understanding of causal terminology. To avoid such possible misunderstandings, I will lay out in advance the underlying theory being assumed whenever I use the word "cause" and its cognates, laying emphasis on where I am making ontological commitments. In addition, this

[2] It has come to be the standard practice in the philosophy of mind to mention Brentano´s assertion that the hallmark of mental systems -- what sets them apart from all other types of systems -- is that they consist of states which are essentially representational. I shall follow standard practice in this regard. Some (notably eliminativists) may maintain that there are no mental states, and that the totality of cognitive capabilites are explainable without recourse to mental/intentional type talk; however, as stated in Chapter 1, I am assuming the contrary. Cognition has an essential mental component, and what distinguishes mental states is their intentionality.

section will aid in fleshing out the sorts of objects that are relevant to a model.

Stated broadly, the goal of a scientific discipline, whether it be astronomy, biology, or psychology, is to identify the rules which capture the regularities of the state transitions of the objects of concern to the discipline (that is, rules with the form S1-->S2, where S1 and S2 are state descriptions of an object or of objects (possibly complex), and "-->" is not logical, but rather nomological implication). For astronomy, the relevant objects are celestial bodies, and the regularities to be captured involve repeated patterns of motion of those bodies. The rules are, however, assumed to be more than mere generalizations, true of the observed state transitions, but possibly not true of unobserved or yet-to-be-observed state transitions. Rather, they are assumed to codify an underlying natural (or, in accordance with philosophical terminology, *nomological*) relationship, such that not only the observed state transitions and the yet-to-be-observed transitions, but also the counterfactually observed transitions, proceed in accordance with the rules. That is, the rules justify sentences of the form: "if it were to be the case that S1, then it would (shortly thereafter) be the case that S2". So, at a minimum, a theory of causation must support counterfactual claims regarding state transitions. A particular state transition is causally-produced when it is an instance of one of these rules. (My treatment of causation specifically discounts event causation -- ie, the analysis of causation at the level of particular events, without the requirement of subsumption under a causal law.) So, s1 is the cause of s2 if and only if s1 is an instance of S1, and s2 is an instance of S2, and it is a causal law that S1-->S2 (this can be read in English as "S1´s are nomologically sufficient for S2´s"), and s1 is followed by s2.

Two potential difficulties arise regarding the interpretation of this. The first involves statistical laws: what if S2´s follow upon S1´s with a probability less than 1.0 (i.e., not every S1 is followed by an S2, but of those that are, the S2 is caused by the S1)? We see such statistical laws in quantum physics. The interpretation of nomological sufficiency for the case of statistical laws is slightly different from the case of non-statistical laws with respect to prediction: we cannot say that if an instance of S1 occurs, an instance of S2 will immediately follow. However, the role that statistical laws play with respect to *explanation* remains the same as with non-statistical laws: after the fact, when an S2 immediately follows an S1, s1 was the cause of s2. We can easily cover this case by extending the meaning of S1-->S2 to include statistical laws: so now, S1-->S2 means "S1´s are nomologically sufficient for S2´s with probability p". While it is an empirical question whether psychological laws are statistical in nature, I will (admittedly, without any attempt at justification) assume that they are not. As my ultimate concern is an explication of *mental* causation, I shall henceforth treat causation and causal laws as non-statistical: to say that S1´s cause S2´s is to say that, given an S1, an S2 will (with probability=1.0) follow, subject to ceteris paribus constraints, as per below.

The second potential difficulty regarding my theory of causation involves the status of non-strict (so called *ceteris paribus*) laws. It is (universally?) accepted that all non-basic laws have a suppressed ceteris paribus proviso: only for the most basic physical laws do S2´s always follow upon S1´s, irrespective of all other facts about the world. Does this mean then that all non-basic laws are not precisely-speaking genuine laws? The consequences of accepting such a view are far-reaching. For the physicalist, this amounts to the capitulation that all "laws" other than those dealing with state transitions in basic physics are not genuine laws, but merely approximations. In particular, psychological laws (and, hence, mental causation) are impossibilities.

One way out of this dilemma is to reject physicalism: there are basic laws quantifying over something other than physical states. In particular, for the non-physicalist who is keen on maintaining the existence of psychological laws under this precise construal of "lawhood", there are basic (hence, potentially strict) laws quantifying over intentional states. This is not an option that I can take, for in describing traditionalism and PDP as cognitive models, I must remain faithful to traditionalism and PDP as they are understood by their proponents: in each case, it is assumed that the causally-interacting objects are ultimately instantiated in physical matter.

This leaves only two options: either accept the claim that there are no causal laws covering state transitions of objects that are not the entities of basic physics, or relax the criteria for causal lawhood to include non-strict (i.e., ceteris paribus) laws. As Jerry Fodor has persuasively argued in Chapter 5 of *A Theory of Content*, choosing the first option means not only that all purported psychological laws are not entitled to the claim to lawhood, but also that *all* purported laws in all disciplines other than basic physics are non-laws. If a discipline is a science by virtue of locating the causal laws relating state transitions of its relevant objects, then all so-called special sciences are not genuine sciences. I, along with Fodor, take this as a reductio of the appropriateness of the equation of "causal laws" with "strict laws". Any adequate theory of causation must allow for non-strict laws.

Here I would like to distinguish two sources of non-strictness in causal laws, roughly characterizable as countervailing tendencies and unsatisfied implementation-level assumptions. By "countervailing tendencies" I have in mind the existence and instantiation of other causal processes that tend to produce the opposite effect as that produced by the causal law in question. An example from Newtonian physics illustrates this source of non-strictness. It is a law that rigid bodies move in the same direction of an applied force with an acceleration equal to the strength of the force divided by the mass of the body. Strictly-speaking, this law, even as applied to middle-sized rigid bodies, is not exceptionless, for there may be other forces exerted on the body that tend to move the body in the opposite direction. Newtonian physics has developed the handy notion of the vectorial summation of forces (which, in reality, is not a *particular* force being applied to the body) to explain the "exceptional cases" to this law. Thus, even though I may apply a 1 Newton force to a rigid ball, it is nomologically possible that the ball does not accelerate in the direction of the application of the force, because in addition to this 1 N force, there is a second 1 N force being applied along the same axis as the first, but in the opposite direction. Consider an example of countervailing tendencies closer to the domain of this work. Suppose that it is a psychological law that if I desire to eat x, and I believe that I have unrestricted disposal rights to x, then I will eat x. Suppose further that I have a desire to be well-respected, and a belief that I will be well-respected if I publicly give x to some needy person. Suppose further that it is also a law that if I have a desire to achieve condition y and see the means to achieve y as readily available to me, then I will (via those means) achieve y. Even though I may desire to eat x myself, the countervailing causal process explains my failing to eat x and my giving of x to the needy person. Just as in the case of countervailing (physical) forces, we see that the mere satisfaction of the nomologically sufficient conditions does not guarantee the obtaining of the effect.

The second source of non-strictness in causal laws (summarized above as the lack of satisfaction of implementation-level assumption) is what is usually intended to be captured under the rubric "ceteris paribus" (at least, as Fodor uses the term). The "ceteris paribus" is intended to capture the fact that (assuming physicalism) all objects are implemented in physical stuff, all causally-governed state transitions of the non-basic objects occur as a function of the causally-governed state transitions of the objects´ implementing constituents, and occasionally, the background assumptions of a causal law (statable only in the vocabulary of the science of the implementing constituents) are not satisfied.[3]

[3] One may argue that the need for ceteris paribus laws shows that the objects postulated by the special sciences do not "carve nature at the joints", but that the objects are merely "close approximations" to the true strict causally

Returning to the formulation of causal lawhood, "S1-->S2" means "S1´s are nomologically sufficient for S2´s, ceteris paribus". When a particular S1, s1, is followed by an S2, s2, s1 is the cause of s2, simpliciter: the non-strictness is only at the level of lawhood, not at the level of particular causal interactions. It is arguable that the statement that psychological laws are non-strict is incompatible with traditionalism (the model of the mind to be discussed in Chapter 3). However, such an argument, while interesting as a counter to the claim that Fodor´s various views relevant to the nature of the mind are consistent, would not be directly relevant to the topic of this work.

So far, I have not specified what makes a state transition description a causal law. One aspect involves the "naturalness" of the connection between the two states. I mentioned previously that in order for a generalization relating two states to be a causal law, it must hold not only for all actual S1´s, but also for all counterfactual S1´s. How is this condition to be understood?[4] Most importantly,

interacting aggregates, and, rather than capitulating to the tendency to accept the status quo in the special sciences, we should withhold the title of "science" from those disciplines which have not yet located the true causally interacting aggregates, on the assumption that, eventually, the special sciences will be able to locate the "natural" objects and the strict causal laws relating their states. While I find this line of argument somewhat compelling, I take the universality of non-strictness of laws in the special sciences to indicate that it is not that the scientists in those disciplines are sloppy, nor that the disciplines are qualitatively less well-developed than basic physics, but rather that the non-strictness (and, hence, the need to accept ceteris paribus laws as laws) is a fact of nature: there are no non-basic objects whose state transitions obey strict laws.

[4] In what follows, I shall adopt a highly modified version of David Lewis´ explication of causation and counterfactual support. As my main concern in this chapter is describing mental causation, I must relax some of his constraints on lawhood. This is because Lewis´ treatment of causal lawhood in *Counterfactuals* is most naturally viewed as applicable only to strict laws, which as argued above, exist only in basic physics. Similarly, his assumption of type-type bridge laws between basic physical states and higher-level states

causation is a natural, rather than conventional relationship: there is a fact-of-the-matter about whether a particular succession of states is causally related and there is a fact-of-the-matter about whether a generalization relating states is a causal law. Specifically, I reject all instrumentalistic construals of causation, whereby the ground for perceived regularities of succession is left as wholly mysterious. Instances of a cause (both actual and counterfactual) necessitate their effect because there is some property of the cause that forces the transition to the effect. For the case of special science laws, the forcing property(ies) are grounded in the causal laws relating the states of the implementing constituents. For the case of the laws in basic physics, the forcing property(ies) are not analyzable: the regress stops here at the level where the forcing of the effect given the cause is a brute fact of nature.[5] A description of the causal structure of the world can be given in terms of possible worlds. I adopt this approach because of the ease with which it can be used as a framework for analysing counterfactuals.[6]

is an issue on which I am trying in this work to remain agnostic. (His reductionistic predilections are clearly expressed in the chapters on the philosophy of mind in his *Philosophical Papers, Vol. 1.*) In making this change, I introduce difficulties not found in his original theory -- this is the price that must be paid in converting a clean but generally inapplicable theory of causation into one that can be used for disciplines other than basic physics. I also diverge from his theory in my analysis of the similarity relation, as described below.

[5] Admittedly, this leaves the ground for causation as mysterious as that resulting from instrumentalism: positing "brute facts of nature" is merely a philosophical device for stopping what would otherwise be an infinite regress. So be it.

[6] I am not clear that I want thereby to commit myself to the existence of these possible worlds (a la Lewis). Rather, all that I think is necessary is that I am committed to the non-conventionality of the closeness ordering of the possible worlds. In particular, there is a matter-of-fact about which possible worlds are close to the actual world, and which are not.

The set of causal laws forms an approximate hierarchy, with the laws of basic physics at the base. Assuming that the current "disciplinization" of science is complete and accurate,[7] each level of the approximate hierarchy corresponds to a scientific discipline. Within each level, the causally-interacting objects (i.e., the objects whose state transitions are quantified over in the causal laws), while implemented in lower level objects, really do exist.[8] However, the causally-interacting objects and the causal laws form a package deal, determinable only a posteriori. A distinct scientific level of analysis exists if both the state transitions of objects at that level are describable by causal laws and the objects and their states fit smoothly into the overall quasi-hierarchy of scientific disciplines. This "smoothness of fit" criterion is the analog of the strength and simplicity criteria for causal lawhood within a level, as described below. Within a level, the causal laws are determined using Lewis´ and Ramsey´s simplicity and strength criteria for lawhood in basic physics, whereby the laws are analogous to axioms, which, in combination with additional axioms describing the initial state of the world, yield the set of facts. Lewis describes it thus:

[7] This assumption is being made at this point only for expository purposes: so as to allow me to use simpler locutions such as "the laws of chemistry", rather than the strictly more correct locution "the laws of chemistry, under the assumption that chemistry is a proper discipline of completed science". Nothing at this point hinges on this assumption, and I shall in later chapters explicitly disassume it when discussing the status of psychology as a scientific discipline.

[8] I know this way of putting the point sounds silly. However, I must say it in order to make clear my view that, even on the strong claim that the causally-interacting objects are *reducible* to their lower level constituents, causally-interacting objects are never "mere artifacts" of a discipline, useful but ontologically-speaking fictitious. If the states of an object are quantified over in a causal law, then that object and these states exist; they are not epiphenomena of their respective constituents.

> Whatever we may or may not ever come to know, there exist (as abstract objects) innumerable true deductive systems: deductively closed, axiomatizable sets of true sentences. Of these deductive systems, some can be axiomatized more *simply* than others. Also, some of them have more *strength*, or *information content*, than others. The virtues of simplicity and strength tend to conflict. Simplicity without strength can be had from pure logic, strength without simplicity from (the deductive closure of) an almanac. ... What we value in a deductive system is a properly balanced combination of simplicity and strength -- as much of both as truth and our way of balancing permit. We can restate Ramsey´s 1928 theory of lawhood as follows: a contingent generalization is a *law of nature* if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength. A generalization is a law at world *i*, likewise, if and only if it appears as a theorem in each of the best deductive systems true at *i*.[9]

This theory of causal lawhood has been widely discussed, and many of its weaknesses have been pointed out. My purpose here is not to enumerate them, but to focus on two of them that are particularly relevant in light of the uses I want to make of the theory. The first of these involves the apparent irrealism presupposed by this theory: nowhere in this definition of causal lawhood is mentioned causation as a *natural* (as opposed to conventional) relation. It is possible that the axioms/laws that form the "best fit" to the data (in terms of simplicity and strength) fail to cut nature at the joints -- either because the "true" causal laws do not conform to our somewhat aesthetically-based criteria of simplicity and strength or because there are no natural joints, hence no "true" (in the realist´s sense) causal laws to be determined. My response to this charge of

[9] *Counterfactuals*, page 73.

irrealism against my adopted theory of causation is as follows: it is a brute assumption of mine that there (really) are causal laws -- I make absolutely no claim to be able to justify it. Given this assumption, the second of the two possibilities is not a problem. What of the first possibility (i.e., that the axioms/laws thus determined fail to cut nature at her (true) joints)? This objection, if left uncountered, would prove to be the undoing to my otherwise thoroughly-realist account of causation. Lewis notes and in some passages accepts this charge, as in the following, where he considers the possibility that there will be no "best set" of generalizations:

> We may hope, or take as an item of faith, that our world is one where certain true deductive systems come out as best, and certain generalizations come out as laws, by *any* remotely reasonable standards -- but we might be unlucky.[10]

In order to remove this potential source of irrealism, I must make another change to Lewis´ original formulation: let the laws be those generalizations that appear as axioms in the deductive systems of the actual *and all nearby possible worlds.* As stated above, my realism extends to the existence of a similarity relation between the actual and all possible worlds. Lewis waffles on this point. For example, there are realist-sounding passages such as:

> It is a fact about a town that it is situated near to one city rather than another, and in the same way it is a fact about our world that its character is such as to make some antecedent worlds [that is, those worlds in which the antecedent of a counterfactual conditional is true] be similar to it, and others not.[11]

[10] *Counterfactuals*, page 74.

[11] *Counterfactuals*, page 69.

However, the more often stated view is less realist with respect to the similarity relation (although he does presuppose that the similarity relation is something more than mere whim) as in:

> ... [T]he relative importance of respects of comparison, and thereby the comparative similarity of worlds, are at least roughly fixed. Not anything goes. It can happen that a counterfactual is true (at a world) according to some permissible systems of spheres but not according to others, so that its truth value will be indeterminate by reason of vagueness. But it can happen also, and often does, that a counterfactual has the same truth value according to all permissible systems of spheres, and so is definitely true or defintely false.[12]

As my aim here is not exegesis of *Counterfactuals*, but rather the description of the causal theory I shall be assuming, I feel free to reject the latter (admittedly, more representative) view in favor of the former, realist, view of the similarity relation. This said, there will be a matter-of-fact about which worlds are nearby to the actual world. This matter-of-factness grounds a matter-of-factness with respect to the truth value of counterfactuals (or, at least the truth value of counterfactuals whose antecedent does not stand in contradiction to the causal laws in the actual world). When I say that I am a realist with regard to causal laws, this is what I mean.

Let us assume that the similarity relation as regards a world and its nearest neighbors (using Lewis´ spheres terminology, the centered world, *i*, and the set of worlds constituting the sphere immediately surrounding *i*) is symmetrical and transitive. That is, if *i* is a nearest neighbor of *j*, then *j* is a nearest neighbor of *i*, and, if *i*

[12] *Counterfactuals*, page 93.

is a nearest neighbor of *j*, and *j* is a nearest neighbor of *k*, then *i* is a nearest neighbor of *k*.[13] Then, a world and its nearest neighbors will have the set of causal laws in common. Stepping back and viewing the ordering that the nearest neighbor relation imposes on the set of possible worlds, we see that they form equivalence classes, each class having the set of causal laws in common. The similarity relation for a world as relates to the possible worlds other than its nearest neighbors may be non-symmetrical and non-transitive.

A second criticism of the counterfactuals via possible worlds formalization of causation involves the possibility that there are multiple but mutually-orthogonal sets of best fit axioms, hence, given Lewis´ initial criterion[14], that there are no causal laws. This is particularly relevant to the topic of mental causation in two respects. First, ontological space must be made for mental states to be causally-efficacious, even presuming a physicalist metaphysics. If psychological generalizations are redundant in the sense of producing the same predictions at a macro-level as those made by the laws of basic physics at the micro-level, then the simplicity criterion will rule them out as possible laws of nature. Lewis views this as being not a flaw, but a feature, as he maintains that the only

[13] This is yet another divergence from Lewis, who explicitly denied the *general* symmetricity of the similarity relation, as noted on page 51 of *Counterfactuals* "This assumption of symmetry for the similarity measure implies a constraint on similarity ordering derived from that measure. ... But that constraint would be unjustified if we suppose that the facts about a world *i* help to determine which respects of similarity and dissimilarity are important in comparing other worlds in respect of similarity to the world *i*." However, the assumption of limited symmetry and transitivity (holding only between a world and its nearest neighbors) is much more well-founded, since, as Lewis himself notes, it is features of the centered world that influence what worlds are similar to it -- and those similar worlds should have similar features.

[14] That to be a law of nature is to be a member in *all* of the best sets.

laws of nature are to be found at the level of basic physics anyway. (This is, I think, his view in the most consistent reading of *Counterfactuals*. He does in other works explicitly mention causal laws other than those of basic physics, however.) I shall discuss the issue of the supposed causal inertness of mental objects in a later section. Suffice it for now to say that this result, if allowed to stand, makes the traditionalism versus PDP debate irrelevant, since neither are in that case potential candidates as mental models. Therefore, I need make yet another change to Lewis´ original theory. The change involves only one word, yet its acceptance transforms the theory from one wherein all the so-called special sciences are not genuine sciences to one where sciences other than basic physics can truly speak of causal laws in terms of their own special vocabularies. This change is as follows: replace Lewis´ definition of "law of nature" with "a contingent generalization is a *law of nature* if and only if it appears as a theorem (or axiom) in **ANY** of the true deductive systems that achieves a best combination of strength and simplicity."

Lewis´ reasoning behind not using this definition is that it leaves open the possibility that events can be causally over-determined. This is not the cornerstore variety of over-determination, the classic example of which is the shattering of a glass after being simultaneously subjected to being struck by a hammer and being exposed to the soprano´s high-C, both of which are alone nomologically-sufficient for the glass´ breaking. Rather, the over-determination is deeper, in that it involves something akin to diverging ways of conceptualizing the world as a causally-describable system. It was this deeper over-determination that troubled Lewis. For me, on the other hand, this form of over-determination is a necessary feature of a layered view of the sciences -- hence, it is to be sought out rather than avoided. This acceptance of over-determination is not inconsistent with a realist understanding of causation, for realism does not imply that only one causal process can be operative in a dynamic system. Consider a particular chemical reaction. The events taking place (broadly construed) can be causally explained at either of two levels of description: the chemical and the atomic (physical). The former recognizes whole molecules and their states as the causally relevant objects and properties. The latter, on the other hand, explains the same event (broadly construed), yet makes no mention of molecules nor of molecular states. Both explanations, however, truly describe the causal processes taking place during that event. Realism alone does not exclude this as a possibility.

There is a related point (the second manner in which it could turn out, on Lewis´ original definition of "law of nature", that there are no laws because there are multiple best sets). Even within a level (i.e., within a scientific discipline), it might be the case that an event is over-determined,[15] perhaps because two theories have laws which quantify over the states of different objects. This possibility is perhaps less obviously compatible with realism than over-determination by virtue of the existence of causal processes at distinct levels. Even here, though, I see no basic antagonism with realism: perhaps the world just is so constructed that there is a

[15] I use the term "event" broadly, such that the same event can be picked out in two different ways of carving the world into objects and their states.

basic indeterminacy in the natural joints that carve it into its causally efficacious objects. The picture of such a possibility is of a dynamic world in a superposition of causal descriptions. This superposition is different from that as normally understood within quantum physics, in that the totality of facts (stated in a causation-neutral way) for the particular moments of time are the same across the superposed causal worlds. With this as a possibility, one needn´t be forced into an exclusive either/or position, whereby one model´s being true implies the other´s being either false or true at a distinct level of description of phenomena. If this seems like a non-sensical feature for a theory of causation to have, like something a phlogiston-theory based researcher would say against mounting evidence in favor of the opposing view ("maybe we can both be right"), I beg to differ. For one thing, it is an empirical matter whether the phlogiston theory was true. It turned out not to be. But more importantly, as the recent developments in physics have shown, our common notions of what sort of theories "make sense" is changeable. Even though the acceptance of over-determination within a level as a genuine possibility consistent with realism is not absolutely necessary (indeed, I nowhere make explicit use of this as a possibility), it does make the overall question that I am addressing in this dissertation more approachable: if I can isolate the question "are the two models qualitatively distinct?" from the need to choose between the two, it will make the conceptual analysis required to answer it easier.

A final set of remarks concerning my understanding vis-a-vis the reductionistic versus supervenience versus strict autonomy of levels views of science is in order. As stated previously, I hold the view that the sciences form a quasi-hierarchy. (The "quasi" is inserted for two reasons: (1) it seems to me that a failure to qualify the word "hierarchy" gives the impression of strict reductionism, which I do not wish to imply, and (2) it may turn out that some sciences (a good candidate being biochemistry), while circumscribed enough to constitute a science, nevertheless cut across the levels defined by other disciplines.) There are then several possibilities with respect to the relationship between the objects whose states are quantified over at the various levels. The first possibility to consider is that the objects are strictly autonomous. Under this assumption, it could happen that two objects in the same world are identical in every physical respect, yet disparate with respect to some or all chemical, biological, or psychological respects: the chemical, biological and psychological levels are wholly autonomous from the physical. As I am assuming a physicalistic metaphysics, this is not a serious option. Supervenience and reductionism are both consistent with physicalism. The source of the general supervenience hypothesis is Donald Davidson´s "Mental Events":

> Mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.[16]

The literature sparked by this hypothesis has been immense, and has produced several distinct flavors (strong versus weak, global

[16] "Mental Events", in *Essays on Action and Events*, page 214.

versus local) of supervenience. Of most concern to me, however, is the relation of supervenience to reductionism, and the appearance of incompatibility of a non-reductive (but supervenient) layering of the sciences with the existence of causal laws at levels other than basic physics. (For a sample expression of this view, see Loar´s *Mind and Meaning*, pages 15-25.)

Supervenience was put forward as a way to make the existence of mental objects compatible with physicalism. Previously, it was believed that physicalism implied reductionism -- the view that object state types outside the purview of basic physics were not only implemented in basic physical stuff, but also identical with basic physical stuff. In particular, non-basic object state types were identical with some (likely very complicated) complex of basic object state types. Were this the case, causal laws relating states of non-basic objects would be strictly-speaking superfluous. Relating this to my theory of causation, the simplicity criterion for lawhood would rule out as potential laws of nature any such redundant generalizations. The fear among philosophers in general, and philosophers of mind in particular, was that reductionism implied the epiphenomenalism of all non-basic object states. (Of special concern to philosophers of mind was the feared epiphenomenalism of mental states, although, as remarked previously, all of the special sciences are in the same boat with respect to the epiphenomenalism of their object states.) Supervenience was seen as a way out -- it allows a connection of implementation of non-basic states in basic state complexes, yet avoids the strict reducibility of the former to the latter.[17] One advantage of reductionism over supervenience is the philosophical cleanness with which one can explain regularities of state transitions among non-basic objects. If those objects (or, more correctly, object-types) are complex aggregations of basic objects, and the causal laws at the level of basic physics happen to be such that the non-basic objects persist, giving the appearance of regularity of state transitions, then some (if not ontological, at least pragmatic) room is made for causal laws among non-basic objects. With supervenience, however, the lack of strict basic-object-state-type/non-basic-object-state-type equations seems to rule out the possibility of non-basic causal interaction, the desire for which was the very thing that led philosophers to embrace it. This is because causal interaction presupposes a causal law, and, without a law-like equation relating non-basic-object-state-types to their implementing basic-object-state-types, there is no mechanism to undergird non-basic causal laws. (Loar argues this point in *Mind and Meaning*, pages 16-17.) Given my previous assumption, however, this line of argument is unsound. Granted that non-basic objects are implemented in basic objects (note here that I am speaking of object *tokens*, not object types), it can still be consistently maintained that there are causal laws relating state transitions of non-basic objects types. For example, suppose I have a certain desire (to drink some water) and a certain belief (that there is a glass containing water in front of me), and suppose it is a psychological law that such beliefs and desires cause water-drinking behavior, then (ceteris paribus, of

[17] This is in theory what supervenience accomplishes, yet there is not unamimity on this point. See, for example, John Heil´s *The Nature of True Minds*, Chapter 3 for an overview of the debate.

course) I will drink the water (that is, that desire and belief will cause my water-drinking behavior). Now, supervenience requires that my belief and desire are implemented physically (either in the current state of my nervous system or in the current state of my nervous system plus certain physical relations with the external world).[18] It is not a requirement of logic, however, that in order for a law to relate the belief/desire state with the behavior, there must be a law-like equation of the belief/desire state type with a basic physical state type and the behavior state type with another basic physical state type. In this sense, supervenience represents a middle ground between the complete autonomy of levels and the complete reducibility of the less-basic level to the more basic. In any event, I do not wish to commit myself to either supervenience or reductionism as the correct understanding of the relationship between the sciences: I take it as an empirical question, well outside the purview of philosophy.

Given the large number of amendations to Lewis´ theory that have been made in the last several pages, it may be useful to summarize my theory of causation without reference to Lewis. (I should reiterate that what I have been trying to do in this section is to lay out a theory, not argue for it. My occasional use of motivating arguments has been intended more than anything as a means of clarifying my own view, particularly in respects in which it diverges from one more commonly held.) My theory of causation is realistic: there are as a matter-of-fact laws of nature, and the objects whose

[18] To use Davidson´s terminology, the mental event described and the physical event that implements it are one and the same event, under different descriptions.

state transitions are quantified over in those laws really do exist. There is a matter-of-fact about whether a particular counterfactual conditional is true. This matter-of-factness is grounded in the existence of a real similarity relation which orders possible worlds as a function of how close they are to the actual world. To be a causal law is to be a member of the set of generalizations that serve as axioms in any of the "best sets" of axioms that describe the actual world and all of its closest possible worlds, where "best" is understood as involving strength and simplicity considerations. Thus, the actual world and all of its nearest neighbors share the same set of laws. Causal laws quantify not only over the transitions of objects in basic physics, but also those of objects in the special sciences.

Given my agnosticism in choosing between reductionism and supervenience, I must separate my description of how I square the strength/simplicity requirement for sets of causal laws, with the existence of causal laws at levels other than that of basic physics.

(1) (Assuming reductionism) In this case, non-basic laws are, when one considers just the actual world, redundant. The state type transitions quantified over in non-basic laws are equivalent to complex state types at the basic level. However, as the set of "facts" (both actual and those at the actual world´s nearest neighbors) that need to be deducible from the set of axioms, given the initial conditions in each of those worlds, may involve differing manners of reductions (i.e., the reduction equations at the actual world may differ from those in the nearby worlds), the door is left open to non-basic causal laws. I am assuming that this is the case, for otherwise

my analysis does not allow non-basic laws (because any non-basic generalizations would be excluded from the "best set" by the simplicity criterion). Admittedly, it may be that the reduction equations in the actual and nearby worlds are identical: it is merely another asumption of mine that they are not. As an example, on my analysis it may be the case that chemical objects are reducible to physical objects in some nearby world in a manner other than the way they are reducible in this world.

(2) (Assuming supervenience) The difficulty here arises not out of a fear that the simplicity requirement will rule out redundant non-basic laws, as per above, but out of a concern that the existence of ceteris paribus laws is threatened: one might imagine that the simplicity and strength criteria would prefer strict laws over non-strict laws, leaving the latter again non-members of the best set. My answer here is that the notions of simplicity and strength are so undeveloped that I cannot say for sure that ceteris paribus laws are obviously ruled-out. The fact that the current state in the special sciences is such that ceteris paribus laws are universally used lends credence to their usefulness. Whether this translates into such laws´ being in the best set is yet another assumption on my part for which I have no argument.

2.2 What is the Mental?

First and foremost -- I am working on the assumption that eliminativism is false: there is such a thing as the mental realm, populated by causally-interacting mental states. In the final analysis, the truth or falsity of eliminativism is an empirical issue, determined by whether generalizations quantifying over state transitions of mental objects manage to make it into the best set; hence, an attempt at an a priori argument against eliminativism is certain to fail right from the get-go. (This is, however, not to say that an a priori argument against the belief in eliminativism suffers the same fate.) The relative success of mental-based explanations of some human and animal behavior lends a modicum of support to my assumption, but it hardly constitutes an argument. This is the last time I shall mention eliminativism as a hypothesis.

As the section title indicates, I shall attempt here to describe my theory of the mental: in particular, what distinguishes mental states from non-mental states? Stated glibly, the mental is the level (or levels) of description corresponding to the true psychology. This needs quite a lot of unpacking. I do not hold the view that science is complete, hence, that psychology as it is currently practiced must correspond to the *true* psychology. However, I accept what I take to be the two underlying assumptions of psychology. The first of these is that psychology fits into the quasi-hierarchy of science. That is, psychology is compatible with physicalism (either via reduction or supervenience of its objects on more basic physical stuff). Also, psychology, like other scientific disciplines, is concerned with the discovery of the causal laws grounding the regularity of succession among the states of its objects.

The second underlying assumption of psychology is that the vast majority of the states over which its causal laws quantify are contentful.[19] Hence, psychology is the science concerned with discovering the causal laws pertaining to being an intentional agent, such that those laws are expressed using intentional terms. Although I want to postpone a detailed explanation of how mental states get their meanings according to the two mental models being considered, I feel compelled at this juncture to at least suggest some possible approaches one might take to answering this question. I do this in order to avoid the appearance of assuming something (i.e., the existence of meaningful states) which is an utter impossibility according to physicalism. Two approaches that have been widely discussed in the literature are the evolutionary role explanation and the causal pathways explanation. Briefly, the first (teleological) view contends that our physical states can come to have meaning by virtue of being correlated with a condition in the external environment. In particular, a physical state comes to refer to an external condition because those ancestors of ours who succeeded (by the evolutionary standard of success: reproduction) did so at least partly in virtue of the correlation of this internal physical state with the survival-relevant environmental condition. The most familiar proponent of such a view of meaning is Ruth Millikan.[20] She summarizes her basic approach to explaining how physical states can come to be meaningful in terms of the proper functions (also called the "teleo-functions") of those states:

[19] Recall, I am allowing that some non-intentional states (eg, pain states) may also play a role in psychological laws; hence, I cannot make the blanket statement that *all* states are meaningful.

[20] See particularly *Language, Thought, and Other Biological Categories*.

> To describe the biological function of an item is ... to describe the role that its ancestors played in a particular historical process, a concrete cyclical process of birth, development, and reproduction extended over a number of generations. It is to tell how earlier items involved in this historical process that are homologous to this functional item characteristically contributed to continuation of the cycle (thus helping, of course, to account for this item´s existence).[21]

She continues:

> The position is that psychological classification is biological classification: hence proceeds by reference to teleo-function. This means that categories such as belief, desire, memory, percept and purposive behavior are biological function categories.[22]

This theory, if successful, grounds the contentfulness of certain physical states naturalistically: it does so without having to deny physicalism (although, more than the complete current physical description of an intentional agent is needed in order to determine what, in particular, a specific physical state means). Millikan remarks on this feature of her theory:

> The position is that intentionality is grounded in external natural relations, Normal and/or proper relations, between representations and representeds, the notions "Normal" and "proper" being defined in terms of evolutionary *history* ... [T]his means that there is not a way of looking just at a present-moment, eg, ... at his neural network patterns, that will reveal even the intentional nature of his ... inner representations, let alone reveal *what* these represent.[23]

[21] "Explanation in Biopsychology" in *Mental Causation*, pages 211-212.

[22] "Explanation in Biopsychology", pages 212-213.

[23] *Language, Thought, and Other Biological Categories*, page 93.

A second candidate theory for the naturalization of content is what I shall call "the causal pathways approach". As a first approximation, a certain physical (brain) state of mine has a particular meaning by virtue of being caused by the entity that it represents. In naive terms, a brain state of mine (for example, a brain state that occurs while I am looking at a cat) means this cat by virtue of being caused by the cat -- counterfactually, had the cat that my brain state represents not been there, that brain state would not have occurred. For cases of a meaningful physical state instantiation not *directly* caused by its referent (i.e., the non-perceptually-based production of a meaningful physical state), the content of the state needs to be explained in other terms than by its immediate distal cause. In fleshing out this general framework, Jerry Fodor develops and refines what it is for a physical state (or, using his terminolgy, a "tokening") to *refer to* something. He does this in terms of the counterfactually-based asymmetric dependence of the causal relation in false tokenings (between the distal cause of the tokening and the tokening) on the causal relation in true tokenings. Formally, his criteria for a tokening, "X" meaning X are:

1. \`Xs cause "X"s´ is a law.

2. Some "X"s are actually caused by Xs.

3. For all Y not = X, if Ys qua Ys actually cause "X"s, then Ys *causing* "X"s is asymmetrically dependent on Xs *causing* "X"s.[24]

[24] *A Theory of Meaning*, page 121.

As stated above, I do not want to argue for or against either of these theories of meaning; rather, I include them to head off a criticism that the very idea of meaningful physical states is absurd.

I now return to my description of the true psychology. This general approach to circumscribing a scientific discipline is not unique to psychology. One could similarly describe biology as the science concerned with discovering the causal laws underlying a subset of processes (i.e., those pertaining to life), such that those laws are expressed in the vocabulary of biology. Some caveats are in order. First, psychology is not about explaining *all* behavior of intentional agents. (For example, it is no criticism of psychology that it cannot explain -- ie, that its laws do not include -- reflex reactions, for such behavior is not intentional per se.) Rather, psychology is the science that seeks the causal laws underlying the behavior of intentional agents qua intentional agents.

A second caveat relates to a general character of the domain of psychological states. While I am not a behaviorist, I do maintain that psychology must (like all other scientific disciplines) ultimately deal with external behavior: its ultimate experimental domain is observable. This is not to say that there will not be many intermediate, non-observable states posited as parts of causal chains; however, the initial cause and final effect must be observable. All of the initial causes, final effects, and intermediate states are the purview of psychology, so long as they play a role in causal laws and are expressed in intentional terms. A second reason for tying psychology down to behavior is that, unlike in other sciences, there is a tendency within psychology to populate the realm of non-observables with states based only upon the subjects´ post-introspective reports. Introspection, however, is notoriously untrustworthy. (This is not a problem for other sciences: it is not even an option for a physicist to ask an electron "How do you feel, spin-wise?") True psychology, properly practiced, cannot assume that certain non-observable states obtain, merely because subjects say that those states obtain (however, subjects´ thus saying so *is* behavior -- hence, in the purview of psychology). This requirement also keeps an illegitimate argument relevant to my topic here from gaining acceptance. The argument goes:

> (P1): Subjects describe their mental states in the language of traditionalism.
>
> (P2): Subjects always introspect and (when attempting to be sincere) describe their mental states accurately.
>
> (P3): The language of traditionalism is not the same as the language of PDP.
>
> (C): PDP cannot be the true model of the mental.

Thus, my way of describing true psychology does not assume that introspection gives an accurate "snapshot" of the subject´s mental state. Whatever mental states take part in causal laws will fall out on their own.

A third question which often arises in the context of discussions of alternative systems as mental models is: what types of beings have minds? I hope that my (biologically-neutral) way of describing psychology will save a lot of quibbling on this topic. If dogs´ (or computers´, or martians´) behavior is non-trivially[25] explainable by reference to causal laws quantifying over intentional objects, then such beings have minds. Similarly, the fear that clearly non-intentional-state-possessing objects (e.g., lecterns) will also be included in the set of things possessing minds, because their behavior is "explainable" in terms of causal laws quantifying over their "mental" states (e.g., their "desire" to remain where they are) can be likewise allayed within my framework, for such entities´ behavior is only trivially "explained" by reference to causal laws quantifying over intentional objects.

[25] What I mean here by inserting the modifier "non-trivially" is something like this: one can assume that the causes and effects in causal laws quantifying over intentional objects will form a network of relations (ie, the objects will be mentioned as causes or effects in *many* causal laws). To cite a particular example using the traditionalist model: my belief that p may be a cause (or one of several collectively nomologically-sufficient conjuncts) in several causal laws, and the effect (or one of several nomologically-necessitated effects) in other causal laws. In order for an entity´s behavior to be *non-trivially* explained by reference to laws quantifying over intentional objects, the entity must be such that at least counterfactually, it can participate in a large subset of these causal laws, where the accessibility relation for the resolution of counterfactuals is limited only to the actual world and its nearest neighbors. (Recall that on my construal, causal laws -- both mental and non-mental -- are shared by the actual world and all of its nearest possible worlds.) Lecterns and the like fail to satisfy this "non-triviality" constraint -- the number of causal laws quantifying over intentional objects in which they can actually and counterfactually participate is very limited. This condition of non-triviality also rules out the existence of so-called "punctate minds": "minds" only capable of having one or a very few intentional items. That such a non-triviality condition is not unique to the discipline of psychology can be seen from the following: certain non-living entities may be seen as exemplifying a biological law. For example, it has been argued by some that the earth´s biosphere+atmosphere as a system is subject to many of the homeostatic laws regulating the behavior of biological entities, yet I (and, I think, most biologists) would not therefore maintain that such a system constitutes a *living* entity, because the number of biological laws that such a system could participate in is so very limited. Similar considerations can be seen in the debate as to whether viruses are living. Previously, it was believed that the number of biological laws in which viruses could potentially participate was very small (in particular, a subset of the laws dealing with reproduction). The consensus has now shifted in favor of viewing viruses as living entities, as their capacity for participation in a much larger subset of biological laws as come to light.

A final clarification: I take the proper role of psychology to be the discovery of *all* of the intentionally-based laws, not just that subset which, for lack of a better way of describing it, I call "the subset corresponding to the activities of our `reason´." It is an empirical issue whether there are intentional laws covering "non-reasonable" or "sub-reasonable" aspects of mentation. This is particularly relevant to the traditionalism/connectionism debate, as one part of Fodor´s language of thought argument against proponents of PDP is that their model cannot explain the "fact" that the belief that if p then q and the belief that p is nomologically sufficient for the belief that q. Whether this rule is a causal law is an empirical issue, just as it is an empirical issue whether the belief that if p then q and the belief that not-p is nomologically sufficient for the belief that not-q.[26] I hope to describe psychology so as not to prejudice either in favor of or against a model that is closely connected with the existence of only rationally-defensible inferences as existent. These caveats have been explicitly given so as to make it clear that I am not building a pro-traditionalism or pro-PDP bias into my construal of psychology -- and, hence, into my construal of mental causation.

2.3 Problems of Applying Notion of Causation to the Mental Realm

There are in general two major potential problems in applying the notion of causation to state transitions in the mental realm. The

[26] In my dealings with UMass freshman, I have found equal empirical support for the latter (fallacious) inference as for the former.

first of these has already been briefly mentioned: the supposed epiphenomenalism of psychological states assuming their dependence on more basic physical states. The second (and, I think, more threatening to the status of psychology as a science) involves the causal efficacy of psychological states qua intentional. In this section I shall describe each potential problem in detail and offer a solution. Following this I enumerate some lesser considerations in applying the notion of causation to the mental realm.

While many philosophers over the centuries have argued that physicalism (whether reductive or supervenient) implies the inefficacy of psychological entities, the locus classicus for this view is the first half of Norman Malcolm´s "The conceivability of mechanism". In this paper, Malcolm describes a hypothetical (but potentially realizable) completed neurophysiology "which is adequate to explain and predict all movements of human bodies except those caused by outside forces".[27] This completed theory makes no mention of intentional states in any of its laws.[28] Thus, even if one could identify generalizations describing psychological state transitions, the states quantified over would be inefficacious.

[27] Page 45.

[28] I should note that Malcolm´s portrayal of the intention/purpose-based rival to this neurophysiological theory is *not* what I (or, I think, most other philosophers) have in mind when speaking of the underlying laws to be discovered by psychology. Indeed, Malcolm himself notes the tautologous nature of the straw-man psychology that he puts forward, as in: "Premises of the other sort [used in psychological laws adverting to intentional states] express a priori connections between intentions (purposes, desires, goals) and behavior." (Page 50.) And a paragraph previously: "Thus the universal premise of a purposive explanation is an a priori principle, not a contingent law." (Page 49.) As stated previously, my construal of the role of psychology is as the discoverer of the *contingent* causal laws relating psychological states. Neverthless, I think the questions Malcolm raises concerning the efficacy of psychological states given the hypothesized completed neurophysiological theory are equally applicable to a more robust picture of psychology.

Malcolm describes the rivalry between the neurophysiological and psychological explanations of a man´s behavior as he climbs up a ladder to fetch his hat:

> Given the antecedent neurological states of his bodily system together with general laws correlating those states with the contractions of muscles and movements of limbs, he would have moved as he did regardless of his desire or intention. If every movement of his was completely accounted for by his antecedent neurophysiological states, ... then it was not true that those movements occurred *because* he wanted or intended to get his hat.[29]

Malcolm is assuming here that allowing determination of the same (broadly-described) event via two causal chains at distinct levels of analysis is a philosophically-unacceptable form of overdetermination. However, as Fodor (and others) have pointed out, this overdetermination of causal transitions is a feature not only of psychological laws, but of all laws in the special sciences. One could just as easily construct an argument against chemistry by substituting the phrase "chemical law" for "psychological law" and the phrase "underlying physical explanation" for "underlying neurophysiological explanation".

I see two reasons for rejecting Malcolm´s conclusion that, assuming such a completed neurophysiology, psychological entities are inefficacious. The first is the reductio ad absurdum that consistency would thus demand that all non-basic entities (interestingly, including neurophysiological entities) are causally

[29] Page 53.

inert. Thus, all the laws put forward by all sciences apart from basic physics are not really laws. We are more willing to give up on the demand for non-overdetermination than on the status of the special sciences as sciences; hence, psychology is saved.

In contrast to this negative argument, there is a positive argument for the salvation of psychology. (While I do not think it originated with Putnam, his name is most closely associated with it.) Were one to insist that only those generalizations stated in the vocabulary of basic physics can count as genuine causal laws, one would miss out on a large number of counterfactual-supporting generalizations relating state transitions of objects, where the properties describing the states are not in theory expressible in the vocabulary of basic physics. Indeed, the objects whose states would partake in such regular transitions would not be mentioned by the causal laws. (This is, I take it, true even if one is a reductionist and assumes that the reduction relations in the actual world are also valid in the nearest neighbor possible worlds.) The description of such counterfactual-supporting, but non-basic, generalizations in the vocabulary of basic physics would constitute a very ungainly generalization indeed. Imagine a generalization encoding some non-basic counterfactual-supporting rule (for example, the rule in biology that, without energy input, a gradient of some substance across a permeable membrane tends to equalize itself) in the vocabulary of basic physics. Such a generalization would most likely be a huge disjunction, in order to account for all the types of membranes, all the ways those membranes could be permeable, all the types of gradients, etc. (I doubt that there is some basic physical property that, for example, all membranes have, that might rein in the number of disjuncts constituting the generalization.) Nowhere in this generalization is there any mention of membranes or permeability, which are necessary entities in seeing this generalization *as a true generalization.* Here the strength and simplicity criteria for the inclusion of a generalization in the set of causal laws argues in favor of the inclusion of such generalizations, for a very large number of counterfactual-supporting regularities would otherwise be missed. And, as per above, the vocabulary specifying the relevant objects and their states participating in such regular transitions must be that of the corresponding level (rather than that of basic physics) in order for the "generalization" to be a generalization at all. So, even assuming a reductionist physicalism, room can be made for causal laws outside of basic physics.

There is a second potential problem in applying the notion of causation to state transitions in the mental realm. It concerns the issue of how to understand the causal efficacy of psychological states qua intentional. More specifically, the problem runs as follows. The intentional content of a mental state does not in general supervene upon the intrinsic physical properties of the object possessing that state. Twin-earth thought experiments illustrate the essential extrinsicness of content, in that they describe a situation in which two people with physically-identical bodies fail to share all of the same intentional states; hence, intentional states require something more for their discrimination. In particular, they require the consideration of certain relational properties between the person and the world (usually understood in terms of the causal histories leading up to the tokening of a particular mental state). One needn´t rely upon such exotica as molecular duplicates on twin-earth to see that this rather homely point is true. What distinguishes the reference of a visual experience of a desert oasis in the veridical case from that of a mirage has nothing to do with the experiencer´s physical state: we can stipulate sameness of physical state in both cases. Rather, what distinguishes the fact that in the one case the reference of the physical state token that implements the percept is the nearby oasis whereas the reference of the other is something else[30] are the extrinsic properties of the experiencer. Granted then that contentful states supervene on both intrinsic *and extrinsic* physical properties of a subject, how can such states participate in causal interactions -- how can a subject´s extrinsic properties be causally relevant? (Note that, unlike the case with the first potential problem described at the beginning of this section, this problem is, among the special sciences, particular to psychology; hence, a Fodor-style reductio won´t help.)

I see here two ways of answering this question. The first route basically acquiesces (i.e., answers "they can´t"): while meaning does indeed supervene on both extrinsic and intrinsic properties, the former are causally inert. This "methodological solipsism" has the advantage of side-stepping this potential problem, but at an enormous cost for psychology as the science of intentional states, for it puts psychology in the untenable position of simultaneously maintaining that there are causal laws relating contentful state

[30] I am not sure what the reference of an illusion should be. In any event, it cannot be the real nearby oasis, because, per supposition, there is no such thing.

types, yet each (tokened) causal state transition is not causal by virtue of being subsumed under this law (for the law makes reference to states in terms of their contents, which, on such a view, are inert), but by virtue of some implementing mechanism. This would set causation as understood in psychology totally apart from causation in the rest of science, where it is the very properties mentioned in the causal law that are causally responsible for its efficacy. To make a parallel with the previously-mentioned causal law of biology dealing with equalization of a gradient across a membrane, it would be as if being a membrane were causally inert, even though it is by virtue of being a membrane that something is subject to this law. Thus, taking this tack with respect to the second problem leaves one no protection from criticism with respect to the first: the reason that the reductio is so intuitively effective is that it is based upon the assumption that psychology is just like the other special sciences; however, on this view, psychology is qualitatively distinct from the other special sciences.

I find the advantage of the methodological solipsistic stance is more than outweighed by its disadvantages (which, when taken to their extreme, render psychology a non-science). Yet, I also take the second problem of mental causation as serious -- hence, as requiring a response. So, how do I explain the causal relevance of mental states qua intentional? Here is an instance in which the specificity with which I described my assumed theory of causation will pay off, because, it allows a rather straight-forward explanation of mental causation. (I am also of the opinion that much of the literature on this subject consists of philosophers talking at cross purposes, because they are each making differing assumptions with respect to the nature of causal interaction.)

Before giving my response I would like to state the problem in what I take to be its most compelling form. Even if we grant that some physical state tokens can have the property of meaning x, that is still a long way from demonstrating that that property is causally relevant, for a physical state that constitutes a cause has many, *many* properties that are not relevant to its forcing the transition to the effect. Perhaps the property of meaning x is one of these. So, to cite a famous example,[31] while the physical state corresponding to the soprano´s singing a high-C has the property of meaning A, that property is causally irrelevant to the glass´ shattering. Maybe the meaningfulness of all physical states (including the states of cognitive agents´ nervous systems) is likewise causally irrelevant. Maybe methodological solipsism is the best that we can get.

A solution to this problem would consist in identifying a causally relevant difference between the property of meaning A in the soprano case and the property of meaning A for a psychological state -- a difference that can in principle leave the door open for the causal relevance of meaning.[32] My causal intuitions tell me that the meaning of the soprano´s words is causally irrelevant -- how is that to be interpreted within the framework of my theory of causation? I understand that as the counterfactual "had the soprano sung something with a different (or with no) meaning, while all other

[31] Dretske, *Explaining Behavior*, page 79.

[32] I should re-emphasize, whether meanings are causally relevant is, I take it, an empirical issue. What I am about here is demonstrating that such a thing is not in principle ruled out by my theory of causation.

properties of the physical state corresponding to the singing remained either unchanged or changed as little as necessary given a change in meaning, the glass would still have shattered". Hence, in the nearest possible world in which she sang something with a different meaning, the glass shattered. (I assume that the nearest possible world in which this is the case is a nearest neighbor of the actual world.) Therefore, there is no causal law relating the meaning of the words with the glass´ shattering.

What about the cases in which it is presumed that the meaning of a physical state token *is* causally relevant? Let´s return to Malcolm´s case of the man climbing a ladder because he wanted to fetch his hat. As mentioned earlier, Malcolm holds that "[g]iven the antecedent neurological state of his bodily system together with general laws correlating these states with the contractions of muscles and the movement of limbs, he would have moved as he did regardless of his desire or intention". Is this the correct counterfactual to use when assessing whether meaning is causally relevant? I think not. Furthermore, I believe this way of construing the problem of mental causation leads unavoidably to methodological solipsism. Compare the two counterfactuals, where A is the agent, N is the physical state possessed by A in the actual world that produced behavior B (broadly construed), and N means M in the actual world:

> C1: If it were to be the case that A was in state N, but N did not mean M, then B would not have been produced.

> C2: If it were to be the case that A was not in a state with meaning M, then B would not have been produced.

Malcolm (and many other philosophers) chose C1 as the correct counterfactual to use in considering whether meaning is causally relevant, whereas I maintain C2 is the correct one. It is obvious how the adoption of C1 as the correct construal leads to difficulties: if it is a law that N´s produce B´s, then in all the nearest neighbors in which A was in state N, B would be produced. The question then becomes: are any of these N-worlds also worlds in which N does not mean M? The answer here is not so clear. If we are convinced by the twin-earth thought experiments that meaning resolution involves consideration of the causal history of a physical state tokening (so N includes not just an "at-this-moment" snapshot of A´s neurological state, but also some of A´s relational properties), one could attempt to argue that there are no nearby possible worlds in which N does not mean M; however, I am not convinced that this must be the case. Therefore, the construal C1 leaves open the door for someone like Malcolm to argue for the causal irrelevance of meaning.

On construal C2, this argument is blocked. Let´s consider how C2 is analysed. We examine the nearest possible world in which possessing the meaning M is not a property had by any of A´s states. This could be so either because A was not in state N (in which case, other things being equal, B would not be produced -- therefore, C2 is true) or because, a la C1, A was in state N, but N did not mean M (which leaves us back in the ambiguous case noted above). I maintain that the first of these two possibilities is closer to the actual world than the second, because the number of important changes needed to get from the actual world to the first is less than that to get to the second. Because the meaning of an intentional state depends not just on the intrinsic state, but also on the causal history of the agent, the move from the actual world to the nearest possible world in which the same intrinsic state had a different meaning would require changing the causal history of the agent, in comparison with the relatively smaller change in the intrinsic state of the agent needed in the nearest possible world in which the agent failed to have that intrinsic state. This change would also require other changes (in particular, the minimum number of changes necessary to accommodate the change in intrinsic state); however, the overall quantitive amount of change is less is this case. While this does not prove that there are *in fact* causal laws quantifying over intentional entities, it at least demonstrates that such a thing is not ruled out; one might rephrase this as: the problem of mental causation is not a conceptual problem (hence, not a problem for philosophers) but rather an empirical issue.

There are a few other lesser difficulties that relate specifically to the application of causation to the mental realm. The first of these involves the possible difference between ceteris paribus conditions in psychological laws and in the laws of other special sciences. As mentioned previously, the ceteris paribus condition is meant to encode the background assumptions that must be satisfied for the cause to be enabled to force the effect, where these assumptions pertain to conditions at the implementation level;

hence, if made explicit, the ceteris paribus conditions would be stated in the vocabulary of the implementing mechanism. In the case of psychological laws, however, it is sometimes assumed that at least some conjuncts in the ceteris paribus clause are themselves also at the psychological level. For example, one often reads in the traditionalistic AI literature of the extreme difficulty of enumerating all of the background beliefs, desires, etc. relevant to a psychological law.[33] I think, though, that this is an inappropriate use of the ceteris paribus condition: those "background" beliefs, desires, etc. do not constitute a background to psychological laws in the same sense as a properly functioning brain constitutes a background. On my view, such background beliefs, desires, etc. belong in the body of the psychological law. If philosophers like Dreyfus are correct in maintaining that the totality of background beliefs, desires, etc. are not enumerable, then the enterprise of traditionalist psychology is called into question. I don´t have any particular counter against

[33] Sometimes this difficulty is viewed as merely pragmatic: it is hard to enumerate them all, but not in theory impossible. Sometimes, though, the "difficulty" is portrayed more as a theoretically insurmountable hurdle, founded on an essential differentness of psychological laws. A proponent of this view is Hubert Dreyfus. He states in his paper "From Micro-worlds to Knowledge Representation: AI at an Impasse":

> My thesis ... is that whenever human behavior is analyzed in terms of rules, these rules must always contain a *ceteris paribus* condition, i.e., they apply "everything else being equal," and what "everything else" and "equal" mean in any specific situation can never be fully spelled out without a regress. Moreover, this *ceteris paribus* condition is not merely an annoyance which shows that the analysis is not yet complete. ... Rather the *ceteris paribus* condition points to a background of practices which are the condition of the possibility of all rulelike activity. In explaining our actions we must always sooner or later fall back on our everyday practices and simply say "this is what we do" or "that´s what it is to be a human being." (Page 92).

Dreyfus´ attack. In any event, the burden of proof is on him to show (rather than merely assume) that this background is indeed either infinite or necessarily regressive. His more recent writings in praise of PDP leads me to believe that he doesn´t take his concerns as decisive against psychology per se, but rather as directed against cognitive psychology as it is currently embodied in its strong traditionalist form. As stated earler, I am not presuming that present psychology is the true psychology; rather, that there *is* a true psychology (i.e., a science that aims at discovering the causal laws explaining human behavior in terms of intentional states). I am not concerned in this paper with arguing for or against one or the other models.

Another consideration in applying causation to the mental realm involves the possibility of there being more than one distinct level of organization of intentional entities that are causally related. One sees a similar phenomenon within the (widely-construed) discipline of biology as the science of entities qua living systems. This discipline encompasses causal laws at the organellular (i.e., pertaining to parts of cells), cellular, organ-level, organism-level, and ecological levels. Perhaps psychology is similarly laminar. One often sees such a hypothesis in the more ecumenically-minded articles describing PDP as a cognitive model.[34] According to this metapsychological thesis, there are two distinct levels of organization of entities qua intentional agents, each of which possesses its own causal laws stated in the vocabulary appropriate

[34] An example is Paul Smolensky´s "On the Proper Treatment of Connectionism".

to that level, such that the traditionalist level is implemented in the PDP level. This view allows that some PDP-modellable behavior is not describable in traditionalist terms, but all traditionalist-modellable behavior is implemented at the PDP level. Both levels are in the domain of psychology because the laws within each model make reference to *meaningful* states.[35] While this possibility is an interesting empirically-decidable issue, it is relevant to the question I am posing in this present work in so far as it requires that the two models correspond to *distinct* levels. Hence, the models are themselves distinct.

The possibility of two psychological levels may appear to pose problems for my theory of causation. Wouldn´t the strength and simplicity criteria rule out the adoption of the generalizations as causal laws at both levels? A related concern involves the mutual dependence (better known as circularity) of the entities and their states quantified over in generalizations, on the one hand, and the generalizations made, on the other. Which object-states are considered potentially causally efficacious determines which generalizations will be made. This is true not only at distinct levels, but also within the same level, when there are two or more competing ways of consistently "carving up" reality, each of which produces a causal web relating the states within that way to one another. We can assume that there is no theoretical reason to prefer

[35] There are also philosophers (noteably Fodor and Pylyshyn) who argue that while it is possible that the model proposed by PDP is a true model describing human behavior, and that it corresponds to the implementation level of traditionalism, it is not itself a model of the mind, because the states related by its causal laws are not intentional states. I shall have much more to say in Chapter 4 on if and how PDP can be viewed as a model of the mind.

one of the competing conceptualizations to another: both sets of entities and generalizations "cover" the facts (broadly construed); although, again, the "facts" (narrowly construed) to be explained differ from one conceptualization to another. Applied to the traditionalism versus PDP debate, this constitutes another ecumenical possibility. Unlike in the case mentioned previously, however, one would not be an implementation of the other. Does this possibility pose a problem within the framework of my theory of causation?

I think that in each case my theory not only handles these potential problems, but also leaves the door open to a philosophical analysis and comparison of traditionalism and PDP to an extent not possible within the framework of a causal theory that allows only one of several competing descriptions of the causally interacting world to be true. (That is, I can isolate the, for present purposes irrelevant, question of which model is correct from the question of how the models differ.) Recall that a generalization is a causal law by virtue of being a member of any[36] of the best sets of generalizations. The competing conceptualizations (if, per supposition, they provide equivalent explanatory power) make it into distinct best sets; hence, both sets of generalizations constitute causal laws. A similar result is obtained when the theories are at least partly at distinct levels (but where it is not the case that one completely reduces to or supervenes upon the other). I view this not as a problematic aspect of my theory of causation, but as a feature, both for providing a

[36] A reminder: This is a departure from Lewis´ original theory, in which a generalization was a causal law by virtue of being a member of *all* of the best sets.

framework for comparing traditionalism and PDP as well as for elucidating the relationship between competing paradigms within a discipline when both paradigms yield equally good strong-yet-simple generalizations.

While the above considerations show that my theory of causation has no theoretical difficulties in dealing with psychological laws, yet, still the doubt remains that it does have certain practical difficulties. In particular, the fact that most psychological laws will relate non-directly-observable states, combined with the relative generosity of my theory of causation in granting lawhood to generalizations (with a concomitant ontological commitment to the entities and their causally-interacting states), produces the fear that there will not be enough restraint placed on which generalizations are causal laws. This fear of a population explosion of causal laws and entities is ungrounded, as the criteria for lawhood will immediately exclude from any best set those generalizations which do not contribute to the explanatory power of the set as a whole. By way of illustration, consider the two sets of generalizations below:

Set 1:

.
.
.
Gn As cause Bs
Gn+1 Bs cause Cs
.
.
.

Set 2:

.
.
.
Gn   As cause Cs
.
.

Assume that Bs are mentioned only in generalizations Gn and Gn+1 within Set 1. If both sets are equally strong (i.e., postulating the causal efficacy of Bs does not produce any increase in explanatory power) then Set 1 is not a best set; hence, Gn and Gn+1 are not laws, and there is no requirement for an ontological commitment to any entities mentioned in B, unless it is mentioned elsewhere in Set 2. Thus, each best set will be minimal relative to a host of sets within a particular paradigm. While this example considers only the simple case of a superfluous intermediate state, rather than a superfluous web of states, I see no reason to doubt that the simplicity criterion will likewise eliminate superfluous states that appear as conjuncts in complex causes and/or effects.

2.4 Analysis of Mental Causation as Providing Model of the Mental

I began this chapter with a brief overview of what I think a model of a domain is, in order to motivate my subsequent wanderings through the topics of causation in general and mental causation in particular. Now it is time to re-examine the notion of modelhood in light of the previous three sections.

The first point to note is that, within common practice in science, the word "model" is used ambiguously to include both merely predicting (or, less pejoratively, simulating) models as well as explanatory models.[37] The former grouping includes those models that are constructed to mirror the state changes of the modelled system, without regard to whether the causal laws producing the state changes in the modelled system are the same as those producing the state changes in the modelling system. My previous example of the desk-top reproduction of the solar system is just such a simulating model. This remains true, even supposing that, by means of gears with carefully chosen ratios, I produce a dynamic desk-top reproduction that mirrors not only the relative positions of the real planets and sun at a particular time, but also the relative velocities and positions of the real planets and sun through time. In that case the toy solar system, while accurately reflecting the location-state transitions of the real system, does *not* do so by virtue of being subjected to the same causal laws underlying the location-state transitions of the real solar system. Likewise, computer models of physical systems (i.e., simulations by means of a computer of the state transitions of the modelled system) are merely predicting, even when the state transition predictions are based upon an encoding of the relevant causal laws underlying the state transitions in the modelled system. No one would say in either the desk-top reproduction case or the computer simulation case that the

[37] A more appropriate way of referring to this sort of model would be to use the phrase "implementing model". However, this usage may lead to confusions when I discuss the implementing level of a model. Hence, I have closen to use the less apt, but also less confusing phrase "explanatory model".

models *implement* the systems in question, where "implement" is understood in terms of the state transitions of the modelling system being produced *by the same causal laws* as those governing the state transitions of the modelled system. Hence, no one would argue that such models are explanatory models. In contrast, many models used in the sciences (the clearest example being animal models within medical research) are based on the assumption that the causal laws forcing state transitions in the modelled system (i.e., in the human) are reproduced in the modelling system (i.e., in the animal).

Unlike the typical computer simulation of the state transitions of a physical system (for example, the simulation of the progression over time of a thunderstorm) it is not a forgone conclusion that a computer model of the mind can be at most a merely predicting model. This is because the causal laws that must be captured in a concrete implementation of a mental model are not laws relating physical state types but intentional state types and behavior. And it is not clear whether or not the intentional state types and behavior that are quantified over in psychological laws are reproducible on a computer. I shall argue in Chapter 3 that traditionalism is committed to the theoretical reproducibility in a computer of both the intentional states quantified over in psychological laws as well as the psychological causal laws themselves; hence, that traditionalism is committed to the possibility of an *explanatory* model of mind implementable in any computational device with certain capabilities. In Chapter 4, I argue that PDP likewise assumes that their systems will constitute an explanatory model of the mind. The relevance of this will become apparent in Chapter 5, when I compare the two models. Traditionalism is committed to the computational nature of mental causal laws -- hence, to their computability. Some have argued[38] that the PDP model proposes (or, at least, does not rule out) non-computable psychological laws. If this is true, then the two models are distinct.

Both traditionalism and PDP offer a general summary of how mental processing works. From these summaries, one can tease out a theory of mental causation for each which describes the gross characteristics of object states that take part in causal interaction and the gross characteristics of the causal relationship. These form the body of the respective mental models. One can thus view a model of the mental realm as an abstract web whose interior "nodes" are the causally efficacious state types and whose "directed connections" are laws relating the partial cause and the partial effect. I have in mind something like the following (this might be a small section of a folk psychological causal web representing a model of the mind).

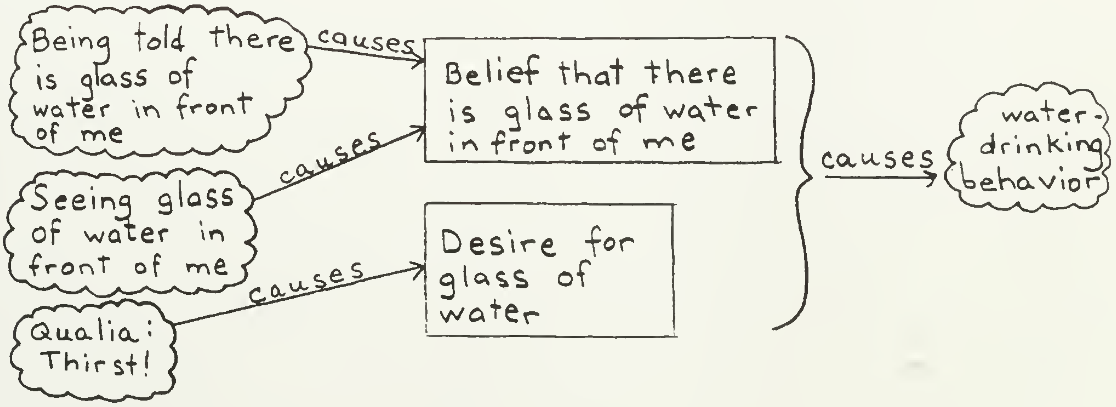

Being told there is glass of water in front of me
causes
Belief that there is glass of water in front of me
Seeing glass of water in front of me
causes
Qualia: Thirst!
causes
Desire for glass of water
causes
water-drinking behavior

Figure 1 -- Portion of folk psychological web

[38] See for example Cummins and Schwartz´ "Connectionism, Computation, and Cognition" in *Connectionism and the Philosophy of Mind.*

In order to avoid misunderstandings I should reiterate that I am not concerned here with what particular intentional states and causal laws are mentioned in the causal web (that is an empirical matter) but with the gross characteristics of the states and laws. For example, what level of reality is represented by the causally efficacious intentional states by each of the two models? By considering the above question, one can determine the ontological commitments made by a particular model. In addition, traditionalism and PDP place constraints on what sorts of causal laws are allowed and/or obligatory.

One final concern that I shall only mention here (but treat in detail in Chapter 4) involves an ambiguity within the PDP literature regarding which level of analysis of a PDP system to equate with *the* model of the mind being offered. Obviously, the laws regulating unit-level state changes and the intentional content of unit-level activation differ from the laws regulating pattern-level state changes and the intentional content of pattern-level activation; hence, the unit-level description of PDP systems produces a distinct model of the mental from the pattern-level description of PDP systems.

I would like now to summarize the most important points of this chapter. I take causation, whether involving object states in basic physics or in any of the special sciences (including psychology) as a real relation. A particular state transition is causal by virtue of being subsumed under a causal law. Psychological laws are distinguished by quantifying over intentional state types. To be a psychological law is to be a generalization that, relative to a consistent paradigm, best encapsulates the regularities of the behavior of humans (in addition to the other higher animals, and any other entities capable of being governed by intentionally-described states) qua intentional agents. It is possible that multiple self-consistent but mutually-incompatible paradigms describing the intentional level exist. That intentional states either supervene upon or are reducible to physical states in no way shows the mere epiphenomenalism of the former.

A model of the mind is first and foremost a theory of mental causation. This theory is brought to light by abstracting away from the particular vocabulary used to describe mental causal interaction and focussing on the characteristics (in terms of possible representational content) of the causally efficacious states and on the constraints placed on the possible causal laws relating these states. This abstract way of viewing a theory of mental causation permits formalization of causation in terms of a relation and relata, thus allowing a formal comparison of two presumably competing theories of the mind.

CHAPTER 3
TRADITIONALISM AS A MODEL OF THE MENTAL

Traditionalism has most often been described in its folk psychological version. As will become apparent later in this chapter, I believe that there is sense to be made of traditionalism sans folk psychology. However, the ubiquity of the examples illustrating traditionalism using folk psychological constructs, and more importantly, the dearth of non-folk psychological examples of traditionalist causal laws, lead me to introduce traditionalism by way of folk psychology.

In this chapter I shall describe what has been, until quite recently, the predominant view of the mental realm within psychology and philosophy in the latter half of this century. This view (variously called "classicism", "computationalism", and "traditionalism" -- I adopt the latter term for the remainder of this work) offers a model of the mind which, among other things, was the first to explain how our folk psychological theory of intelligent agenthood may be realized, without contradicting an underlying physicalist metaphysics. This fact is, I believe, the main reason for traditionalism´s popularity. The first section of the chapter is devoted to a description of traditionalism as it has emerged with the beginnings of AI (artificial intelligence) in the 1940´s. Along the way I make explicit some of the (oft-unmentioned) assumptions inherent in traditionalism, and enumerate some of the various flavors in which traditionalism comes -- this as an aid in identifying the absolute minimal commitments of traditionalism. The second section of this chapter takes up the topic of traditionalism as providing a model of the mind. In particular, I consider the questions: What ontological commitments are made? What form would causal laws posited by traditionalism take? Given that traditionalism proposes that representational states are causally efficacious, what level of reality do such states represent?

3.1 What is Traditionalism?

Folk psychology as a theory of intelligent agenthood has been around for a long time. While I dare not hazard to guess how long, it is clear that it pre-dates the advent of traditionalism. According to folk psychology, the behavior of certain entities (including humans and the other higher animals) is explainable by reference to the beliefs, desires, etc. of those entities: ie, those beliefs, desires, etc. are causally relevant to the behavior. Thus, one encounters in folk psychology such putative causal laws as:

> **If** A desires to drink some water, and
> A believes that there is a glass of water in front of A,
> **then** (ceteris paribus)
> A engages in water-drinking behavior.

There are several problems with folk psychology that led philosophers and psychologists to doubt that it could ever constitute a serious (i.e., scientific) theory of mind. The most obvious (philosophical) problem is how to square the causally efficacious mental states posited by folk psychology with physicalism; hence, how to find a place for the folk psychological ontology within the scientific quasi-hierarchy. (It is an assumption made by virtually all mainstream philosophers that any respectable scientific theory must presuppose physicalism -- to buck this assumption is to be immediately branded "fringe".) How could something mental like a belief be causally efficacious? Also, how could something as patently non-physical as meaning play a role in a causal interaction? A third concern (one whose consequences can be seen in the particular psychological theory put forward in contraposition to folk psychology in the latter 19th and earlier 20th centuries) involves how folk psychology could ever be transformed into a science, given that the causally efficacious states it posits -- ie, beliefs, desires, etc. -- are non-observable.

One found in associationism and behaviorism an attempt to formulate a theory that can explain the behavior of intelligent agents without recourse to their hypothesized mental states. Behind this movement lay the hope that purely physical causal pathways would be discovered (most likely, via the brain) linking stimuli and response. Were this the case, philosophical concerns about making the existence of mental states consonant with physicalism would be avoided -- there would be no mental states in the usual sense (i.e., as identifiable using a non-physicalist vocabulary).

Two events (or, more precisely expressed, two movements) in the mid-twentieth century turned the tide of favor within psychology and philosophy against the associationistic/behavioristic approach to explaining intelligent behavior and (back) towards a belief/desire-based psychology.[1] The first of these was the feeling (identified most strongly with Chomsky´s attack on behaviorism) that the associationist approach could not explain the facts of human behavior.[2] For later purposes in my comparison of traditionalism and PDP it is important to note that a working hypothesis of behaviorism was the learned nature of concepts; the attacks against behaviorism were more often than not attacks against this hypothesis. Indeed, Chomsky´s attack against behaviorism related to the inability of the latter to explain how children could learn the concepts expressed within the language, not how they could learn the language itself. Thus, in psychology in the post-behaviorist era, there has been a return to a Cartesian view of concepts: they are *atomic* entities that are built into the mind. (When speaking of concepts entertained by natural creatures, this corresponds to the thesis that concepts are innate.)

The second movement leading to the decline of behaviorism was the progress in AI in making the very idea of causally

[1] I want to reiterate that traditionalism need not be identified with folk psychology. (Indeed, traditionalism is silent on whether the causally efficacious mental states are beliefs, desires, and/or something else.) If anything, folk psychology is best viewed as one among many specifications of traditionalism. In its pure form, traditionalism is silent on which particular generalizations are causal laws. Folk psychology, on the other hand, includes (perhaps even consists solely in) the set of commonsense generalizations purportedly providing explanations and predictions of the behavior of mind-possessing entities. Many of these commonsense generalizations have been called into question by various schools within traditionalist psychology, and many additional generalizations not a part of the folk psychological repertoire have been advanced. A clear example of the latter category is Freud´s psychological theory, which, while belief/desire-based (hence, consonant with traditionalism) is not typically considered a part of folk psychology.

[2] The identified short-comings of behaviorism were based on a posteriori considerations: it was not that behaviorism was *in principle* incorrect as a theory of intelligent behavior, but rather that it failed to explain certain aspects of human behavior -- in particular, human language acquisition.

efficacious mental states conceivable within the framework of physicalism. The importance of the computer metaphor in bringing about the hegemony of traditionalism within cognitive psychology cannot be overstated. Philosophers saw in the computer an existence proof that one aspect of the mind-body problem (i.e., how a meaningful state could be causally efficacious) was readily solvable: the meaningful states were also (token identical with) physical states whose physical/syntactic properties were sufficiently related to their meaning, so that the transition from one physical state to the next mirrors the transition from one meaningful state to the next. In the case of the computer, the meaningfulness was derived -- so, the analogy between computer and human was not exact. However, it was believed that at least part of the mind-body problem was solved; all that remained was the naturalization of original intentionality. (This is a project that is still ongoing. See my Chapter 2, Section 3.) Giving the historical roots of traditionalism points out one of its underlying assumptions: any mental state, in order to be genuinely efficacious, *must be explicitly represented in a physical state.* This holds true both for systems (like the computer) with derived intentionality as well as for systems with original intentionality. Even the most abstract (i.e., remote from details of physical implementation) cognitive psychological diagram of the mind, with its belief boxes and arrows showing the flow of information, is based on this assumption: there must be a set of physical states and physical pathways that instantiate the depicted mental states. Jerry Fodor, perhaps the most unambiguous proponent of traditionalism, makes clear this underlying assumption:

> So, then, what exactly *is* RTM [his version of traditionalism] minimally committed to by way of explicit representation? ... According to RTM, mental processes are transformations of mental representations. The rules which determine the course of such transitions may, but needn´t, be themselves explicitly represented. But the mental contents (the `thoughts´, as it were) that get transformed *must be* explicitly represented or the theory is simply false. To put it another way: if the occurrence of a thought is an episode in a mental process, then RTM is committed to the explicit representation of the content of the thought.[3]

With this assumption as background, the psychologist, like any researcher in the other natural sciences, need not concern herself further with the particulars of implementation, but can remain conceptually isolated within the vocabulary of beliefs and desires.

The historical relationship between traditionalism and AI points out another of traditionalism´s groundlying assumptions: the computability of the function governing mental state transitions.[4] For present purposes, it is most fruitful to define computability in terms of rule-governedness of manipulation. Using the vocabulary transferred over from the computer metaphor, the representational

[3] *A Theory of Content*, pp. 23-24.

[4] Computability theory as a subdiscipline of computer science pre-dates by several decades the construction of the first electronic computing devices in the 1940´s; hence, the definition of what constitutes a computable function is given, not in terms of the modern von Neumann-style computer (with its CPU, instruction registers, and addressable memory), but in terms of the Turing machine. A function is computable if and only if it is Turing-computable (ie, if and only if there is a Turing machine that can, for each element in the domain of the function, return the function´s output for that element). It just so happens that the computational power of the universal Turing machine and the (non-resource bounded) von Neumann-style computer are the same: any Turing-computable function is von Neumann-computable and vice versa.

states being manipulated are explicitly stored, and the program that refers to and transforms these states corresponds to formal rules governing the manner of manipulation. (The computability thesis carries with it a set of restrictions on the form that mental causal laws can take. I shall discuss this topic later.)

We see in traditionalism an even stronger interpretation of the computer metaphor than as a mere analogy to aid in clarifying mental vocabulary. From the beginnings of AI it has been a thesis that, once the rules governing mentally-describable state transitions were discovered and encoded, a computer program implementing these mentally-describable states and their corresponding rules would not only simulate a mind, but implement a mind. In the vocabulary introduced in Chapter 2, the thesis is that an *explanatory* model of the mind (rather than a merely predicting model) is in theory achievable. Even the name first mentioned by John McCarthy in the 1950´s for the fledgling field reflects this assumption: notice the distinction between the import of the phrases "*artificial* intelligence" and "*fake* intelligence" -- the former implies that genuine intelligence, albeit via human-manufactured entities, is the goal, not a simulation of intelligence. This, along with the often tacit assumption that only things with minds can have genuine intelligence, implies that part of the goal of AI is the production of a mind. Perhaps the most well-developed espousal of this strong computational theory of mind view is to be found in Pylyshyn´s *Computation and Cognition*. A typical passage is:

> As we see below, in the case of cognitive psychology, explanatory adequacy depends on a stronger sense of

> equivalence [than mere correspondence of I/O behavior], particularly on knowing the details of the process at a suitable level of abstraction. What, then, recommends computation as the appropriate vehicle for *that* task? To provide a framework for discussing this question, let us first look at computation from a more abstract point of view. That will help bring out further similarities in the relationship of computational devices and computational processes, on the one hand, and brains and cognitive processes, on the other. It is the failure to distinguish computation as a type of process from the particular physical form it takes in current computing machines that has prevented many people from taking computation as a literal account of mental process. If we understand computation at a fairly general level (as, in fact, it is understood in theoretical computer science), we can see that the idea that mental processing is computation is indeed a serious empirical hypothesis rather than a metaphor.[5]

A similar statement *equating* mental processes with computational processes can be found in the writings of other traditionalists, such as Fodor. For example:

> There are, as it happens, some reasonably persuasive theories about the nature of such mechanisms [dealing with mental phenomena]. The one I like best says that the mechanisms that implement intentional laws are computational.[6]

One can see now the truly pivotal role that the emergence of AI has played with regard to saving a place for causally efficacious mental states within the framework of physicalism. Computers (like brains) are physical devices, subject to physical (and chemical and

[5] Page 55.
[6] *A Theory of Content*, page 145.

thermodynamic, etc.) laws. While embodying a computation process, a computer has, in addition to its physical (and chemical and thermodynamic, etc.) states, certain computational states. Computational states possess some interesting properties, three of which are particularly relevant for present purposes. The first is that, while implemented in a physical medium, such a state qua computational is not physical. Rather, the property that makes it the computational state that it is is its functional role within the context of an abstract process. This functional role is understood in terms of the relation between this state and its preceding and following computational states. A particular physical state implements a particular computational state by virtue of being a token of a member of an equivalence class of physical state types that are related to other equivalence classes of physical state types in the same way as the corresponding computational states are related to one another. A second relevant property of computational states is the rule-governedness of their succession upon one another. One can identify, by means of an algorithm (at the abstract level) or a program (at a more concrete level), the rules that govern the manipulation of data structures.

Finally, and perhaps most importantly in seeing the contribution of the computational theory of mind to the partial solution of the mind/body problem, computational states are intentional states. As already mentioned, their intentionality is strictly derived from the original intentionality of the observer/creator of the computational process. When embodied in a concrete computational device, the physical characteristics of the physical state that implements a particular computational state bear a relation to the meaning of that state, such that the physical laws that force the transition from the current physical state to the next physical state that is a token of a member of a distinct equivalence class constituting another computational state are isomorphic to the rules governing transitions between the computational states. This correspondence is "built in" to the computer: the designer designs it so that this correspondence between the computer´s physical properties and its computational properties (when it is engaged in a computational process) obtains. (I am finding it very difficult to express this in English -- see the diagram for a pictorial representation of this relationship between physical states and computational states.)

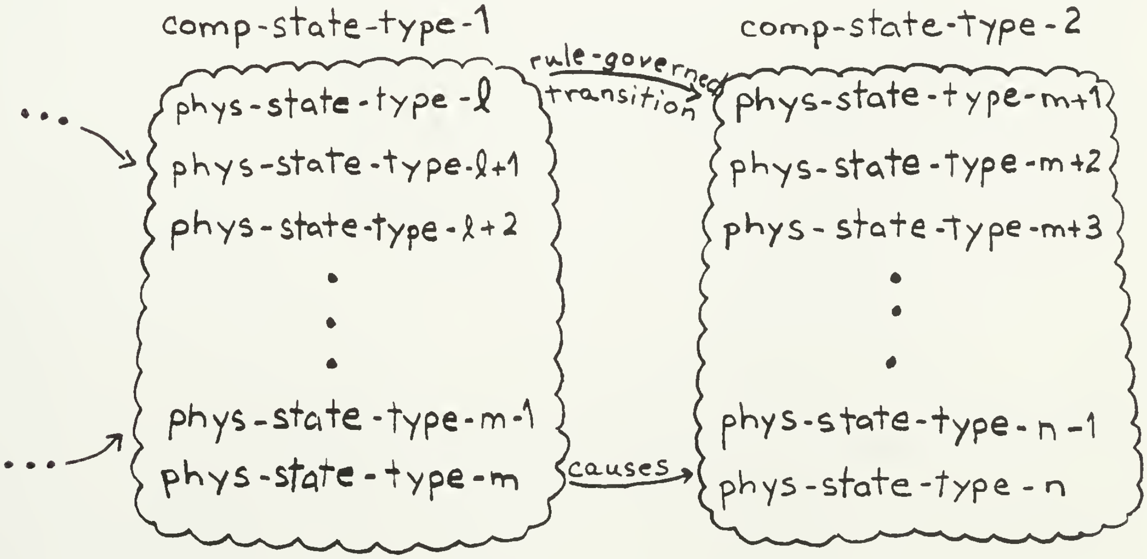


Figure 2 -- Relationship between physical and computational states

We see one single computational state transition depicted in the diagram (i.e., the transition from comp-state-type-1 to comp-state-type-2). At the computational level of description, the transition is governed by one of the rules that constitute the (abstract) computational system. It is possible to implement a computational system in a physical system (e.g., in a von Neumann-style computer). Each computational state type corresponds to an equivalence class of physical state types, as shown. Computers, as artificial devices, are designed to take advantage of relevant physical causal laws to allow relatively easy implementation of computational systems: one can guarantee (ceteris paribus) that the computer´s physical state transits from one of the states in phys-state-type-l ... phys-state-type-m to one of the states in phys-state-type-m+1 ... phys-state-type-n if and only if the computational system´s computational state transits from comp-state-type-1 to comp-state-type-2. A particular physical state of a computer (e.g., a token of phys-state-type-l) is meaningful by virtue of instantiating a computational state type (for this case, it inherits the intentional content of comp-state-type-1). For the case of a computer, we can guarantee a correspondence between computational states and equivalence classes of physical states only because the (human) designer of the computer has built in the correspondence. What guarantees such a correspondence between computational/mental states and equivalence classes of physical states in the human, or any other natural being? How a traditionalist responds to this question depends upon which method for the naturalization of original intentionality is assumed. A traditionalist leaning towards the evolutionary approach to naturalization will answer that, in the history of a species, it has offered selectional advantage to have a body (or, more narrowly, a nervous system) whose physical states follow upon one another in the manner of Figure 2. Whether or not such a correspondence between classes of physical states of the nervous system and computational states obtains is an empirical issue which is not yet decided.

While the computational-nature-of-mind thesis is in one sense liberating to the study of mental phenomena (in that it frees the psychologist from a concern for the implementational details of the computational/mental states), it can also be seen as constraining. This is because it limits the candidate functions describing the mental realm to the set of computable functions. There are, however, many functions which are known to be non-computable (i.e., there is no computational system that computes these functions). Perhaps the mind is a system that implements a non-computable function. The ramifications of this possibility are only recently becoming understood within psychology. Although rarely mentioned explicitly in traditionalist writings, the computability assumption is so integral that it cannot be removed from traditionalism without destroying the integrity of the entire model. This is because the intentional content of the physical state implementing a mental state is determined by the computational state type that it is a token of. (I am taking as the received view among traditionalists the thesis that computational statehood carries with it wide content. As mentioned in Chapter 2, I assume that psychological phenomena are identified by their subsumption under causal laws that advert to (wide) representational content.) But something´s being an instance of a computational state type makes sense only within the context of a computational process -- a process which, at a minimum, can be described by formal rules.

This restriction to the set of computable functions constrains the sorts of learning in which a traditionalist system may engage. In particular, learning must be confined to changes in the manipulated structures, for changes to the program take us outside of the realm of computation. This restriction makes sense, given how the program is interpretted within tradtiionalism: the program encodes the mental causal laws, which themselves remain unaffected by learning. In order to stay within the guidelines set by the computationalist assumption, the program is unalterable. Some may object that computers running self-altering programs are not only conceivable but also actual, and this is certainly true. However, in implementing such a function, the computer is not implementing a computational process. (In general, a physical computer is capable of performing many tasks other than computing functions. For example, a computer can implement the function $d = 0.5\ a\ t^{**}2$ when I drop it out a window; however, the computer´s states relevant to its implementation of this function (namely, its displacement from its location of release) are non-computational. It is important to keep in mind that what the computationalist assumption (and its accompanying restriction) buys for traditionalism is representational content.

As we shall see in Chapter 4, PDP systems, while most often *in practice* limited to the implementation of computable functions, are not *in theory* thus limited. Whether PDP researchers believe that the possibility of implementing non-computable functions is an important feature of PDP systems vis-a-vis mental modelling is not an issue on which there is any consensus in the literature. I shall discuss this topic in more detail in Chapter 5, Section 4.

In the last several pages I have described the set of hypotheses within traditionalism stemming from its AI roots: (1) mental states are computational states and (2) these mental/computational states must be explicitly represented. I would like now to examine more closely the features of these states. A major assumption within traditionalism is that mental states are often structured:[7] they are composed of parts, each of which, in combination with its "semantic" position within the state, contributes something to the overall meaning of the mental state. The meaning is nothing over and beyond the synthesized meaning of its parts. In addition, the meaningfully-relevant parts correspond to physically-isolable structural parts of the physical state that implements the mental state. Additionally, any meaning-relevant position is also reflected by some physical relationship. So, if a mental state consists of three parts (mental-part1, mental-part2, and mental-part3) such that this specific order is important (i.e., a different ordering of the parts would constitute a different mental state), then the physical state that implements it will also have three parts (phys-part1, phys-part2, and phys-part3), such that phys-part1 implements mental-part1, phys-part2 implements mental-part2, and phys-part3

[7] On this view, not all mental states need be structured, but the vast majority will be.

implements mental-part3 and this ordering of the parts is somehow encoded. When asked to give a reason for insisting on *structured* representations, traditionalists most often cite Fodor´s language of thought argument. As I discuss this argument in Section 2 of this chapter (in the context of an examination of the level of reality represented by mental states), I shall not do more than merely mention it here.

Traditionalism is the theory of mind most often assumed in mainstream (non-PDP) AI and cognitive psychology. It is well advised, therefore, to sample some of the writings from researchers in these fields, if for no other reason than to prove that traditionalism is not just a theory of interest to philosophers. Perhaps the earliest sign of an inclination toward traditionalism within the field that would later develop into AI is to be found in Alan Turing´s "Computing Machinery and Intelligence", first published in 1950, in which he refers to the "human computer" and compares processing in the digital computer with processing in the (human) mind.[8] Another important figure in the early history of AI, Allen Newell, describes the close relationship between psychology and computer science:

> [My purpose] is to call your attention to the use of symbolic models in many places through out experimental psychology. ... I maintain that a shift

[8] This paper originally appeared in *Mind* LIX, in October 1950, pp. 433-460. It is reprinted in *The Philosophy of Artificial Intelligence*, edited by M. Boden.

> in the Zeitgeist in psychology has taken place toward a view of man as an information processor.[9]

And, a few pages later in the same article:

> In the discussion of the possible relationship of information processing models to psychology we opted for the use of such models as detailed theories of behavior, rather than, say, metaphors or exercises in the discipline of operationalism.[10]

In a later work, Newell gives his own theory of human cognition. He writes:

> At this point I wish to be explicit that humans are symbol systems. ... They might be other kinds of systems [eg, biological systems] as well, but at least they are symbol systems.[11]

He defines "symbol system" in terms of being a "form of universal computational system."[12] I am taking Newell´s views as representative of those among researchers in AI, as is corroborated by Haugeland, in:

> Formal systems [ie, computational systems] can be interpreted: their tokens ["token" here is being used in a slightly restricted sense. A token is not any old instance of a type, but refers specifically to the

[9] "Remarks on the Relationship between Artificial Intelligence and Cognitive Psychology" in *Theoretical Approaches to Non-Numerical Problem Solving,* page 376.

[10] "Remarks on the Relationship between Artificial Intelligence and Cognitive Psychology", page 378.

[11] *Unified Theories of Cognition,* page 113.

[12] *Unified Theories of Cognition,* page 76.

> physically instantiated objects (i.e., concrete realizations of computational states) manipulated by a system] can be assigned meanings and taken as symbols about the outside world. ... [I]f artificial intelligence is right, the mind itself is a (special) interpreted formal system.[13]

Traditionalism is also the model of choice among cognitive psychologists. I take John Anderson as typical of that group.

> Production systems are particularly general in that they claim to be computationally universal -- capable of modelling all cognitive activity.[14]

One point of contention amongst traditionalists relates to the issue of the requirement for the *explicit* representation of the rules/program governing the transition from one computational state to another. As the above-quoted passage from Fodor demonstrates,[15] he is willing to allow that the rules may be built in. Thus, it may be that mental processing corresponds to computational processing on a dedicated, rather than on a general purpose computer. Newell and Simon, on the other hand, understand the computational nature of mind thesis as requiring a separation into universal computational device and particular program running on that general purpose device -- thus, the rules, along with the manipulated symbols, must be explicitly represented. I do not take this difference as crucial to the understanding of traditionalism, for

[13] *Artificial Intelligence,* pp. 99-100.

[14] *The Architecture of Cognition,* page 13.

[15] "The rules which determine the course of such transitions may, but needn´t, be themselves explicitly represented." *A Theory of Content,* page 24.

all three agree on the computational nature of mind, and on the thesis that a special purpose device (e.g., a special purpose Turing machine) can implement a computational system. Hence, I side with Fodor in allowing non-explicit rules within traditionalism. The issue is often brought up in the context of comparisons between traditionalism and PDP, in that the latter not only permits non-explicit rules, but requires it by the very nature of PDP systems. I shall have more to say on this topic in Chapters 4 and 5. Suffice it for now to note that, as I do not believe the explicit representation of rules is a necesary condition for a traditionalist model, I will tend to be dismissive of this line of argument in distinguishing the two models.

I am able at this point to state what I take to be the defining marks of traditionalism: a realist understanding of mental causation, the equation of mental processes with computational processes, and the structured nature of the mental/computational representations being manipulated. Tienson has summed up this view quite nicely as follows:

> The "twin minimal commitments" of [traditionalism]: syntactically structured representations and structure sensitive, rule-governed computational processes.[16]

It may be useful to examine traditionalism in light of various criticisms of it raised by dissenting philosophers; I do this in order to make crystal clear some of its ramifications. The most famous opponent of traditionalism is John Searle. In his oft-cited article

[16] From the introduction to *Connectionism and the Philosophy of Mind*, page 23.

"Minds, Brains, and Programs"[17] his rejection of traditionalism can be summed up by his rejection of the computational theory of mind: he states that "no reason has been given to suppose that when I understand English [he uses understanding English here as an archetype representing *all* mental processing] I am operating with any formal program at all."[18] Rather, he claims, causal laws relevant to mental goings-on (including those relevant for establishing a mental token as meaningful) reach down to the biological level; hence, it is in theory impossible to implement an explanatory model of the mind in a computer.

Steven Stich, on the other hand, accepts the computational theory of mind (in a slightly modified form), but rejects the traditionalist hypothesis that the mental tokens thus manipulated are *representations* (i.e., are intentional). He summarizes the conclusion of his arguments against the strong representational theory of mind (i.e., a theory according to which mental laws advert to *contentful* states):

> The question at hand is whether the notion of belief and related folk psychological notions will find a comfortable home in cognitive science. One view that urges an affirmative answer is the Strong Representational Theory of Mind, which sees a mature cognitive science postulating representational states and adverting to content in its generalizations. ... [T]he cognitive scientist is ill advised to adopt the Strong RTM paradigm.[19]

[17] Originally published in *The Behavioral and Brain Sciences* 3 (1980), pp. 417-424. Reproduced in *The Philosophy of Artificial Intelligence*, edited by M. Boden.

[18] *The Philosophy of Artificial Intelligence*, page 71.

[19] *From Folk Psychology to Cognitive Science*, page 160.

He puts forward an alternative theory of the mind (dubbed the Syntactic Theory of Mind) which is computational, yet "[avoids] any appeal to content in cognitive generalizations."[20]

A third route to the rejection of traditionalism is to adopt an instrumentalist construal of mental state tokens. Under the rubric "instrumentalism" I include both the standard interpretation (a la Dennett) and (for lack of a better label) the non-explicit representationalists. Both groups reject the *realist* assumption of traditionalism -- they reject the thesis that there corresponds, for each mental state token, an implementing physical state token.

A final group of cognitive scientists (most of whom are experimental psychologists working in the field of mental imagery--eg, Kosslyn and Shepard) argue against traditionalism on the grounds that it gets the nature of the structure of mental representations wrong. They base their opposition on a set of famous psychological experiments,[21] which purport to show that (at least some) mental processing consists of rule-governed manipulation of pictorial, rather than quasi-linguistic, tokens. (That is, they maintain that the manner in which the part/whole relation of some mental tokens is to be understood is in terms of the part/whole relation typical of pictures or images, rather than that of complex sentences.) I believe that Fodor´s language of thought argument (to be discussed in Section 2 of this chapter) presupposes that the structure of causally efficacious mental tokens is quasi-

[20] *From Folk Psychology to Cognitive Science*, page 160.

[21] See, for example, *Mental Images and their Transformations*, by R. Shepard and L. Cooper.

linguistic. While I would not want to base my interpretation of traditionalism solely on the views of Fodor, the centrality of the LOT argument in the traditionalism versus PDP debate inclines me to exclude, at least in the context of this work, pictorialists from the traditionalist camp.

3.2 Traditionalism as a Model of the Mental

In this section, I convert traditionalism (as defined by a commitment to a realist theory of mental causation, in which mental processes are computational processes, such that the mental/computational representations manipulated are structured) into a model of the mind. Before I begin, I would like to reiterate that a model of the mind is an abstract model, in that it makes no mention of concrete mental causal laws. In particular, I want to explicitly distinguish traditionalism and folk psychology, because the latter includes a set of presumed causal laws. Indeed, the above description of traditionalism is even silent on whether causally efficacious mental tokens are beliefs, desires, and/or any other type of attitude that one typically finds in folk psychology. What exactly the true mental tokens would be, if not beliefs, desires, etc., is unclear; however, traditionalism in its purest form allows that they might be some other sort of representational states.

As described in Chapter 2, Section 4, a model corresponding to a particular theory of the mind is intimately related to the form that the causal laws take within that theory. Hence, I begin this section with an analysis of mental causation according to traditionalism.

While traditionalism is committed to the physical instantiation of all mental states, mental causal laws quantify over states qua intentional -- ie, a state takes part in a mental causally-determined transition in part because it bears a particular meaning. (It may also be that some other aspects of the state are causally relevant -- for example, that it instantiates a certain attitude of the agent towards the representational content. Using the belief/desire-based specification of traditionalism, the physical state may instantiate a *belief* that p. This is a distinct mental state type from a *desire* that p, even though the representational content of the two states is identical. Presumably, mental causal laws advert to both content *and* attitude.) The rules that specify the transition from one computational state to another *are* the mental causal laws.

One important aspect of causal lawhood (both on my particular construal given in Chapter 2, Section 1 and in general) involves the analysis of counterfactuals. How are they to be understood a la traditionalism? Consider a particular hypothesized mental causal law:

C´s cause E´s

where C and E are mental state types, each of which may be a complex conjunction composed of more basic mental state types. Suppose that a particular instance (i.e., a token) of C (call it "c") occurred in the actual world, and that it was followed by a particular instance of E (call it "e"). Given the relationship between mental states and physical states according to traditionalism, this means that there was a token of some physical state type which implemented c, in that it was a member of the equivalence class of physical state types that form the computational/mental state C. A similar relationship holds between e and the physical state that implements it. Call the state type of the physical state that implements c "Phys-C", and call the physical state token that, on this particular occasion, implements c, "phys-c". Similarly, Phys-E is the physical state type and phys-e is the token of the physical state that implements e. As mentioned, we are assuming that "C´s cause E´s" is a law, and that c (an actual token) caused e. How are we to understand the following counterfactual?

> If a C hadn´t occurred,
> then an E would not have occurred.

We first consider the nearest possible world in which this instance of C did not occur (presumably, this world is one of the nearest neighbors of the actual world). Now, there are three possibilities to consider with respect to the properties at this possible world. Either:

(a) phys-c occurred, but phys-c in this world is not a token of C, or

(b) phys-c did not occur, but some other state (call it phys-o) did occur, and phys-o is an instance of Phys-O, which is a member of the equivalence class constituting C in the actual world, but not a member of the equivalence class forming C in this possible world, or

(c) phys-c did not occur, and no other token of a state that is a member of the equivalence class constituting C in the actual world occurred.

Recall in Chapter 2, in the discussion of analysis of counterfactuals in general, I argued that the possibility represented by (a) above is *not* the possible world to consider in analysing the corresponding counterfactual, for it is farther away from the actual world than (b) or (c). We can see the similarity of this case with the one used in Chapter 2 by considering the following: phys-c´s instantiation of the mental state type C gives to phys-c in the actual world its meaning. So, possibility (a) is analogous to the case from Chapter 2 in which the man was in the same physical state, yet that physical state did not correspond to the belief and desire that led him to climb the ladder in the actual world. But the difference between the actual world and this world needed to effect the change in meaning for the identical physical state is greater than the difference between the actual and some other world in which the meaningful state is altered because the physical state is altered. Thus, whether a token of E occurs in the possible world corresponding to (a) is irrelevant.

Now consider possibility (b). The antecedent of the counterfactual is satisfied only when Phys-O, while a member of the equivalence class forming C in the actual world, is *not* a member of the equivalence class forming C in this possible world. Thus, we have the same situation as that described for the above case: the equivalence class forming the computational/mental state is not constant across the possible worlds -- ie, from the actual world to the possible world described by (b). And, as above, the quantity of change needed to effect this is greater than the quantity of change needed to go from the actual world to the possible world corresponding to (c). As above, therefore, whether a token of E occurs in this possible world is irrelevant.

So, this leaves us with (c) as the world to consider when analyzing the counterfactual: neither phys-c nor any other physical state token that is an instance of a physical state type that constitutes the equivalence class C occurs. If, as hypothesized, the E in the actual world was caused by the C in the actual world, then the physical mechanism implementing this interaction was phys-c´s causing phys-e. (This is so, because, according to traditionalism, phys-c´s being a token of C is conditioned upon its nomic ability to produce a physical state (in this case, phys-e) that is an instance of a member of the equivalence class that constitutes E). But, according to possibility (c), there is no reason to believe that an E would be caused, unless E was overdetermined in the actual world. (That is, it is also a causal law that, for example, C*´s cause E´s, and both an instance of C as well as an instance of C* (call it "phys-c*") occurred, and both were nomologically sufficient for the production of phys-e). In this case, however, we would not say that the C caused the E, but rather that both the C and the C* caused the E. Hence, we needn´t consider this possibility. So, in this, the closest possible world, the counterfactual turns out true, as expected.

Up until this point, I have been somewhat sloppy in my characterization of computational/mental states. I would like to think that my sloppiness is merely a reflection of the sloppiness to be found in traditionalist writings on this topic -- given that I have often been engaged in summarizing the views of others in the first part of this chapter, it is reasonable that I should adopt the vocabulary used in the literature. Now, however, I must of necessity adopt a more precise and standardized vocabulary. According to the classical computer science definition, a computational state of a computing device consists of the *complete* computational state of the device at a time step. Using the Turing machine as archetype of a computing device, a computational state is the triple:

<current-machine-state,[22]
tape-contents,
position of read/write head>[23]

Thus, the Turing machine progresses from one computational state to another at each time step.[24] A concrete Turing machine, were one to construct it,[25] would be a physical device with physical components (a tape, a R/W head, and some sort of controller that could store the current machine state and perform the physical actions corresponding to the formal actions specified in the machine state tansition table). Corresponding to each computational state is a (very large) equivalence class of physical states that can instantiate it. One can think of this physical diversity as the consequence of implementing a quantized device in a world whose ultimate level of

[22] One of a finite number of states accessed by the state transition table. The machine state makes no mention of head position or items stored on the tape.

[23] An equivalent formalism for describing the computational state of a Turing machine is: <current-machine-state, contents of tape to left of R/W head, contents of tape from R/W head to the right (inclusive)>.

[24] Turing machines, like digital computers, are assumed to have a clock that synchronizes all the changes necessary in going from one computational state to the next. Thus, computational states are quantized: it makes no sense to ask, for example, "What state is the machine in as the R/W head moves from square 201 to square 200?".

[25] Technically-speaking, this is not possible, because a Turing machine has access to an unbounded amount of tape. As no one believes the (non-resource bounded) Turing machine is true in all respects in depicting the attributes of the mind, this short-coming can be glossed over.

quantization (if there is one) is much lower than that of the device. So, there might be many distinct ways of being in the state <S24, 0010000..., 3>. Some of the diversity results from the various precise physical configurations corresponding to the R/W head scanning the 3rd square of the tape. There are many other sources of diversity. (As mentioned, in general, the equivalence class of physical state types will be very large.) A von Neumann-style computer, while differing in some of the above details, produces an analogous picture. The traditionalist thesis is that so too does the human nervous system (or, more generally, the nervous system of any creature possessing a mind).

There is, however, a slight problem of terminology in squaring this (monolithic) view of a computational state with the more finely-grained usage outside of the theory of computation literature. Namely, psychologists and philosophers talk as though it is a proper part of the complete computational/mental state of an entity that is causally responsible for some change of mental state or some behavior. Returning to the folk psychological example, it is only my desire for water and my belief that there is a glass of water in front of me that causes my water-drinking behavior. I have many, many other beliefs, desires, etc., that play absolutely no role in this causal sequence. In particular, my belief that 2+2=4 is causally irrelevant to my drinking. However, on the monolithic view imported from computer science, it is my *whole* computational/mental state that caused my water-drinking behavior.

Is there a way to make psychological and philosophical usage of the computational theory of mind consistent with computational theory simpliciter? What traditionalism-cum-folk-psychology needs is a way of subdividing this monolithic computational/mental state into substates that correspond to beliefs, desires, etc., while preserving the individuation of these substates along computational lines. There is, I think, one avenue open to the traditionalist. Recall that one of the commitments of traditionalism is to structured parts: the parts of the mental state (i.e., the individual conjuncts constituting the monolithic computational/mental state) correspond to physically-isolable parts of the physical state implementing the mental state. (While usually intended --eg, in Fodor´s argument from systematicity of mental representation[26] -- to cover the part/whole relation between single propositions and their constituents, the structured nature of mental states is a thesis that also applies to conjunctive mental states.[27]) So, on this theory, each conjunct is itself physically-isolable: within each physical state type that is a member of the equivalence class that constitutes a mental state is a physically-isolable "sub"-state type, tokens of which implement the corresponding mental state parts. The "sub"-state types are themselves physical state types. We have the following picture:

[26] See, for example, the section on the systematicity of cognitive representation in Fodor and Pylyshyn´s "Connectionism and Cognitive Architecture: A Critical Analysis", pp. 37-41.

[27] See, for example, the section of the systematicity of inference from the same work, pp. 46-48.

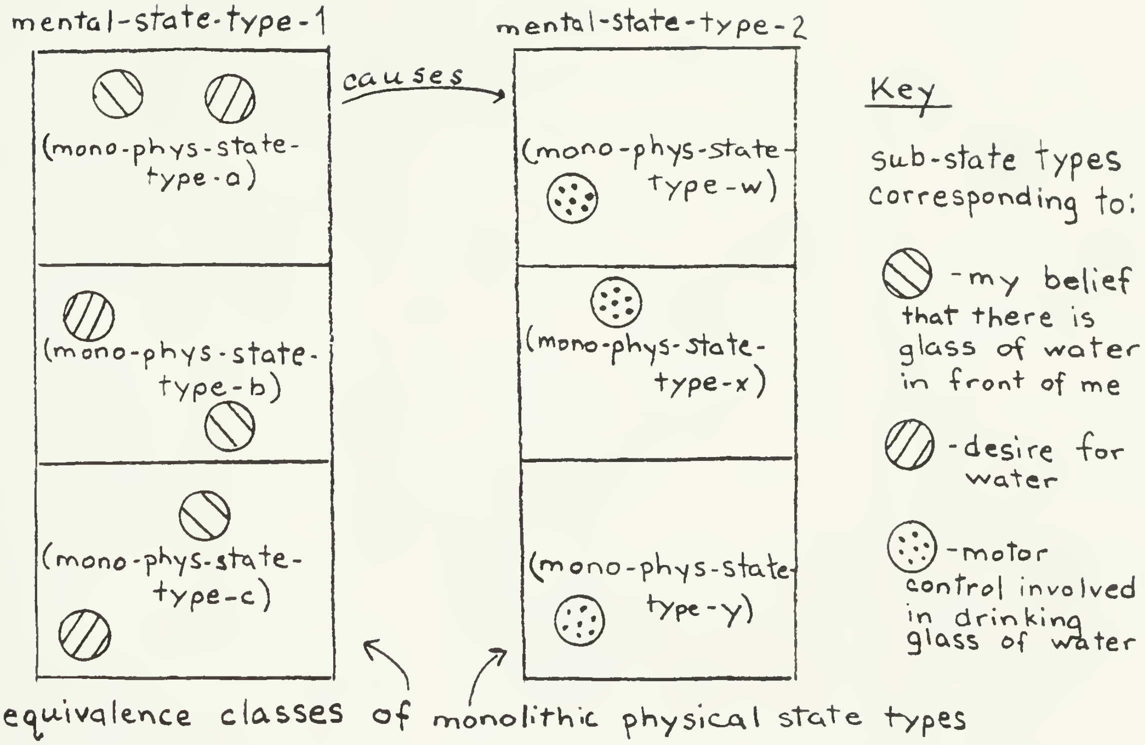


Figure 3 -- Relationship of monolithic and sub-state types

To say that my belief that there is a glass of water in front of me and my desire for water caused my water-drinking behavior is to say that, for each of the monolithic physical state types that is a member of the mental state type, only a subset of that monolithic physical state is, strictly-speaking, necessary for the production of water-drinking behavior. (In particular, when mental-state-type-1 is instantiated by a token of monolithic-phys-state-type-b, only the subset of b identified by the two circles within b are necessary to cause the circle in monolithic-phys-state-type-w.[28]) Clearly, how

[28] The subset need not be limited to spatial parts of the monolithic physical state; rather, it is a subpart that is isolable using physical vocabulary. Spatially-isolable parts are only one among many parts thus isolable.

structured and causally-isolable the monolithic-phys-state-type parts are is an empirical issue.

In the idealized case where such parts are perfectly causally-isolable, a picture emerges of multiple computational processes running in parallel, implemented in a single sequence of monolithic physical state tokens. It is analogous to a parallel-processing computer executing several programs simultaneously. There are two computational levels: one corresponding to the overall computational process encompassing all of the subprocesses, and a second level corresponding to many separate computational processes, one for each individual program running in parallel. I think it is this picture that best fits the terminology adopted by mainstream traditionalists. (It should be noted that parallelism as described above does not increase the computational power to a level above that of a serial von Neumann-style computer: both (non-resource bounded) serial and parallel computers can compute exactly the same set of functions as a Turing machine.) An aside: this picture, in conjunction with the further thesis that the various processes running in parallel are relatively compartmentalized, results in faculty psychology, a la Fodor´s *Modularity of Mind.*

It is interesting to note the relationship between this picture and my comments in Chapter 2 on how to interpret "ceteris paribus" in the context of mental causal laws. Recall that, in that discussion, I distinguished two interpretations assigned to ceteris paribus clauses. The first supposes that ceteris paribus clauses, if cashed out, would be seen to encapsulate a bunch of background beliefs and desires that, while strictly-speaking necessary, are omitted because this background has proven recalcitrant to enumeration. I rejected that explication of "ceteris paribus" as insufficient: if the background beliefs and desires are nomologically necessary to produce the effect, then they belong in the body of the causal law. The second interpretation of the role of ceteris paribus clauses also has it that they stand for the background assumptions causally necessary for the effect. However, on this interpretation, these assumptions are at the level of the mechanism implementing the causal law (hence, if cashed out, they would be stated in the vocabulary of the discipline implementing the causally-related states). For example, suppose that a desire for water and a belief that there is a glass of water cause water-drinking behavior, ceteris paribus. Most traditionalists take neural hardware as either the implementing level, or perhaps, the implementing level of the implementing level of mental states. (In any event, not too far below mental states in the scientific quasi-hierarchy.) The ceteris paribus clause is not satisfied when the neurological hardware is not functioning as assumed. Were I to have particular tokens of the above-mentioned belief and desire, yet, just as my water-drinking behavior was about to commence, I suffered a serious stroke, or was shot in the head, the ceteris paribus conditions would not be satisfied, and water-drinking behavior would not be caused. Similarly, were the motor end of my central nervous system to sudddenly become damaged, the expected water-drinking behavior would not commence. This is, I think, the only interpretation that can be consistently maintained. Within the context of traditionalism, it means something like this: each monolithic physical state type that is a member of mental-state-type-1 shares the "sub"-state types that implement the belief and desire, as shown. Strictly-speaking, these "sub"-states are not nomologically sufficient for the production of the "sub"-state-type corresponding to the water-drinking behavior. Rather, in at least one of the monolithic-phys-state types, there is an additional part of the physical state that, in conjunction with the instantiation of the belief and desire, cause the effect. However, this part is not shared by each of the monolithic physical state types constituting the equivalence class in such a way that it could either form a new computational subpart or be consistently encompassed within the belief or desire. So, the boundaries of the belief and desire, if interpretted as surrounding the parts of the monolithic physical state types causally sufficient for production of the effect, "leak" a bit. Looking back to my example of failing to engage in water-drinking behavior because of damage to my motor control system, the current state of my motor control system is not a part of the relevant belief or desire, yet is included in the monolithic physical state type describing my current physical state. This aspect of my physical state prevents the water-drinking behavior: it is in this sense that I say that the boundaries of the belief and desire "leak" in order to accommodate all of the causally relevant parts of my monolithic physical state.

Is this fatal for traditionalism? Does it show that those sneaking suspicions about belief/desire psychology in particular, and the computational nature of mind thesis in general, were justified after all? One possible line of argument against traditionalism, based on these considerations, is that the computational states which play such a central role within traditionalism could not be isolated if they "leak"; hence, there would be no physically-isolable units to be the bearers of meaning. I really don´t know whether such an argument could show that traditionalism is incoherent. My inclination is to pull a Fodor-style reductio: but the other special sciences suffer from the same "causal leakiness". In any event, I am not here so much concerned with attacking or defending traditionalism, so further discussion of this topic will have to be postponed until another occasion.

I cannot, however, postpone any longer an examination of Fodor´s language of thought argument and its consequences for traditionalism. In particular, I shall focus on what the LOT argument has to say about the level of reality that is represented in causally efficacious mental states. Fodor´s argument[29] is supposed to show that certain features of the cognitive capabilities of humans and other mind-possessing entities are best explained by postulating a language of thought, whereby mental representations possess combinatorial syntax and semantics, and the processes that manipulate those representations are sensitive to their structure. His argument can be summarized as follows: traditionalism is the model of the mind that best explains certain empirical features of cognition. These features of cognition are: (1) the systematicity of inference, (2) the systematicity of mental representation, and (3) the productivity of mental representation. I shall focus on the first two

[29] Actually, the LOT argument did not originate with Fodor, but rather, with Chomsky. I associate it with Fodor because, particularly in the context of the traditionalism versus PDP debate, he has been the most vocal promulgator of it.

of these on the way to arguing that the level of reality represented by causally efficacious mental states and their causally efficacious parts are propositions and the concepts expressed by words in our natural (i.e., public) language.

In arguing for traditionalism on the basis of the systematicity of inference, Fodor makes it quite clear that the causally relevant parts of some complex mental representations are propositions. To say that inference is systematic is to say that representations with logically similar forms are all processed in the same manner. Fodor and Pylyshyn cite a particular example: "it´s a psychological law that thoughts that P&Q tend to cause thoughts that P and thoughts that Q, all else being equal."[30] In order for this to be true, the parts that this rule must be sensitive to are the conjuncts that constitute the overall mental representation that P&Q. So, one level of reality represented by causally efficacious mental states corresponds to propositions.

The argument for traditionalism based on the systematicity of mental representation presupposes that some mental states have parts which correspond to concepts expressed by words in our public language. There are, I think, four reasons for asserting this. First, the overall structure of the argument from systematicity of mental representation is basically an argument from analogy with public language: you never find someone who is a native speaker of a language who can understand (for example) "John loves the girl" but cannot understand "The girl loves John". The ubiquity of this phenomenon is explained by the fact that the well-formed sentences

[30] "Connectionism and Cognitive Architecture: A Critical Analysis", page 46.

of a public language are not primitive, but rather are composed of elements (i.e., words) according to certain rules (i.e., the grammar). Just so, you never find someone who can think "John loves the girl" but cannot think "The girl loves John". If an analogous explanation of this phenomenon of mental representation is to work, one must assume that the parts of the mental representation correspond to the words used to express the proposition that John loves the girl.

Secondly, I cite one version of the argument from the systematicity of mental representation:

> A fast argument is that cognitive capacities must be *at least* as systematic as linguistic [public language] capacities, since the function of language is to express thought. ... You can´t have it that language expresses thought *and* that language is systematic unless you also have it that thought is as systematic as language is.[31]

By closely tying the systematicity of mental representation with the systematicity of public language, Fodor commits himself to the view that the parts that are necessary to explain the systematicity in the language of thought form a one-to-one correspondence with the parts (i.e., the words) that are necessary to explain the systematicity in public language.

A third reason for identifying the parts of mental representations with the concepts expressed by words is that, in each case, when Fodor illustrates what he means by systematicity of mental representations with a particular example, the parts of the

[31] *Psychosematics*, page 151.

mental representation correspond to the concepts expressed by the words which collectively express the proposition.

Finally, Fodor and Pylyshyn cite the example of existential introduction as an aspect of the systematicity of inference that must be explained. They write:

> We can reconstruct such truth preserving inferences as *if Rover bites then something bites* on the assumption that (a) the sentence `Rover bites´ is of the syntactic type Fa, (b) the sentence `something bites´ is of the syntactic type Ex(Fx) and (c) every formula of the first type entails a corresponding formula of the second type (where the notion `corresponding formula´ is cashed syntactically; roughly the two formulas must differ only in that the one has an existentially bound variable at the syntactic position that is occupied by a constant in the other).[32]

In order to explain this systematicity of inference, the syntactic parts of the proposition `Rover bites´ that is represented in the mind must be `Rover´ and `bites´ -- otherwise, the systematicity remains a mystery.

These considerations make it clear that, at least in the view of Fodor and Pylyshyn, traditionalism is committed to the theses that:

> (1) many mental representations are complex structures, with a combinatorial syntax and semantics,
> (2) those mental representations that are conjunctive have causally relevant parts that correspond to the individual propositions that make up the conjunction, and

[32] "Connectionism and Cognitive Architecture: A Critical Analysis", page 29.

> (3) those mental representations that correspond to individual propositions themselves have causally relevant parts that correspond to the concepts expressed by the words that form the sentence expressing the proposition.

Thus, the level of reality represented by the causally efficacious mental representations are propositions[33] and the concepts expressed by words in the public language. As noted previously, traditionalism per se is not committed to this close linkage between the language of thought and the public language. However, given the general endorsement of the LOT argument by most traditionalists, and given the lack of an alternative thesis among traditionalists as to the nature of the parts of mental representations, I shall henceforth accept these assumptions of the LOT argument as descriptive of traditionalism in general.

One point of comparison I shall use in Chapter 5 in trying to distinguish traditionalism and PDP as models of the mind is the ontological commitments inherent is each. Therefore, I end this chapter with an enumeration of the ontological commitments made within traditionalism. First and foremost, traditionalism, while based on a physicalist metaphysics, assumes that there are causally efficacious mental states. These mental states are explicitly instantiated in physical states, presumably in the physical states of the brain. Each meaningful physical state has its particular meaning by virtue of its functional role. Mental causal laws advert to the content of these states, whereby the "units" of content are the

[33] Conjunctive sentences express (complex) propositions, on my use of the word "proposition".

concepts expressed by words in public language and propositions; thus, mental causal laws quantify over states which can represent reality at the level of word-concepts and/or the level of propositions. The computability assumption inherent in traditionalism places restrictions on the form that mental causal laws can take. In particular, they must be formally specifiable. In order not to transgress the underlying physicalism, the physical states implementing the mental states must have a structure that mirrors the structure of the meaningful units of the mental state. For example, the mental state that represents "John loves the girl" has causally relevant parts corresponding to "John" and "loves" and "the" and "girl". The physical state that implements this must likewise have causally relevant parts, one of which represents "John", another "loves", another "the", and another "girl".[34] Furthermore, this physical state must encode the structure of the sentence "John loves the girl" (i.e., it must capture in a way that is causally relevant that John is the actor and the girl is the recipient of the loving relation).

[34] It is consistent with the views stated by Fodor and Pylyshyn that some groups of words function as a unit, for example, "the girl" may function as a unit, such that there is no causally relevant part of the physical state implementing this mental state that corresponds to "the" alone. All that is necessary for my above analysis to go through is that, by and large, there is a correspondence between the words in the sentence expressing the proposition and the causally relevant parts of the physical state instantiating it.

CHAPTER 4

PDP AS A MODEL OF THE MENTAL

In this chapter, I present parallel distributed processing as a model of the mind. In doing so, I must restrict myself to one version of PDP (or, more precisely, one version of PDP´s self-image). This is because researchers within this field display a wide variety of views about such basic issues as what PDP systems are understood as modelling. Perhaps because of PDP´s relative youth as a research endeavour, or perhaps because the researchers who have of late flocked to PDP represent by-and-large two distinct ways of describing intelligent activity (i.e., from the field of psychology, with its "mind-centered" approach to explaining intelligent behavior and from the field of neuroscience, with its "brain-centered" approach), the literature shows no consensus on even this fundamental question.[1] Similarly, a myriad of less-basic but still important issues regarding the "correct" understanding of PDP have yet to be resolved (or, quoting one of the more useful analogies from PDP research, the field is still in the process of settling into its stable state). Along the way, I shall hint at the variety of opinion within PDP (particularly in

[1] My choice of the neutral name "parallel distributed processing" over the more common, but also more partisan names "neural networks" or "neural network processing" is quite intentional -- the latter gives, I think, the strong impression that the entity being modelled is the brain. If the mind is simultaneously modelled, it is only coincidentally so. Also, this name gives more information on the nature of such systems than does the name connectionism.

As an historical aside, the name "PDP" derives from the title of perhaps the most influential work in the recent past of this paradigm (ie, *Parallel Distributed Processing*, Volumes 1 and 2), which, in turn, repeats the name of the research group responsible for its publication.

the first and second sections); however, I will often lapse into that mode of speech which presumes a unified view. I merely want to warn the reader that this mode does not reflect true unity of opinion.

The chapter is divided into four sections. In Section 1, I briefly describe the history of PDP from the 1940´s onwards. The purpose of doing this is to shake the reader out of the mindset that traditionalism is the only model of the mind at present. The second section provides an introduction to PDP from a "syntactic" perspective (i.e., one that describes PDP networks qua isolated, arepresentational systems). This section will serve to bring the reader up-to-speed with regard to PDP, so that a non-superficial analysis of PDP as a model of the mind can proceed in Sections 3 and 4. Any reader who is already knowledgeable about PDP may wish to skip the first and second sections, as the terminology that I adopt for later use is the literature standard. (A disclaimer: Given my purposes in describing PDP in so far as it provides a general model of the mind, I consider myself justified in overlooking many of the technical details of such systems. As even the most cursory perusal of a work dealing with the mathematical basis of PDP systems will show, a considerable amount of background knowledge -- in linear algebra, multivariate calculus and differential equations -- is necessary to understand in detail the dynamics of PDP systems. While I have dutifully read the proofs -- with greater or lesser comprehension: my academic background includes all of the above-mentioned prerequisites -- my general feeling is that such detail is unnecessary for gaining an understanding of the philosophically interesting features of PDP systems, and would, if included here, only confuse any reader without such a background.) The latter two sections correspond to Sections 1 and 2 in Chapter 3. In Section 3, I present one version of PDP qua representational system which, among the alternative versions being circulated, offers I think the best hope of providing a coherent model of the mind. Section 4 is occupied with the actual description of PDP as a model of the mind in light of my comments in Chapter 2, Section 4 vis-a-vis what it means to be a model of the mind.

4.1 History of PDP

I, along with most commentators, begin the history of PDP in the early 1940´s with the work of McCulloch and Pitts. They demonstrated that networks consisting of many simple processing units were capable of non-trivial computation. Their motivation, like that of the other early researchers in the field that would become PDP, was in understanding how the brain could implement the mind. In order to understand the import of their work, one must imagine oneself back in the 1940´s. It had been clear for centuries that, for humans, the possession of an intact brain was a necessary condition for the possession of a mind. Neuroscience at that time was far enough advanced that the gross features of the brain (as consisting of a huge number of highly-interconnected cells which passed signals amongst themselves) were well known. However, the huge conceptual gulf separating the activities of the brain from those of the mind seemed unbridgeable. (As I shall report in Section 3 of this chapter, PDP as a model of the mind, like traditionalism, downplays or ignores altogether those features and capabilities of the mind not directly relevant to information processing.) McCulloch and Pitts´ work showed that the sort of input/output processing that single neurons were capable of could, within the context of a system of many interconnected neurons, support computation. In particular, there exists, for every Turing-computable function, a system of interconnected simple processing units which can instantiate that function. Non-trivial processing, indeed.

There was, however, still a crucial piece missing in the spanning of the brain/mind gulf: McCulloch and Pitts never developed a method by which their systems could *learn*.[2] For each separate computable function to be instantiated, the system had to be designed with the correct interconnections of processing units. For all but the most trivial tasks, this is practically impossible. Neurons in functioning brains, on the other hand, display the ability to change their patterns of connectivity and to learn thereby. Presumably, this brain-learning went hand-in-hand with the learning that one could discern at the level of mind. The researcher Donald Hebb published an influential work in 1949 which provided a theory of how systems of interconnected neurons (and, not coincidentally, systems of artificial simple processing units) could

[2] I am not suggesting that a solution to the learning problem would mean that all philosophical issues on this score would also be solved. Far from it. Rather, the presumed centrality of learning in the acquisition of representational content for natural creatures requires that the mere ability to instantiate computable functions is not sufficient for the possession of intrinsic intentionality.

learn. With this piece of the puzzle in place, work could begin on developing new learning rules,[3] and running experiments on systems of interconnected simple processing units. While there were many researchers engaged in this project, the name that is most often given as representative of the work within PDP during the ensuing decade or so is Frank Rosenblatt. He was responsible for developing a learning rule for changing the pattern of connectivity between units in a restricted class of networks. It was proven that the system employing this learning rule was guaranteed to converge (with appropriate exposure to training instances) to a pattern of connectivity which solved the given problem (i.e., instantiated the desired function) if such a solution pattern of connectivity existed. Along the way, he published results showing that the type of network he used in his research was capable of instantiating (hence, capable of learning) functions corresponding to non-trivial classification tasks. (At this stage in its development, it would certainly be premature to say that Rosenblatt´s systems displayed full-blown intelligence.)

Unfortunately for Rosenblatt, the type of network that he used in his research was too simple to instantiate what seemed very basic functions. As mentioned above, the perceptron convergence theorem[4] showed that convergence was guaranteed if the network was, in fact, capable of instantiating that function. However, the

[3] Hebb´s contribution consisted, not in the discovery of a particular learning rule, but rather in the illucidation of a framework for learning in which particular learning rules could be developed.

[4] "Perceptron" was the name chosen by Rosenblatt to identify his network-type. These networks consist of a single layer of units, where each unit computes an output based on the inputs to the system.

class of functions that perceptrons could instantiate was considerably less than the set of Turing computable functions. This is because Rosenblatt´s learning rule was only applicable to a limited subset of all the possible types of networks. For example, it was known that the sorts of networks used by Rosenblatt were capable of instantiating (hence, learning) only those functions describeable as classification of vectors into linearly separable sets.[5]

The publication of Minsky's and Papert's *Perceptrons* in 1969 marks the end of the first epoch in the history of PDP. I have included it here for several reasons. First, it emphasizes the fact that a major motivation in the development of PDP has been the prospect of bridging the mind/brain gap. Traditionalism, by being so remote from neurophysiology, threatens to produce a psychology not only isolated from its implementing levels, but also (so the fear goes) irreconcilable with them. Second, the above history sets the stage for understanding the central thesis of PDP: namely, that "intelligence emerges from the interaction of large numbers of simple processing units."[6]

The recent history of PDP begins in the 1980´s with the development of a learning rule that is applicable to a more general class of networks than that studied by Rosenblatt. In particular, a (still circumscribed) class of multi-layered networks can now be effectively trained. However, it is not guaranteed that the net will finally converge to a connectivity pattern that solves the problem,

[5] For present purposes, it is not so important that the reader understand exactly what this entails. It is sufficient to note that many functions, for example, XOR (exclusive-or) fall outside of this domain.

[6] *Parallel Distributed Processing*, Vol. 1, page ix.

even if one is possible. Rather, convergence depends on a variety of factors, including the initial state (prior to training) of the network connectivity and the nature of the problem space under consideration. Current PDP research is directed at refining the learning procedure to increase the probability of converging on a solution, at fine-tuning learning parameters to speed convergence, and at developing a new class of learning rule that is more neurophysiologically plausible. (More on this below.)

As mentioned in the introduction to this chapter, there is little agreement at present on some of the most basic issues regarding PDP research. The most troublesome source of contention is also the most basic: what are PDP systems modelling, the brain or the mind (or, perhaps, some as yet unnamed level in between)? My view on this issue is clear, given my present purposes: PDP is assumed to be a model of the mind (i.e., a model of the domain that encapsulates the causal laws that quantify over contentful states). However, as previously promised, I also want to recognize the diversity within the field with regard to this question. It is often the case that, even within the context of the writings of one and the same author, this ambivalence is easily discernible. Consider, for example, the passage from the Preface to Rumelhart and McClelland´s *Parallel Distributed Processing*:

> We are cognitive psychologists and we hope, primarily, to present PDP models to the community of cognitive psychologists as alternatives to the [traditionalist] models that have dominated cognitive psychology for the past decade or so. We also, however, see ourselves as studying architectures for computation and methods for artificial intelligence.

> ... Also, the PDP approach provides a set of tools for developing models of the neurophysiological basis of human information processing ... [7]

There are unequivocal passages from this same set of authors which clearly enunciate a mind-modelling understanding of PDP, as in:

> ... [T]he operations in our models can be characterized as "neurally inspired". We wish to replace the "computer metaphor" as a model of mind with the "brain metaphor" as model of mind.[8]

and:

> We have not, by and large, focused on the kinds of constraints which arise from detailed analyses of particular circuitry and organs of the brain. Rather we have found that information concerning *brain-style* processing has itself been very provocative in our model building efforts. Thus, we have, by and large, not focused on *neural modeling* (i.e., the modeling of neurons), but rather we have focused on *neurally inspired* modeling of cognitive processing.[9]

On the other hand, there are also copious passages from the same work which describe PDP as modelling something other than the mind (presumably either the brain or a level between the brain and mind), such as:

> Parallel distributed processing models offer alternatives to serial models of the microstructure of

[7] *Parallel Distributed Processing,* page xi.

[8] *Parallel Distributed Processing,* page 75.

[9] *Parallel Distributed Processing,* page 130.

> cognition. ... What PDP models do is describe the internal structure of the larger units, just as subatomic physics describes the internal structure of the atoms that form the constituents of larger units of chemical structure.[10]

and:

> It would be wrong to view distributed representations as an *alternative* to representational schemes like semantic networks or production systems that have been found useful in cognitive psychology and artificial intelligence. It is more fruitful to view them as one way of implementing these more abstract schemes in parallel networks, but with one proviso: Distributed representations give rise to some powerful and unexpected emergent properties. The properties can therefore be taken as primitives when working in a more abstract formalism.[11]

With such diversity of views espoused within the same text, it is small wonder that the field as a whole is also not of one mind on this issue.

A further muddying of the waters results from the fact that PDP systems as currently structured are, many claim, *very* far from neurophysiological plausibility.[12] In some cases, PDP systems fail to model neural mechanisms or properties of neurons known to exist. For example, an analog of non-synaptic communication between neurons is wholly lacking, whereas it is known that such communication (implemented by the dispersal of chemicals into

[10] *Parallel Distributed Processing,* page 12.

[11] *Parallel Distributed Processing,* page 78.

[12] For a catalogue of such discrepancies, see pages 136-138 of Rumelhart and McClelland´s *Parallel Distributed Processing,* Vol. I.

diffuse regions of the brain) plays an important role in learning. Also, the finer details of neural spikes are omitted. It has lately been conjectured that such "details" are what allow for the binding of the patterns of excitation stemming from sensory stimulation from multiple modalities into a single object. PDP systems, construed as models of the brain, are also guilty of postulating mechanisms which are known not to exist. The most important among these is the hypothesized need for interneural connections which can propagate an error signal back in the direction opposite to that of the normal flow of information. All of these considerations taken together underscore the difficulty in assigning a unified objective to PDP as a field of research.

As already mentioned, the initial motivation for a PDP approach to cognition was to bridge the gap between the brain and mind. The line of reasoning suggesting PDP as a model providing a means to this end has already been hinted at: the mind is implemented in a physical medium (this is just the familiar physicalist thesis that everything that exists must ultimately be physical in nature). Empirical evidence suggests that the nervous system of a creature is a key component of its mind.[13] In scientific investigations of a domain in general, it often helps in refining the causal laws at the level of that domain if one understands the causal laws of the implementing domain. For example, the laws of chemistry constrain the set of possible laws dealing with transport

[13] This way of putting it leaves open the possibility that the extracorporeal environment of a creature may also be a part of the physical implementation of the mind. Thus, it is not ruled out that relational states of a creature play a role in mental causal laws.

of a substance across a membrane. Thus, a biologist investigating the causal laws concerning transport of glucose across the mitochondrial membrane can automatically discount many possible candidates for laws of biology which may be consistent with the phenomena when considered in isolation, but which contradict known chemical laws. Just so, knowledge of the laws of neuroscience may help constrain the set of psychological laws consistent with the psychological data. Perhaps because of the dualistic nature of much of the theorizing about the mind in which philosophers have engaged in the past (back before psychology broke off as an independent discipline), many physicalists cast doubt on the reconcilibility of the existence of a mind (with causally efficacious states) and the rest of science. In addition, the failure of folk psychological states (such as beliefs) to dovetail nicely with modern neurophysiological theories has only increased the scepticism on the part of many (both philosophers and scientists) that traditionalism (at least in its most familiar folk psychological guise) could ever be vindicated as a science. Such sceptics will only acknowledge psychology as a science when its states are shown to at least supervene on neurophysiological states plus certain physically-realized relational states. The recent rise in popularity of PDP is, I think, attributable to the widespread view that it is a more likely candidate than traditionalism to find a place in the scientific quasi-hierarchy. If one listens to the "mind-modelling" contingent among PDP researchers, this hope seems justified. The states quantified over in PDP system laws are representational;[14] hence, it is

[14] Although, as we shall see in Section 4 of this chapter, the most commonly

legitimate to say PDP systems are modelling the mind. Furthermore, at least on the surface, PDP promises to tie into neuroscience. In responding to the charge that PDP systems lack neural realism, Rumelhart and McClelland enunciate just such a construal of the aim of their research:

> [There are] two different ways in which PDP models can be related to actual neurophysiological processes, apart from the possibility that they might actually be intended to model what is known about the behavior of real neural circuitry. ... First, they might be intended as idealizations. An alternative [the one that they espouse] is that they might be intended to provide a higher level of description, but one that could be mapped on to a real neurophysiological implementation. ... Specifically with regard to the word recognition model [described previously, but not reproduced here], we do not claim that there are individual neurons that stand for visual feature, letter, and word units, or that they are connected together just as we proposed in that model. Rather, we really suppose that the various abstract informational states -- such as, for example, the state in which the perceptual system is entertaining the hypothesis that the second letter in a word is either an H or an A -- can give rise to other informational states that are contigent upon them.[15]

used interpretation scheme among PDP researchers has it that units alone do not represent, but rather take part in patterns of activation over many units which collectively have representational content.

[15] *Parallel Distributed Processing*, page 138.

4.2 A Syntactic Description of PDP

In this section, I shall describe PDP on the assumption that the reader knows nothing about such systems. I begin at the level of the unit: what sorts if functions can it instantiate? How is it typically connected (via its inputs and output) with other units? Following this is a description of network behavior, including a discussion of various learning rules and their convergence characteristics, and a very cursory examination of the mathematical basis of PDP systems. I then run through a simple example showing how such a network behaves, and end with a "syntactic" description of the sorts of tasks that PDP systems can perform. A "semantic" account of PDP is the topic of Section 3.

The building block of PDP systems is the unit.[16] A single unit (depicted in Figure 4) is, abstractly considered, a function over numbers.

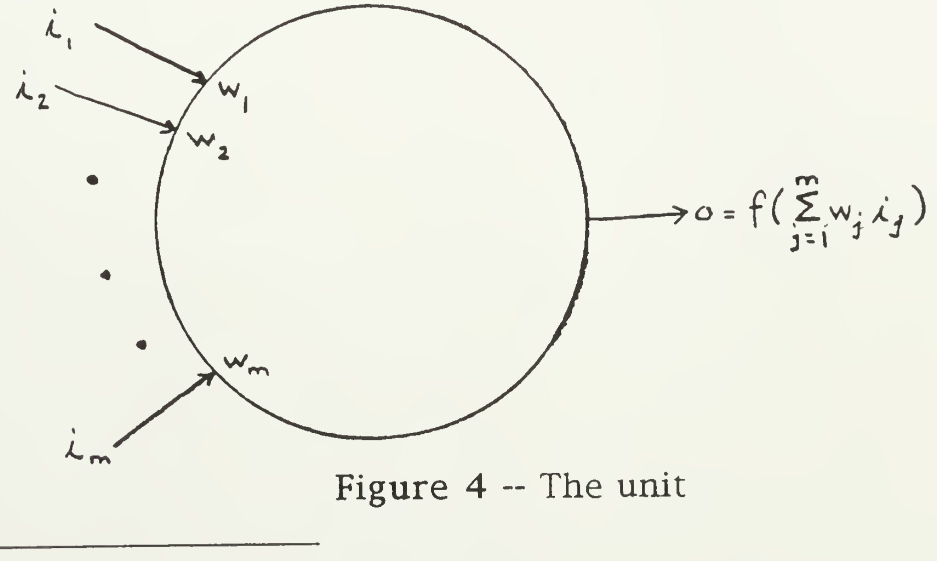
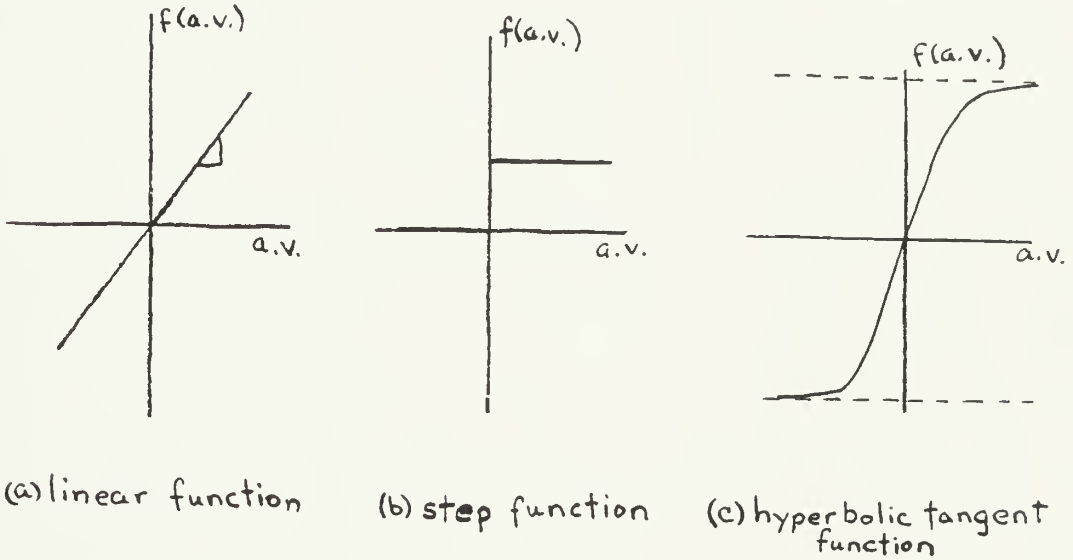


Figure 4 -- The unit

[16] Because of PDP´s history as emerging from neuroscience, one often sees units referred to as "neurons". Similarly, connections (to be discussed shortly) are sometime called "synapses". As with my choice of the more neutral name "PDP", I have continued, whenever possible, to choose names least likely to reinforce the "PDP as neural model" view.

In particular, it maps the sum of the *(input x weight)* products to another number. While the notion of an input is intuitively clear, the meaning of "weight" may not be. In its neuron-modelling guise, the weight corresponds to a measure of the synaptic efficacy: ie, how easily a particular pre-synaptically produced signal is passed onto this post-synaptic unit. Using neural-neutral terminology, the weight, wj, corresponds to the strength of the connection between whatever produced ij (call it "producer-of-j") and this unit. If producer-of-j is strongly connected to this unit, then the magnitude of wj will be relatively large. If producer-of-j is only weakly connected, then the magnitude of wj will be relatively small. If producer-of-j exerts an inhibitory influence on this unit (i.e., a large ij makes it less likely that this unit will produce a large output), then wj will be negative. In general, weights that are positive are called "excitatory", and weights that are negative, "inhibitory". A weight of zero signifies that producer-of-j exerts no influence on the unit; it is as if producer-of-j and this unit were not connected to one another at all.

There exists, within this framework, great variety on several points. First, some PDP systems employ units which receive as input only a subset of the integers (for example, an input line may take on the value of 0 or 1). Some PDP systems, on the other hand, allow the inputs to range over all of the real numbers, or perhaps some subset of the reals (for example, values between 0.0 and 1.0). There is also a wide range of possibilities with respect to the weights. They may be restricted to the integers or a subset thereof, or they may take on real values. The sum Swjij is called the activation value of the unit, and it may be likewise restricted in the values it can assume, as per above. The output of the unit is a function, *f*, of the activation value of that unit. Three of the most common choices for *f* are shown in Figure 5.



(a) linear function (b) step function (c) hyperbolic tangent function

Figure 5 -- Types of unit output functions

Alternative (a) depicts a linear function: the output is either identical to the activation value (in which case the slope is 1), or is a multiple of the activation value. Alternative (b) shows a typical step function: the output is 0 until a certain threshold is reached, at which point it changes to 1 for that and all greater activation values. While I have, for simplicity´s sake, shown the step function as changing from 0 to 1 at an activation value of 0, this is only one of many possible step functions. Another step function may be -1 for all activation values less than 5, and +1 for all activation values of 5 or greater. Alternative (c), a squashing function, is seen quite often in PDP systems. It has several advantages over a simple step function[17] and over a linear function.[18] (While I have drawn the functions (a), (b), and (c) as having real-valued domains, and for (a), an unrestricted real-valued range, this is not universally the case. As already mentioned, often the possible values of the i´s and w´s restrict the corresponding domain (and range) to some subset of the reals or integers.)

Figure 6 depicts a 3-layered feed-forward network consisting of 9 units.

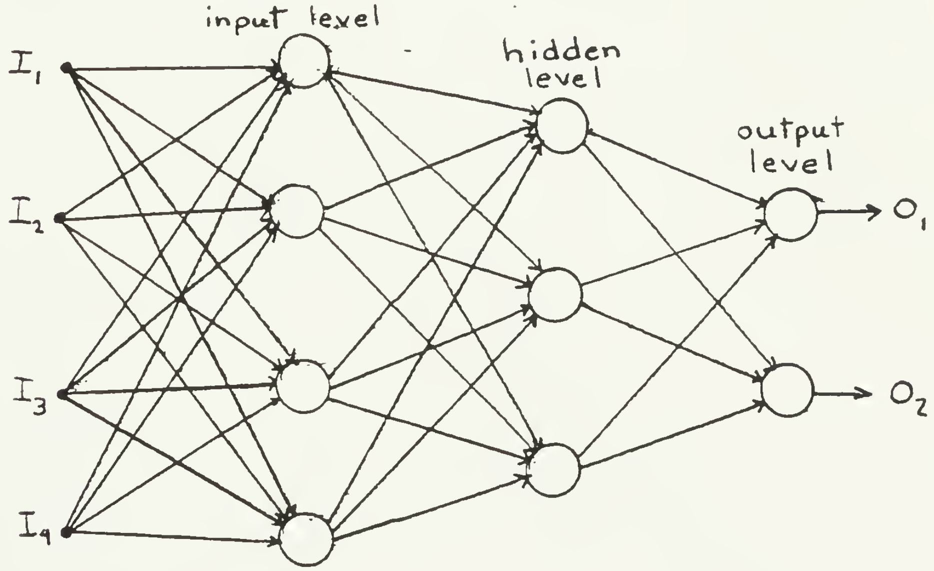


Figure 6 -- 3-layered feed-forward network

[17] A major advantage of the squashing function within the context of multi-layered nets capable of learning is that, because it is continuously differentiable, the most general learning algorithm yet developed for such nets (back-propagation of the error signal in a direction opposite that of forward information flow) is applicable to a network consisting of such units.

[18] In general, a net consisting of units using linear functions has a more limited capacity than one consisting of either of the two depicted non-linear functions.

The inputs to the overall system (I1 ... I4) are supplied by the "environment" -- they could be the outputs from some other network(s), or they could be some signal coming from the environment, as normally understood. Each of the 4 input-level units receives each input signal, and produces an output, as described above. The output from the input-level units forms the input for the hidden-level units. (They are called "hidden" because they are not directly connected to the environment, either via their input or their output.) Similarly, the output-level units receive as input the output from the hidden-level units. The overall output of the system is the output of the units on this level. There are several things to note about this particular PDP system. It is a feed-forward network. That means that information flows only in one direction. Had it been the case that, for example, a hidden unit´s output supplied the input to a unit on the input level (thus producing a loop), the net would no longer be feed-forward. Also, had the net allowed for an output signal from one level to loop back and form the input either for that unit itself or any other unit on the same level, it would no longer be feed-forward. (Non-feed-forward nets are also known as "recurrent nets".) Whether a feed-forward net or a recurrent net is the appropriate choice depends on the task to be performed by the net. (More on this later.) Note also that each unit sends its output to all and only the units on the next lower level. This condition on feed-forward networks, if satisfied, simplifies the analysis of the network behavior. There is, however, no principled reason why a feed-forward net must be thus fully-connected. Most often, all of the units within a network are identical with respect to their allowable domain and range, and the function, *f*, which maps the activation level to an output. As with the full-connectivity condition, there is no principled reason for this: such a condition merely simplifies the mathematical analysis of the network behavior.

Thus far, my description of PDP networks has been as static systems. The manner in which the temporal aspects of processing is modelled varies. For the feed-forward case, one can simply assume that the inputs to the system do not change, and that each unit continuously computes its output function, so that, once the inputs to a unit cease changing, its output remains constant. Thus, the overall output of the system eventually achieves a constant value. For recurrent nets in general, however, such a constant output condition cannot be guaranteed: the dynamics of some networks are such that the overall output never reaches a stable value, even though the system´s inputs remain constant. It is useful here to view the network, not just as a monolithic structure from inputs to outputs, but as a system constructed from individual units. This system-as-units level of description will allow us to consider the overall state of the system as the complex object consisting of the outputs (or, in some cases, the activation values) of each unit. (An identical perspective for viewing network behavior of recurrent nets is that whereby the output of each unit is a part of the overall system output.) It is known that, in the general case, such systems may never settle into a stable configuration. However, for a subset of the class of recurrent nets,[19] such stability is guaranteed.

To get a feel for the dynamics of general recurrent nets, it may be useful to consider some of the properties of the behavior of this limited class of recurrent nets. Each possible state of the net with p units is a p-dimensional vector,[20] where each of the p items corresponds to the output of one of the units. (It is sometimes more useful to consider, not the output of each unit, but their activation values, in describing the overall network state.) A simple way to understand the behavior of such nets is with a 3-D space analogy. Imagine a topographical map depicting the contours of some wholly self-contained group of idealized watersheds:[21] every drop of precipitation that falls in the watershed ends up in a body of water with no outflowing stream. Each body of water in this watershed is a local minimum with respect to elevation, and corresponds to the stable state for each drop of water that falls within the watershed for this body of water. There may be many such bodies of water depicted by the topographical map.

The behavior of an individual water drop is analogous to the behavior through time of a recurrent net with the above-

[19] For example, networks with (1) all units using the step function shown in figure 5b, and (2) networks such that the weight from unit-n to unit-m is equal to the weight from unit-m to unit-n, for every unit-n and -m) will always settle into a stable configuration when the system inputs are held constant.

[20] A vector is a mathematical object with p "slots" for numbers, whereby the particular ordering of the slots is encoded. So, for example, <1.2, -5.3> is a 2-dimensional vector whose first item is 1.2 and whose second is -5.3. This vector is distinct from both <-5.3, 1.2> and <1.2, -5.3, 0>

[21] Idealized to the extent that raindrops do not soak into the ground, but rather roll on the surface under the influence of gravity and friction, as they minimize their energy level.

enumerated properties in the following way. Each local minimum in elevation (i.e., each body of water) acts as an attractor for all the drops of water landing in its watershed. No matter where they land within that watershed, they end up at the same local minimum. Just so, the space of possible network states can be divided up into mutually exclusive and collectively exhaustive sets, each of which has a "local minimum" or attractor state associated with it. As with the journey of each water drop, the evolution of the network state may pass through many (non-stable) states on its way to its stable attractor. Obviously, the analogy fails in many places. For example, the network configuration is not transversing 3-D space, but a p-dimensional space of unit outputs. Also, the initial state of a water drop with respect to its position within the group of watersheds determines which local minimum it will settle into; whereas, for the network, it is the input vector (and, depending on the update rule used, perhaps also the initial configuration of the net) which determine the attractor-state settled into. Furthermore, not all recurrent nets display this stability. (In particular, recurrent nets with non-symmetric weights often do not.) Even with these sources of disanalogy, I think that the group of watersheds picture is a useful one to keep in mind when trying to understand the dynamics of PDP networks.

Thus far, I have said nothing about how such systems could learn. If a network produces (or settles into) the wrong output (as judged by an external observer -- the researcher, perhaps), how is it possible to change the network so that, the next time it is presented with that input, it produces the correct output (or, at least, one "closer" to the correct output than its previous one, in some as yet undefined sense)? Hebb´s work in the 1940´s suggested that learning could occur by means of the changing of the units´ weights according to certain rules. There are two broad paradigms of learning within PDP: supervised learning and unsupervised learning. In supervised learning, some external entity must be available to compare the produced output with the correct or expected output for that input, and provide the network with information, so that, if the produced output is wrong, it can change its weights so as to increase the chances of producing the correct output on the next occasion that that input is given. Supervised learning further subdivides into two subtypes corresponding to how much information is provided by the external entity. In learning with a teacher, the net is supplied with the correct output. In learning with a critic (also called "reinforcement learning") the supervisor gives the net less information: it either informs the net as to whether the produced output was correct or incorrect, or informs the net as to the degree of wrongness of the produced output. (I shall return to the topic of supervised learning shortly, and discuss in general terms how such learning proceeds.) The second learning paradigm is unsupervised learning.[22] This style of learning is appropriate when the network is to learn, not a fixed input/output relation, but rather the regularities in the set of input items that allow them to be effectively categorized. An important aspect (particularly in light of the use made of PDP as providing a model of the mind) of most PDP learning rules is that generalization

[22] Neurobiologists often refer to this form of learning as "Hebbian learning".

is automatic. The network not only improves its performance in producing the correct output for a given input as learning progresses, but also improves in its ability to make "reasonable" generalizations with respect to the correct output corresponding to an input on which it has not yet been trained. (Obviously, "reasonable" in this context needs some serious explication, which I take up in Section 3.) Generalization is automatic in the sense that no additional learning rules need be used above and beyond those associated with general learning.

I shall now look in more detail at learning in PDP systems. I begin with a consideration of Rosenblatt´s perceptron convergence procedure, which gives the learning rule for a feed-forward uni-level network of units within the supervised learning paradigm in which a teacher is available.

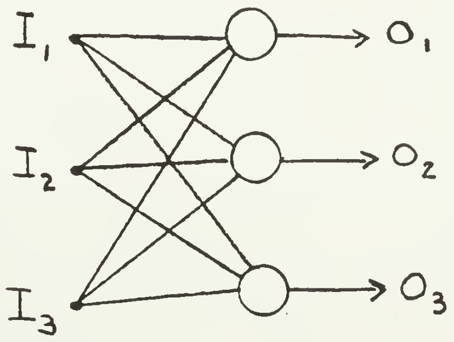


Figure 7 -- 3 unit network capable of learning

Suppose the initial weights connecting the three inputs to each of the three units are set to small, random numbers. The training sequence is begun as an input vector is supplied via the input lines. The supervisor then looks at O1, O2, and O3 to see if they are the output that should be produced for this input. Suppose that they are not; hence, the weights must somehow be adjusted. To be more specific, suppose that the input is (1,1,1) and that the correct output is (1,0,0), and the output given by the net is (0,0,1) -- ie, O2 produced a zero, as it should have; however, O1 and O3 both produced the wrong value. Described qualitatively, we want to leave the weights connecting the inputs and the middle unit unchanged, but change the weights connecting the inputs to the first and third unit. The first unit produced a zero when it should have produced a 1, so we need to increase the weights connecting the inputs to this unit. For the third unit, the produced output was 1 when it should have been 0; therefore, the weights connecting it and the inputs should be decreased.[23] After adjusting the weights as per above, the network is presented with a new input (for example, (1,0,0)), and once again the supervisor checks to see whether the output produced is correct or not. If the latter, then the weights are adjusted again. The attentive reader has probably already noticed that it is possible that the readjustments made on this second training pass may interfere with the learning that occurred as the weights were changed after the first learning pass. Obviously, then, we need to adjust weights in such a way as to guarantee that, if enough training passes are made,

[23] The need for an increase or a decrease in a weight is, of course, relative to the function, *f*, from activation value to output, as well as to the absolute value and sign of the input associated with that weight. I chose the input (1,1,1) to avoid having to specifically mention this relativity in the above discussion. However, the reader should keep it in mind. In addition, this fact is reflected in the perceptron learning procedure, described below.

the network weights will eventually converge to a set that produces the correct output for every input.[24] Otherwise, we may keep supplying inputs and adjusting weights, only to have the next weight adjustment wipe out the previous learning. The perceptron learning procedure tells us by how much the weights connecting the inputs to a unit must be adjusted, so as to guarantee eventual convergence. In particular, the weights should be adjusted in accordance with the following equation:

$$\text{delta-wj} = \text{eta} \times (E - O) \times j \qquad \text{(Eq. 4.1)}$$

where "delta-wj" indicates the required *change* in the weight connecting producer-of-j to this unit. j is the value on the input line from producer-of-j, O is the actual output produced, E is the expected (or correct) output, and eta is the learning rate (a positive number that influences the convergence properties of the unit -- eg, how fast it converges).[25] The perceptron convergence theorem states that, if the above learning rule is consistently followed on a sufficiently large set of training examples, then the network weights will converge to a set that solves the problem.

As the capacity of such uni-layered networks to instantiate different functions is very limited, it is necessary to have a learning procedure which can be applied to multi-layer nets. For this case,

[24] This, of course, assumes that the problem is solveable by this network in the first place. As mentioned above, many problems are not.

[25] Strictly speaking, the above learning rule is a generalization of Rosenblatt´s perceptron learning rule, for which the convergence theorem is valid. However, since we won´t be looking at the convergence proof, this version suffices.

however, a difficulty not encountered with single layer nets arises: namely, how much should the weight connecting one unit to another be changed, given that there is more than one unit separating the network input from the network output. In the case of a 2-layer network (no hidden units), we cannot in general know whether an incorrect output was given because the weights connecting the output layer and the input layer were incorrect and/or because the weights connecting the input layer and the inputs were incorrect. This is called the "credit assignment problem", and its recalcitrance until the development of the back-propagation learning algorithm in the 1980´s meant that, at least with respect to functional capacity, PDP remained stagnant during the 60´s and 70´s. As with my description of the perceptron learning procedure above, I shall begin with a fairly high-level view of supervised learning with a teacher in multi-layered networks, and save the mathematical particulars until later. I then take up the topics of supervised learning with a critic and unsupervised learning in multi-layered networks.

Suppose that the network depicted in Figure 6 has weights initialized to small random numbers, and that the function computed by each unit is the squashing function with output between -1 and +1, as shown in Figure 5c. In order to accommodate the back-propagation of an error signal, we must embellish the network with a communication line (one for each normal forward line) that passes information in the direction opposite to that of the forward line. So, for each forward connection from unit-n to unit-m, there must also be a "learning line connection" from unit-m to unit-n. Note that this learning line connection plays a role only in learning, not in the processing that ensues between introduction of the system input to the input level units and production of the system output by the output units.[26] Training begins as the network is presented with a new input. The supervisor then checks the output after a certain period of time.[27] If the output is incorrect, the weights (those connecting the hidden level units to the output level units, those connecting the input level units to the hidden level units, and those connecting the system inputs to the input level units) need to be changed. The supervisor tells each output level unit the output that it should have produced. The weights connecting the hidden level units to the output level units are then adjusted as determind by the back-prop learning rule, described below. Then, an error signal is passed back from the output level units to the hidden level units, and the same weight-adjustment procedure is repeated on the weights connecting the input level units to the hidden level units. (For the general case of a network with multiple hidden levels, this procedure can be repeated indefinitely.) Finally, an error signal is passed back from the hidden level units to the input level units, and the weights connecting the system input lines to the input level units are adjusted. After this whole procedure is completed, the

[26] One of the major objections against PDP qua neurally plausible framework is its frequent use of a learning technique with a need for this sort of backward flowing information line. In particular, neurobiologists claim that no such means for back-propagating an error signal exists between real neurons. This has led some in the PDP research community to reject the back-prop technique for learning in favor of a more neurally plausible approach. More on this below.

[27] Given that this is a feed-forward net, we know that the output will be stable. However, as the link from input to output involves passing successively through three units, we must wait long enough for the signal changes to trickle down to the output level.

process starts again as the next input vector is supplied to the network. The actual back-prop learning rule used is as follows:

delta-wjk = eta x delj x Ok (Eq. 4.2)

where delta-wjk is the change of weight from the kth to the jth unit, eta is again the learning rate, Ok is the actual output of unit k (received as input to unit j via this connection), and delj is the back-propagated error signal. For the case where unit j is on the output level, delj is simply the difference between the expected and actual output for that unit (i.e., Ej - Oj) times the derivative of the unit´s function, *f*, at unit j´s activation value. (Recall that *f* in such multi-layered networks using this version of back-prop must be a continuously differentiable function; hence, the popularity of the squashing function (Figure 5c).) Once the changes in weights for the output level units are made, the delj for the hidden layer is computed using the equation:

delj(non-output-level) = f´j(av-of-j) x S(over-l) dellxwlj (Eq. 4.3)

where f´j is again the derivative of fj, and l in the sum ranges over all of the units to which unit j sends its output. The above equation is then used to compute the del´s for the next left-most layer, until finally the input level units´ weights have been adjusted.

Perhaps even this level of mathematical detail in the preceding exposition is greater than strictly necessary for understanding PDP as a model of the mind. I have included it because I want to make explicit a feature of this form of learning that will become important in later discussions (particularly in Section 4 of this chapter). Notice that no unit has any information about the global state of the network. During the forward pass, each unit (except for the input level units) receives information of the state of the other units only from its immediate (backward-facing) neighbors in the form of the output from those neighbors. Similarly, no unit gives its output to any other than its immediate (forward-facing) neighbors. Even considering the backward flow of information that occurs in the back-prop technique, each unit receives information only on the state of its (forward-facing) neighbors. This so-called locality constraint on PDP systems will turn out to be relevant to a later discussion of the relationship between mental causal laws and the contentful states quantified over by them.

An analogy similar to the group of watersheds analogy for describing the behavior of recurrent nets as they settle into a stable state for a particular input is useful here. In this case, however, the dynamics to be described involves how the network´s weight-state changes as learning progresses. (In the previous example, the dynamics being mirrored in the analogy were the network´s output-of-each-unit state (or the activation-value-of-each-unit state).) As before, one can picture the weight-state space of a network as a group of mutually exclusive and collectively exhaustive, idealized watersheds. The x-y plane (perpendicular to the axis measuring elevation) corresponds to the location of the network in weight space, and the elevation corresponds to the "degree of wrongness" of a particular system output vector for the given input vector. The aim of learning is to get to a place of zero elevation, where there is no difference between the correct system output and the produced system output. Since each input vector has a distinct topographical map corresponding to it, the overall goal is to find a weight state that has zero elevation in each topographical map. A good rule-of-thumb in situations in which you want to decrease your elevation from its current (non-zero) value to a value of zero is to head downhill, taking as your preferred route that with the steepest incline. Learning in PDP systems does exactly that. It uses a technique called "gradient descent" that involves calculating the direction of steepest incline at a location, and changes the system variables (in this case, the weights of the network) so as to move the system in this direction of steepest descent. After each training pass, the weights are updated and a new input is supplied. If the elevation (i.e., system error) is non-zero, the direction of steepest incline is again calculated and the weights are again adjusted to move the system in this direction. (Just to reiterate, this is only an analogy. In particular, there is no one or no unit calculating the gradient, etc. It is merely a property of the (strictly local) back-prop procedure that it performs gradient descent in the weight-state space.)

Unfortunately, gradient descent has some drawbacks as a universal learning technique. Foremost among these is that it cannot guarantee convergence of the network weights to a set that produces the correct output vector for each input vector, even if the network is in theory capable of instantiating this I/O relation. (This is in distinction to the perceptron learning procedure, for which convergence is guaranteed, so long as the network is in theory capable of instantiating the I/O relation.) The problem can be easily visualized using the topographical analogy. Along with the (presumed) global minimum with zero elevation, there may be one or more wholly enclosed watersheds with a local minimum greater than zero. (Alpine lakes, perhaps.) If the initial weight-state of the network puts it in the equivalent of the watershed for an Alpine lake, performing gradient descent will change its weight vector in the direction of the local minimum. It is possible that the true global minimum is in the opposite direction, so that "learning" actually pulls the system in a direction away from a solution.

For any reader having difficulties visualizing this, I have a real-life story that captures the "problem of local minima" perfectly. I wanted to hike up Mt. Toby, but lacked a map. A friend offerred the following bit of advice: "Well, Mt. Toby is the tallest mountain in this part of Massachusetts, so if you consistently walk uphill, you´re bound to find it." This friend failed to take into account something that I only found out after several hours of walking: while Mt. Toby is indeed the tallest mountain in the area, there is a little mountain (one of the peaks of Bull Hill) in between my location at that time and Mt. Toby. Thus, a local maximum separated me from my ultimate goal. Following the friend´s advice allowed me to find, not the desired, global maximum, but this undesired, local one. Just so, gradient descent aims for whatever local minimum is in the vicinity, irrespective of whether it also constitutes a global minimum.

There are several other shortcomings of gradient descent. For example, if the space being searched is describable as steeply sloped, with very small valleys, gradient descent tends to "overshoot" the minima. Similarly, using gradient descent in a space that is nearly flat results in very slow convergence. One thrust of current research within PDP is to refine the search technique (as implemented by the back-prop procedure) to mitigate some of these shortcomings. A discussion of the particulars of these efforts would take us outside the scope of this work, however. I should at least note in passing that back-prop learning procedures for recurrent multi-layer networks are available, although, as with back-prop for feed-forward nets, convergence of the weights to a solution set is not guaranteed.

As promised, I shall now illustrate network behavior with a concrete example, both during the forward flow as the system outputs are being computed, as well as during the back-propagation of the error during the weight-update phase of the learning cycle. In the example, I use a 2-layered feed-forward network with 2 input lines as depicted in Figure 8. Each of the three units uses a squashing function as its *f*, also shown in Figure 8.[28]

[28] Squashing functions, even those passing through the origin and asymptotically approaching -1 in the direction of decreasing activation value and +1 in the direction of increasing activation value, can be distinguished based on how quickly they approach their asymptotes. The squashing function depicted is much less steep at the origin than what one sees in the typical case. However, for the purpose of illustration, it is preferable.

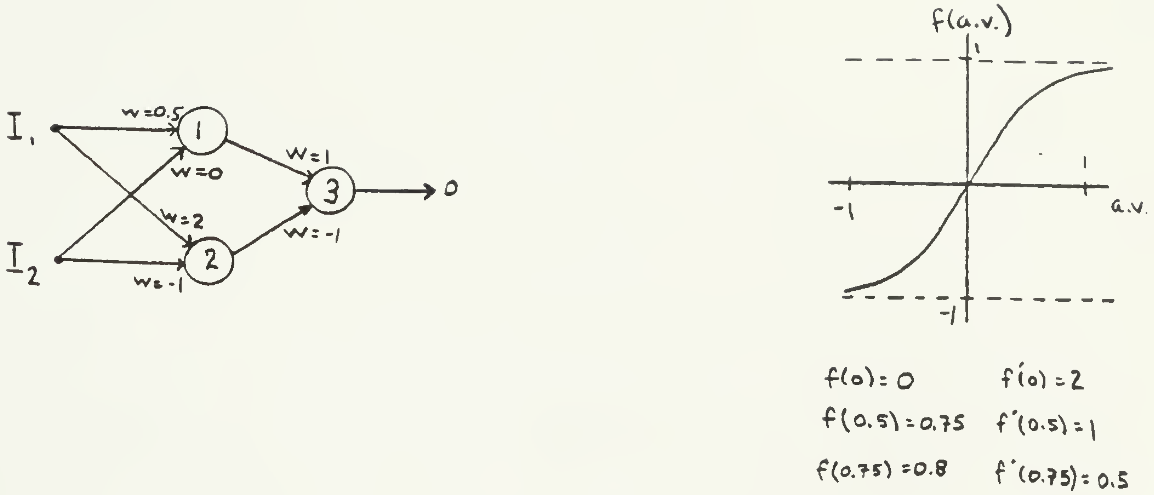


Figure 8 -- Learning example in 3-unit network

We shall use a learning rate of 0.1. The initial weights are as shown. Training begins with the vector (1,2). The activation values and outputs for the units are:

av-1 = (0.5)(1)+(0)(2) = 0.5 :: f(0.5) = 0.75 (Eq. 4.4a)
av-2 = (2)(1)+(-1)(2) = 0 :: f(0) = 0 (Eq. 4.4b)
av-3 = (1)(0.75)+(-1)(0) = 0.75 :: f(0.75) = 0.8 (Eq. 4.4c)

The overall output of the network is thus 0.8. Suppose that the correct output for this input vector is not 0.8, but -1.2. So, the network weights need to be changed. Computing the delta-w´s for the 3 units yields:

delta-w3,1 = 0.1(-1.2-0.8)(0.5)(0.75) = -0.075 (Eq. 4.5a)
delta-w3,2 = 0.1(-1.2-0.8)(0.5)(0) = 0 (Eq. 4.5b)

delta-w2,I1 = 0.1(2)(-1)(-1)(1) = 0.2 (Eq. 4.5c)
delta-w2,I2 = 0.1(2)(-1)(-1)(2) = 0.4 (Eq. 4.5d)
delta-w1,I1 = 0.1(1)(-1)(1)(1) = -0.1 (Eq. 4.5e)
delta-w1,I2 = 0.1(1)(-1)(1)(2) = -0.2 (Eq. 4.5f)

Therefore, the new network weights are as shown in Figure 9.

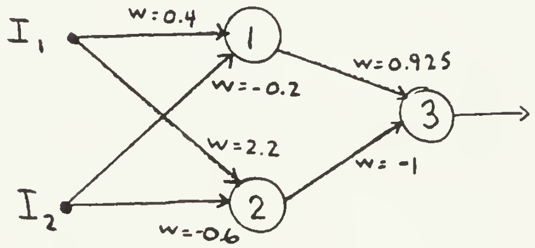


Figure 9 -- Network, post-learning

(If we were to now re-try the input vector (1,2) on this network, the system output would be -1.0; while still not correct, it is at least closer.) Normally, training on such networks proceeds by presenting the system with an input, cycling throught the learning phase, then presenting the network with a new input. Usually, one must cycle through the training sets many times in order for the network to converge to a weight vector that yields the correct output for each input. I shall, however, end the example here, with the hope that the reader has gained at least an intuitive feel for processing in PDP systems using supervised learning with a teacher.

As already mentioned, there is another form of supervised learning in multi-layered nets (so-called "learning with a critic") which has received attention in PDP research. In this case, the supervisor checks to see if the produced output is correct, and, if not, either informs the system of this fact (without, however, supplying the system with the correct output), or supplies the system with a measure of the wrongness of the produced output. Learning with a critic is also referred to as reinforcement learning to stress both its gross characteristics and its greater plausibility as the form of learning most often used by creatures operating in the real world. (Neither I, nor anyone else within the scope of my reading, would hazard to guess what percentage of learning in humans is described as learning with a teacher, versus learning with a critic. Clearly, both occur. Perhaps then, a complete model of the mind -- and its concomitant model of learning -- must include the capability for supervised learning of both types.) As one might guess, convergence of such systems is much slower than for learning with a teacher (if, indeed, the system converges at all), as less information is available to the network to aid in changing weights efficiently.

An interesting approach to reinforcement learning is the models and critics approach, in which the PDP system consists of two distinct sub-networks, one of which forms a model of the reinforcement signal[29] (I shall call this the "model" sub-network), and the other of which performs learning with a teacher as described above, but with the feature that the "correct output" is supplied, not by the supervisor, but rather by the model of the reinforcement signal. (I henceforth call this the "being-taught" sub-network.) The first learning stage in the models and critics approach

[29] In this version of learning with a critic, the net receives a measure of the closeness of its output to the correct one, rather than a mere "correct"/"incorrect" signal.

to reinforcement learning proceeds as follows. An input vector is presented to the "being-taught" sub-network and to the "model" sub-network. The former produces an output. The output is then passed both to the supervisor for reinforcement information as well as to the "model" sub-network. The "model" sub-network then produces its estimate of the reinforcement signal for the given output and system input, and compares it with the actual reinforcement signal supplied by the supervisor. (Thus, the "input" to the "model" sub-network consists of both the input to the system and the output of the "being-taught" sub-network.) Back-propagating the error signal (i.e., the difference between the produced and actual reinforcement signal) back through the "model" sub-network improves the ability of this network to predict the reinforcement signal corresponding to an I/O pair. The weights in the "being-taught" sub-network can be changed at random, since, in this first phase, it is the "model" sub-network that is being trained, not the "being-taught" sub-network. Once the "model" sub-network´s estimates of the reinforcement signal are accurate enough, the second stage of learning begins. Again, an input is presented to both the "being-taught" and "model" sub-networks. The former produces an output, which is sent as before to the supervisor and to the "model" sub-network, which, in turn, produces an estimate of the reinforcement signal. Now, something different from Stage1 learning occurs. An error signal equal to (0 - maximum-reinforcement-signal) is back-propagated through the "model" sub-network, but without the usual updating of weights. Rather, this back-prop is performed with the purpose of producing a guess as to what the output that the "being-taught" sub-network gave should have been. This guess is then used to perform learning a la supervised learning with a teacher on the "being-taught" sub-network. Stage2 learning continues as additional inputs are given successively, and the whole process is repeated.

Unlike the two types of supervised learning described above, in the unsupervised learning paradigm, there is no signal from an external source indicating whether the produced output is correct or not. This mode of learning is therefore not appropriate in cases where a particular I/O relationship is to be learned. Rather, it is used when regularities within a set of input vectors must be identified, so that future input vectors can be classified as belonging to one of the discovered classes, each of which corresponds to a regularity type. This can be implemented in one of two net types. In the first type (called "winner-take-all"), the desired result is to train a network to classify inputs as belonging to one of several mutually exclusive and collectively exhaustive types. The name derives from the fact that the membership is indicated by the production of a 1 on the output line corresponding to the input pattern´s type, and a 0 on all other output lines. The applications of this type of learning within cognitive processing are ubiquitous, particularly in the area of perception, where, for example, a particular input visual vector needs to be classified as an instance of a particular object-type (for example, as a human face). The second type of unsupervised learning involves discovering regularities in the input data, so that future input vectors can be classified in terms of their similarity (with respect to the discovered regularities) to the inputs presented during training. Within the sphere of theories of cognitive-level classification, this is a way in which a fuzzy categorization scheme could be implemented. One particular subtype of fuzzy categorization is Wittgenstein´s "family resemblance" theory, whereby the presence or absence of certain features makes the object more or less exemplary of a given type. Various learning rules have been developed for training networks within the unsupervised paradigm. However, given the fact that I have already gone into some detail in describing learning in the supervised paradigm and the fact that the basic principles remain the same, I shall omit further discussion of learning rules within this paradigm. (I shall, however, pause to reemphasize that no external teacher or critic is used or needed in unsupervised learning; hence, there is no error signal to back-propagate. Instead, learning proceeds by changing weights based only on the state of the unit and the inputs received from and weights associated with its backward-facing neighbors. This has led many within the brain-modelling camp of PDP to adopt this paradigm of learning, as it does not require the neurobiologically implausible passing of information against the normal forward flow.)

I would like to end this section with an overview of four important classes of tasks that PDP systems can (be taught to) do.[30] In the first class, called auto-association, the network is presented with a set of input vectors during the learning phase. The task to be performed involves re-producing (at the output) the input vector

[30] The following is based on the discussion in Rumelhart and McClelland´s *Parallel Distributed Processing,* Vol. 1, pages 159-161.

most closely resembling the given one. The ability to do this is very useful when a network is operating in an environment in which the input is noisy (i.e., has occasionally spurious values on an input line or lines) or in which the input is sometimes incomplete. Some possible application areas include content-addressable memory and functioning as a front-end to some other network within a noisy environment.

The second class of task is similar to the auto-associator, except that, rather than the input vector itself, some other vector paired with that input during training is to be re-produced. Learning consists of repeated presentation of the sets of two patterns to be associated, so the number of input lines during learning must equal the sum of the dimensionality of the vectors to be associated. After learning, presentation of the first of any of the now-associated vector pairs should result in the production at the output of the other. The most obvious domain of applicability is one in which a network is to guide action, in such a way that one thing is to be done after another. For example, the learning of skilled motor behavior involves the learning of complex sequences of individual movements, all concatenated together. Such a sequence of vectors (each corresponding to a single movement) can be associated, such that the initial movement starts a cascade that produces each of the others in turn. With appropriate feedback connections, a network can learn a sequence consisting of many individual vectors. Obviously, the time of production of each item in the sequence may have certain constraints in order for the sequence as a whole to achieve a necessary level of fluidity (as, for example, when a skilled pianist performs an arpeggio with a particular tempo). Indeed, for some tasks, time is crucial not only to fluidity, but also to success. If I tried to run by producing each of the individual muscle contractions and relaxations associated with the running gait, but I produced each item in the sequence 1 second apart rather than the (more appropriate) 1 msec apart, I would likely topple over. Even this time constraint can be built into the system, if the delay characteristics of each unit are known.

A third task is that of classification. Here the network is trained to classify a set of input patterns, so that future presentation of either a wholly novel input vector or one slightly distorted from a previously encountered input vector results in correct classification. Within this task-type, it is assumed that there exists some predetermined classification scheme, so that the initial learning period consists of supervised learning on the training set.

A fourth task that PDP systems can learn to perform is regularity detection. This typically occurs in the unsupervised learning paradigm, and involves the extraction and encoding within the network of regularities within the training set, so that future novel inputs can be classified by means of the learned regularities.

As an aside, it is interesting to note that the PDP literature describing experiments run involving the third and fourth task-types often have a common feature: surprise on the part of the researcher with regard to the regularities in the data seized upon by the network to accomplish the task. What often happens is that the researcher examines the post-training network only to discover that it has uncovered syntactic regularities in the training data not previously noticed by the researcher. In particularly complex networks, it is sometimes even the case that the researcher cannot figure out how the network is performing the task, although its high level of performance after training demonstrates that it has isolated regularities relevant to the overall problem to be solved. This issue crops up again when I examine generalization in PDP.

4.3 PDP as Currently Practiced

The "mind-modelling" contingent among PDP researchers hold several key assumptions in common with their traditionalist counterparts. One of these is that mental activity is a certain kind of processing. In this regard, the views of hard-core computationalists like Pylyshyn (quoted in Chapter 3) are also applicable to PDP theorists: the mind is the instantiation of a particular process, whereby not only the I/O behavior, but also the means by which the I/O behavior is brought about, is important. To be a mind is to instantiate the mental process. The two camps part company (or, at least, appear to -- in a sense, this entire dissertation is concerned with figuring out whether they do indeed part company) in their respective further elaborations of the details of this mental process. Thus, PDP, like traditionalism, is committed to the explanatory, rather than the merely simulating, nature of their model. A second consequence of this view is the subordinate status within the theory of the mind given to mental phenomena such as consciousness. If consciousness is a by-product of mental processsing, then it may accompany an instantiated mental process; the property of being conscious is, however, neither a necessary nor a sufficient condition for being a mind.

Another key assumption held in common by both PDP and traditionalism is the contentfulness of mental states. In the previous section, I confined myself to a description of PDP as arepresentational. In depicting net behavior, the vocabulary used was that of activation values and connection weights -- terms that make no reference to anything outside of the network. Thus described, PDP is not very interesting for a philosopher of mind. This section, on the other hand, will deal with PDP systems qua representational systems.

I begin this task, as usual, with a survey of quotations, showing that my interpretation on this score is, if not universally consented to, at least consistent with the view of an established camp within the PDP literature. Rumelhart and McClelland clearly understand their networks as possessing representations (i.e., states picked out by virtue of being about something external to themselves). In one passage, they occupy themselves with distinguishing their approach to cognitive modelling from that of the behaviorists.[31]

> ... [T]here is a crucial difference between our models and the radical behaviorism of Skinner and his followers. In our models, we are explicitly concerned with the problem of internal representation and mental processing, whereas the radical behaviorist explicitly denies the scientific utility and even the validity of the consideration of these constructs. The

[31] Perhaps it is even debatable whether the term "cognitive model" is applicable to behavioristic theories of intelligent behavior.

> training of hidden units is ... the construction of internal representations. The models ... concern internal mechanisms for activating and acquiring the ability to activate appropriate internal representations. In this sense, our models must be seen as ... strongly committed to the study of representation and process.[32]

Within the same work the authors devote an entire chapter to arguing that distributed *representation* (the sort that most PDP researchers use) is superior to localized representation (seen within some PDP networks, but more commonly associated with the traditionalist approach to representation). (I shall return to the topic of distributed versus local interpretation schema later in this section.) A second work within the PDP paradigm that has greatly influenced how (in particular) philosophers understand PDP, its assumptions and goals, is Paul Smolensky´s "On the Proper Treatment of Connectionism". He likewise enunciates a construal of PDP according to which research has as a focus gaining a better understanding of the concept of representation within cognition. A sample passage is:

> Hidden units support internal representations of elements of the problem domain, and networks that train their hidden units are in effect learning effective subconceptual representations of the domain. If we can analyze the representations that such networks develop, we can perhaps obtain principles of subconceptual representation for various problem domains.[33]

[32] *Parallel Distributed Processing,* page 121.

[33] Smolensky´s "On the Proper Treatment of Connectionism", page 8.

It is not enough, however, to state that PDP is committed to the representationality of certain network states. In order for PDP to constitute a genuine model of the mind, it must be the case that the content of these states plays a role in causal interactions. I shall have much more to say on this later.

If one is to take the above-quoted passages at their word, that PDP is concerned with representation, then there must be an explanation consistent with PDP principles that explains not only how PDP system states can, in general, be contentful, but also how particular contents are obtained. In other words, how can PDP answer the question: "why does this particular state have this particular meaning?" We saw in Chapter 3 that traditionalists have a story to tell (as Fodor would say) about content. Namely, the nervous system is a computer that implements a certain computational process. The process is defined in terms of a set of computational states and the rule-governed transitions between those states. Thus, each physical state that is a token of a physical state type that is a member of an equivalence class of state types constituting the computational state acquires its content from its corresponding computational state. This approach to explaining content inheritance is not open to the PDP theorist, however, for their system states (or, more precisely, the distributed system states) are not describable as implementing a computational process.[34] Thus, I must start back at the beginning with an examination of representation in general.[35]

I adopt the terminology of Dretske´s theory of mental representation (as described in his *Explaining Behavior*). He distinguishes three types of representational systems.

Type I representational systems.

Representational systems of type I are those in which the entities in the system both have no intrinsic power to represent and have their reference stipulated by the user of the system. Dretske describes a representation system of type I:

> Let this dime on the table be Oscar Robertson, let this nickel (heads uppermost) be Kareem Abdul-Jabbar, and let this nickel (tails uppermost) be the opposing center. ... With this bit of stage setting I can now, by moving coins ... around the table, represent the positions and movements of these players. I can use these objects to describe a basketball play I once witnessed.[36]

[34] The parenthetical remark must, for the time-being, remain somewhat cryptic. An explication of it and the philosophical exploration of its ramifications for representation a la PDP will take up a considerable part of Section 4 of this chapter. Unfortunately, since writing (and reading) a paper is a serial process, I must of necessity start somewhere, while making statements whose meaning will not become clear until later.

[35] In what follows, I am *not* making assumptions that in any way contradict what I have said in Chapter 3. Rather, in Chapter 3 I could skip such an examination because the equation of certain physical states with certain computational states "bootstrapped" representation -- or, at least, representation derived in terms of the purposes of an external observer of the system. I shall remark at the end of this introduction to representation how intentionality-via-computational-states fits into this scheme.

[36] *Explaining Behavior*, page 52-53.

Type II representational systems.

In contrast, representational systems of type II are only *singly* conventional: we assign a function to an element which has an intrinsic capability to indicate, and thereby determine what the indicator *represents.* Taking another example from Dretske, a typical fuel gauge in a car can indicate many things: the amount of fuel in the tank, the downward force on the bolts attaching the tank to the car, etc. *We* (the users of the system) determine what the indicator represents by assigning it a function -- in the fuel gauge example, we assign the gauge the function of representing the amount of fuel in the tank (and not the downward force on the bolts), because it suits our purposes.

Type III representational systems.

Representational systems of type III have no conventional aspect: no agent outside the system is needed to assign the representational function of elements within the system. Such systems "are ones which have *their own* intrinsic indicator functions, functions that derive from the way the indicators are developed and used *by the systems of which they are a part.*"[37] Dretske describes a system which embodies type III representation:

> Some marine bacteria have internal magnets, magnetosomes, that function like compass needles, aligning themselves (and, as a result, the bacterium) ... toward geomagnetic north. Since these organisms are capable of living only in the absence of oxygen, and since movements toward geomagnetic north will take the ... bacteria away from the oxygen-rich and therefore toxic surface water and toward the

[37] *Explaining Behavior,* page 62.

> comparatively oxygen-free sediment at the bottom, it is not unreasonable to speculate ... that *the function* of this primitive sensory system is to indicate the whereabouts of benign (i.e., anaerobic) environments.[38]

Representational systems of type III, because they do not rely upon conventional assignments of representational content, serve as the grounding for all intentionality. The elements of type I and II representational systems refer because we (i.e., human cognizers), by virtue of possessing the capability for type III representation, can stop the regress of derived intentionality: we ground all type I and type II representation with our underived intentionality.

Type I and II representation are easily explained, because there is (by supposition) an agent outside the system to assign a function to a representational element in the system; this is not the case with type III systems, for which the function is assigned by the system itself in the way in which an indicator is developed and used. What does it mean for an indicator to be so assigned? Dretske hypothesizes that the assignment takes place when the indicator comes to play a role in the causal sequence of an agent´s behavior. In the bacteria example, the direction pointed to by the magnetosomes represents "benign environment this way" because the indicator has been harnessed (via evolution) by the bacteria for its advantageous results of allowing the bacteria to live (and hence, to reproduce). To make this relationship between the indicator and its representational function clearer, consider a slightly altered

[38] *Explaining Behavior*, page 63.

example. Suppose that biologists discover that oxygen is *not* toxic for the bacteria in question. It turns out that the selectional advantage to the bacteria of possessing a magnetosome playing a certain causal role in the bacteria´s behavior is that it draws the bacteria toward the iron-rich sediment at the bottom, and away from the iron-poor surface water. (Let us suppose that the bacteria feeds on iron.) In this scenario, the magnetosome represents "good feeding ground this way". In fact, the magnetosome has represented this all along, even though we mistakenly conjectured that the selectional advantage offered by the magnetosome had something to do with the relative toxicity of the water for the bacteria. This illustrates that the representational function of an element in a type III system is intrinsic: it represents what it represents in its environment irrespective of the intentional states of systems other than itself.

(Thus, looking back to my depiction of representational capacity a la traditionalism, we see that computational states of an artificial device have type II representationality. To move from this derived intentionality to the original intentionality possessed by computational states implemented in natural objects -- eg, nervous systems -- there must be a causal story to tell about the advantage gained by a creature at having this state which is correlated with some external state-of-affairs. As I said in Chapter 3: "a traditionalist leaning towards the evolutionary approach to naturalization [of content] will answer that, in the history of a species, it has offered selectional advantage to have a body (or, more narrowly, a nervous system) whose physical states follow upon one another in the manner of Figure 2.")

A further distinction divides general type III representation into two classes: those resulting immediately from evolution and those gained as a result of learning during the lifetime of the representational system. Dretske allows only the latter to hold title to genuine intentional mental statehood. He explains this distinction in terms of whether the behavior of an individual depends upon what the internal state means or upon a particular genetic make-up which was selected for what the internal state means (as in the case of instinctive behavior). The bacteria is an instance of the latter; it swims in the direction pointed to by its magnetosome, not because of what the state of the magnetosome means *for it*, but rather because it has a genetic make-up which predisposes it to act in that way. According to Dretske, there is a qualitative difference between learned versus inherited behavioral dispositions. Briefly, this distinction is based upon the manner in which the representational element comes to play its role in the causal sequence leading up to the behavior. In the case of inherited dispositions, this occurs because the ancestors of the organism (system) gained selectional advantage by virtue of having a causal sequence where the representational element played this role. This is not the case for learned dispositions: one can give more than a selectional explanation for why a representational element means what it does. According to Dretske: "What explains why, during learning, R [an internal registration of a type of object´s presence] was recruited as, made into, a cause of M [a particular behavior] is the fact that R was a *sign* of O [the object type] and the organism had a need to coordinate behavior -- in this case evasive movements M -- with the presence of O. ... Hence, the internal sign of O (namely R) was made into a cause of M."[39]

Given the recent controversy surrounding the question of whether Dretske really has explained meaning,[40] I feel obliged to defend my adopted account of meaning against the charge that it is circular. First, though, a summarization of the charge: this account fails because it is committed to the following three theses:

> (1) *X*´s explanatory role is *X*´s causal role.
> (2) A state *C* has an explanatory role in virtue of having meaning.
> (3) A state *C* has meaning in virtue of having a causal role.[41]

The second thesis is the goal of Dretske´s (and my) whole project: to explain how it is that meaning is relevant. The third thesis encapsulates the means by which this goal is to be achieved -- namely, to ground the meaning of a state in terms of its causal role. I admit that Dretske´s account of the explanatoriness of meaning (as depicted by theses 1-3) appears circular; however, I, along with Dretske, want to distinguish "causal role" as it appears in the third thesis, from "causal role" as it appears in the second thesis (under the substitution of "explanatory role" with "causal role"). As Dretske argues:

[39] *Explaining Behavior*, page 19.
[40] See, for example, L. Baker´s "Dretske on the Explanatory Role of Belief" and Dretske´s reply "How Beliefs Explain: Reply to Baker".
[41] L. Baker, "Dretske on the Explanatory Role of Belief", page 100.

> ... *[C]urrent behavior*, the causal process that the meaning of *C* is called upon to explain (as structuring cause) need not (and typically will not) be the same sort of causal process as that which was responsible (during learning) for *C*´s acquiring that meaning. *C* got the function of indicating *F* (hence, this meaning) by being recruited to cause *M*, but what its having this meaning is (typically) called on to explain is its causing *N*, quite a different movement. And even if it *is* called on to explain the production of *M* (the same type of movement that it was recruited during learning to cause), it wasn´t its causing *M* that conferred an indicator function on *C*. It was its causing *something*, some movement *or other* (whatever movements were rewarded in the conditions *C* indicates). So the causal process (behavior) being explained by meaning is *never* the causal process underlying the meaning that explains it.[42]

How does Dretske´s account of the explanatoriness of meaning apply to the issue of intentionality in PDP systems? This question is highly relevant, for, if PDP wants to be a serious contender for a model of the mind, then it must be able to provide a principled explanation of how its systems´ states can be contentful. I would here like to deflect a possible objection that a PDP system, at least as currently embodied in the artificial computer science laboratory, cannot possess intentional states because the inputs to the system come not immediately from the environment of representable objects, but rather mediately via the researcher. So, the objection goes, this mediate-interaction version of PDP is not a possible model of the mind, as such systems lack the capacity for representation.[43]

[42] "How Beliefs Explain: Reply to Baker", page 115.

[43] One also hears this objection raised against traditionalism. My argument that the objection does not ultimately bear fruit applies equally well to PDP and traditionalism as potential models of the mind.

My response to such an opponent of PDP qua mental model is an unqualified "yes and no". First, the "yes". I agree that the mediateness of the stimuli to the system in such artificial environments is relevant to the obtaining of intentional mental states (in Dretske´s sense of the phrase). The relationship of the input supplied to the system with the real environment (of the researcher) depends upon (i.e., is mediated by) the researcher. If there is a counterfactual-supporting correspondence between a particular input vector´s being supplied to the system and a particular state-of-affairs, the counterfactual support relies on certain of the mental states of the researcher. The objection begins with the thesis that representational content is determined by the causal chain which results in the production of the representation. When the causal "distance" separating the object purportedly represented and the purported representation becomes too great, the latter loses its representational content. When aimed at the relationship between input to a system and objects or states-or-affairs said to be represented therein in such an artificial set-up, the causal distance is too great to support representation. Proponents of this view measure causal distance not in purely quantitative terms -- eg, how many causal laws need be invoked to get from A to B -- but in qualitative terms. In the case at hand, the causal distance is too great because the causal chain passes through the mental states of the researcher. A little reflection convinces one that *this* cannot be an objection against representational status, for the causal chain separating a person sitting for a portrait and the painted portrait likewise passes through the mental states of the painter. We would not, however, say that the painting therefore fails to represent the paintee. Perhaps, then, this objection is not so much directed at the ability of the input to represent, but at the ability of the input to support original intentionality in states of the system entered subsequent to receipt of the input. With this interpretation of the objection, it is not so clear that the painting case can serve as a counterexample, for it is not obvious (at least, my intuitions do not register a decisive response) that original intentionality is had by a creature exposed *only* to paintings. In order to thwart the objection, one can consider only the PDP systems which receive input (relatively) directly from the environment (mediated only by the necessary converters -- for example, a television camera that converts the light energy impinging on the lens to a "brightness at a point" matrix of numbers) as candidate models of the mind. (Thus, the camera functions as an artificial eye.)

With the above proviso, we can pose the question: how do the states of a PDP system come to have content? More precisely, how do they come to be intentional mental states? To answer this question, we must re-examine learning within PDP systems, concentrating this time on the representational aspects of the process. In what follows, I shall focus on supervised learning with a critic. I do this for several reasons. First, it makes it possible to get the researcher (and all other supervisory cognitive agents) out of the learning loop: the reinforcement signal, like the input, can be supplied by the environment itself. Thus, the objection that the system has only type II representation, because the actual content of the states is supplied by the supervisor (in her judging of the appropriateness of the produced response and supplying of the correct one) is thwarted. A second reason for favoring learning with a critic is that it is the predominant mode of learning in natural creatures -- it is the exceptional case in which an external source is available to supply information detailing the correct response to a particular situation. As I am in the next few pages concerned primarily with explaining how PDP states can be representational, limiting myself to learning with a critic is justified. However, as I mentioned previously, an adequate model of the mind must also allow for exceptional cases (such as supervised learning with a teacher); hence, mechanisms supporting both forms of learning must be present. To simplify the exposition, I can merely assume that, while content is determined in both types of learning cycles, the content during learning with a teacher cycles is parasitic on the content acquired during the more common learning with a critic cycles.

So, the general framework has the environment (via converters) supplying input vectors, and the environment (via the reinforcement signal) supplying the feedback on the adequacy of the produced output. The front-end of this set-up is fairly straightforward: it is easy to construct (or imagine) audio and video equipment pointing out at the world, converting the inflowing information into a segmented signal capable of being used as input to a PDP network. It is perhaps less obvious how the other end (i.e., the reinforcement signal) is constructed. Consider first the manner in which reinforcement information is supplied in natural creatures. The creature performs a particular action in the presence of a stimulus. If the action is immediately followed by a relatively pleasurable experience (for example, a cessation of an unpleasant thirst with a neutral feeling of equanimity), this serves as a reinforcement signal, which tends to produce changes in the creature such that, in the future, it is more likely to perform that action in the presence of that (and similar) stimuli. Contrarily, when the creature performs an action in the presence of a stimulus that is followed by a relatively unpleasurable experience, the causal pathways linking stimulus and response will change so as to make that response less likely in the presence of that stimulus.[44] This explanation works for natural creatures, because they come equipped with (at least a rudimentary) pleasurableness detector -- warm, fuzzy sensations are pleasurable, whereas sharp, obtrusive sensations are not. Evolution has supplied these detectors to aid in the survival and reproduction of their possessors. This is because the sorts of behaviors resulting in warm, fuzzy sensations (e.g., eating) tend to be those that also aid ultimately in reproduction; whereas the sorts of activities resulting in sharp, obtrusive sensations (e.g., burning oneself) tend to be those that have a deleterious effect on reproduction. Does it make sense to say that PDP systems likewise come equipped with a detector which, like the above-described pleasurableness detector, can supply a reinforcement signal given environmental conditions (including the conditions of the system´s

[44] I am using this vocabulary, not to emphasize a connection with behaviorism in general or instrumental conditioning in particular, but rather as semantically-neutral descriptors of the signal received, input and output of the system -- whether that system be natural or artificial. Given that I have yet to argue that this process can result in intentional states on the part of the learning creature, it seems premature to refer to an input as "perception of the presence of x".

physical parts)? While they would not be the product of the evolution of the system´s forebears, a system could certainly be equipped with such detectors. Does this make a difference to the status of PDP as a mental model?[45] I think not. I say this because the purpose of the reinforcement signal is ultimately to establish type III representational content, which in turn is necessary for subsumption under mental causal laws. Qua mental model, a PDP system need only be able to participate in such content-adverting laws -- whether the means by which content is determined within PDP systems differs from the means employed within natural creatures is irrelevant. (In other words, the means is merely an implementation detail, not pertinent to PDP´s status as a model of the mind.) If I can explain how these artificial detectors can be used to supply a reinforcement signal for training the net, and, in the process, give content to certain of the system states, that is enough.

Strangely, among all of the recent works on PDP (written both by PDP researchers and by philosophers), I have yet to encounter a detailed explanation of how, exactly, PDP states represent.[46] This lacuna is particularly striking, given that the word "representation"

[45] It may seem as though I am getting rather far afield, but it is a common assumption (among German neurobiologists, at least) that the *biological nature* of a mind-possessing creature is important. (One can also see such a concern among some American philosophers -- eg, Searle -- although his argument for the importance of biology takes a slightly different tack.) For those espousing this view, merely building in a "reinforcement module" won´t do. Hence, my argument.

[46] I here and henceforth shall mean type III representation by "representation". No difficulties in explaining type II representational content in PDP states arise, as the content is assigned by the researcher in the act of labelling a unit (for localized representation) or a pattern of activation (for distributed representation). Clearly, though, the act of labelling does not make a difference *to the network* -- it goes on processing quite oblivious to what is assigned to its states: the labels play no role in the ensuing procession of states.

as describing such states is so freely used. (My personal theory as to why this topic has not been addressed relates back to my conviction that the writing in this field occurs independently of a developed theory of causation -- and, without such a theory, it is not possible to give a detailed account of representation.) So, let´s start at the beginning (again). We have a network with a fixed architecture (e.g., number of units, available communication lines), a hard-wired learning rule, and weights initialized to small, random numbers. Figure 10 gives an overview of the system and its relation to the environment.

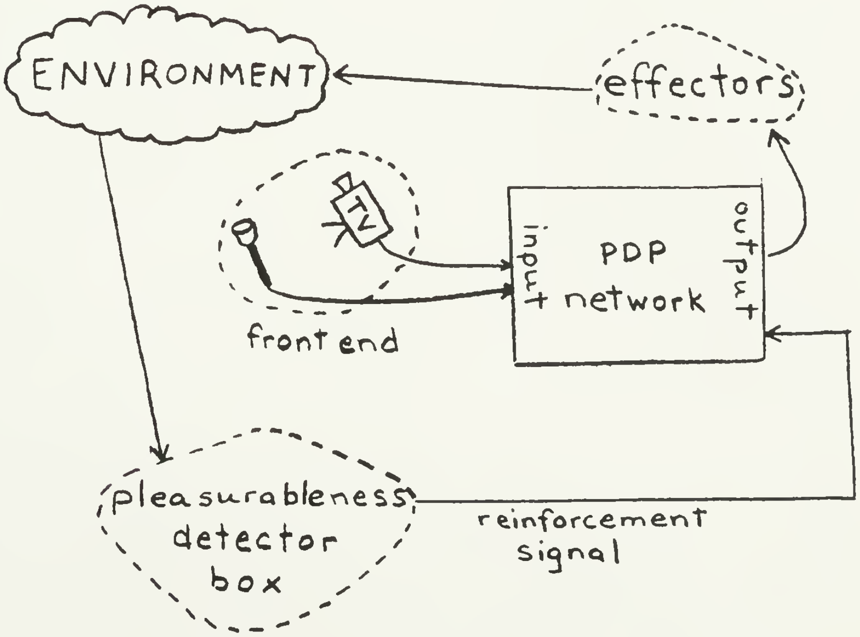


Figure 10 -- PDP network in system capable of supporting representational states

As the system is initialized (i.e., prior to training), the network states do not have representational content. The audio-visual front-end receives information from the environment, and transforms it into a format suitable for use as input to the network. The introduction of this input vector produces changes in the activation values of the input level units, then subsequently the rest of the units in the system. All current PDP systems presuppose temporal quantization: the input signal does not change continuously, but rather changes at discrete time steps. We can stipulate that the audio and video devices likewise sample the information available from the environment at discrete points in time; to simplify the description of network behavior, we can assume that the time interval between input vectors is large enough to allow use of the learning algorithm described in Section 2 of this chapter. The network output is used to drive effectors that (for example) manipulate objects in the environment. A typical effector is a robot arm. Other types of effectors are also possible, for example, a transportation sub-system capable of changing the location of the system relative to the environment. The effectors force changes in the environment, which, in turn, may cause an increment or a decrement in the pleasurableness value returned by one or several of the pleasurableness detectors.

As with the effectors, there are many possible types of pleasurableness detectors. What they all have in common is the ability to produce a signal measuring some factor relating to the hospitableness of the environmental conditions for the continued functioning of the system (i.e., the network plus front-end plus effectors). One type of pleasurableness detector often seen in mobile robots is a simple volt-meter measuring the energy reserve available in the robot´s storage battery. The higher the energy reserve, the greater the pleasure signal. This is loosely equivalent to a hunger detector (which takes into account food in the digestive system and energy reserve in the form of blood sugar and stored fat) in natural creatures. One can imagine other types of detectors based loosely on the sorts of detectors selected for by natural evolution -- temperature detectors, strain gauges, etc. The overall output of the pleasurableness detector box (i.e., the reinforcement signal) is the difference between the overall pleasurableness at the current time step minus that at the previous time step, where the overall pleasurableness at a time step is some function of the outputs from each of the individual pleasurableness detector apparati. With this access to a reinforcement signal, the network can use the learning procedure described in Section 2 to force a change of weights in a manner so as to decrease the probability of doing the "wrong" thing relative to the output of the pleasurableness detector box. If, as supposed, the output of this detector box is correlated with the continued-functioning-of-the-system expectancy, this means that learning will result in a system whose output (and, via the effectors, behavior) is decreasingly likely to be deleterious to the continued functioning of the system.

While this sounds well and good, it is open to the opponent of the possession by PDP systems of type III representational states to argue as follows: although the researcher is in one sense out of the learning loop, in that she need not even be present during the actual learning phase; still, various of her choices in the design of the overall system have as a consequence that the system achieves only type II representation. Such an opponent may point to any of several design choices as critical in this regard -- these include: (1) front-end, (2) network architecture, (3) learning rule, (4) effectors, and/or (5) pleasurableness detectors. Does the fact that these choices were made by a cognitive agent (i.e., the researcher) and not by evolution make a difference to the type of representational status had by the network´s representational states? In answering this, it is important to keep in mind that the representational states in question are *not* the states of the various devices forming the interface between the network and the environment, but rather the states internal to the network. The network achieves type III representation if some of its individual states achieve type III representation. A particular representational state is type III if its causal role (i.e., its place in the chain of states during processing in the network in regard to the other states within the chain) derives from the history of the system -- in particular, from the way that changes in its causal role aided in the increase of survival expectancy (as measured by the pleasurableness detectors) during the learning phase. That the choice with respect to design parameters was explicitly made by a cognitive agent is therefore irrelevant to the type III status of the network´s representational states. A simple thought experiment also produces the same conclusion. Imagine a person born with a perfectly normal brain, but lacking all of the five sense organs. Imagine further that this person is supplied with an artificial eye much like the TV camera forming part of the front-end of the overall PDP system. The mere facts that the eye is artificial and that a physician chose which particular artificial eye to use do not prevent the person from becoming a type III representational system after learning. I think that similar thought experiments questioning the relevance of an external cognitive decision-maker to type III representational status can likewise be given for each of the other four design choices, although, particularly for the case of the architectural design, the scenario to be imagined would be very far-fetched. Just so, the fact that various design decisions in the construction of the PDP system were made by the researcher does not prevent the system from becoming a type III representational system. Thus, we have satisfied one of Dretske´s conditions for the presence of intentional mental states.

Here, I would like to distinguish two types of states used to describe the "current state" of a PDP network: (1) the current weight state (i.e., the matrix of weight values connecting the units to one another) and (2) the current activation value vector (i.e., the vector whose elements are the activation values for each of the units), or, alternatively, the current output vector (i.e., the vector whose elements are the output values for each of the units). All three are candidates for contentful states. In what follows, I use the terms "weight state" and "activation value state" in such a way as to avoid committing myself one way or another on the question of whether such states are atomic or complex with respect to content. As with the discussion in Chapter 3 dealing with the supposed representational complexity of the monolithic computational states, I want to remain non-committal (at least, at this stage of the discussion) with regard to whether it is only the network weight state (or network activation value state) as a whole that can carry content, or whether sub-states of the network weight state (or network activation value state) can also be contentful. In general, the writings of most PDP researchers indicate a willingness to allow sub-states of network states to be contentful.

Now it is time to look inside the network as learning progresses, in order to isolate exactly where the changes are being made and how those changes bring about representationality. As already stated, prior to training, none of the network´s states are representational.[47] In what follows, I present a scenario showing how a network could come to possess a representational state. The network is presented with an input vector. Let us suppose that the scene encoded by the input vector is of a fire. (Clearly, this level of description is that of an outside observer -- the system does not yet possess a representational vocabulary at all, much less one capable of distinguishing a fire from other potential objects in the environment.) As the network´s weights at the beginning of the learning phase are randomly assigned, the network´s output, and, hence, the system´s behavior, will likewise be random. Suppose that the behavior produced just so happens to move the system further away from the fire (perhaps by a command to the transportation sub-system to move in a particular direction which in this case happens to be away from the fire). So we have an input/output pair

[47] Obviously, I am here speaking only of type III representation. It is possible that the researcher, by a clever choice of weights, has produced a network with type II representational states.

(i.e., input vector encoding of the fire scene and "move away" command), and a change in the environment relative to the robot. The change in environment results from the relative movement of the fire. Suppose that one of the pleasurableness detectors supplied to the system is a thermometer measuring the temperature of the air near the system surface, calibrated so as to return the values shown below:

Value	Temperature Range
3	60-75 F
2	40-59 F, 76-80 F
1	20-39 F, 81-95 F
0	5-19 F, 96-110 F
.	.
.	.
.	.

Figure 11 -- Hypothetical pleasurableness values returned as a function of temperature

Suppose further that the only pleasurableness detector that changes value in the time step after execution of the behavior is this thermometer, which measures a temperature drop from 62 F to 58 F.[48] So, the reinforcement signal received for the I/O pair is

[48] Obviously, the example I am presenting, with its many suppositions and happy coincidences, is not very realistic as an actual sequence of events. This fact does not, however, detract from its usefulness as a summarization of the (for present purposes) relevant changes leading to a representation of a fire. A more likely sequence of events -- for example, one in which no change of pleasurableness level is evoked from any of the detectors over many time steps -- would lead to the same representational state. Describing the process would, however, necessarily include many irrelevant details.

negative, thus forcing a change in weights to make this output less likely in the presence of this and similar inputs. The next input vector is again directed at the fire scene; although it will have changed from the first vector in that the fire is further from the front-end. Now suppose the output produced for this input is such as to direct the transportation sub-system to drive the system as a whole closer to the fire. The 60 F temperature threshold separating the "ideal" and "suboptimal" ranges is crossed, and the reinforcement signal becomes positive, which tends to increase the likelihood of continuing to approach the fire. This is repeated. At some point, the 75 F threshold will be crossed, resulting in a negative reinforcement signal, and a corresponding refinement of the "approach-fire" behavior.

During this group of learning cycles, the input vector has been changing, although, by supposition, it has at each stage been directed at the fire scene. The first change (resulting from the initial "retreat" behavior) had the fire taking up a smaller area of the visual field, with decreased overall brightness, and a decrease in the decibel-level of "fire-type" noise. The negative reinforcement signal brought about a change in behavior from "retreat" to "approach", which in turn produced changes in the environment relative to the system. In particular, the portion of the visual field occupied by the fire increased, as did the brightness level and "fire-type" noise.[49]

[49] I would assume that the "crackliness" property of the sound of a fire is isolatable in the frequency profile of an audio reproduction of a fire, and that the rapidly fluctuating brightness and color characteristics of a fire are likewise isolatable in the visual reproduction of a fire. In order to capture these aspects of the *continuously*-changing environment, it may make most sense to have each input vector be not an encoding of the input-at-a-moment

The "approach" behavior led eventually to an increase of the temperature to a value greater than the maximum ideal temperature, thus producing a negative reinforcement signal and a change from "approach" to "retreat" behavior. This in turn led to further changes in the input vector (i.e., decreased fire area, decreased brightness, decreased level of "fire-type" noise). Obviously, other changes in the input vector not directly related to the fire were simultaneously occurring. For example, the background visual signal was changing as the system moved, as was the background noise. Other additional changes were also taking place. Perhaps other mobile creatures entered and/or exited the perceived environment during this time period. Hence, viewing this sample sequence of events in isolation, it is not immediately obvious that it is the aspects of the input directly related to the fire that are important in the changes in reinforcement signal. However, suppose that the system has repeated encounters with fires. The non-fire-related particulars will change with each encounter, while the fire-related effects on the pleasurableness detector (and, in turn, on the overall reinforcement signal) will remain constant. Thus, the change in weights forced by these repeated fire-encounters will tend to make the approach-retreat behavior depend only on the visual and auditory properties of the fire, and not on the non-fire-related circumstances.

Looking back to Dretske´s explanation of the determination of representational content, we see that all of the conditions for the

alone, but rather an encoding of the input-at-a-moment and an encoding of the variability between time steps of the signal received by the front-end.

acquisition of intentional mental statehood are satisfied in the above-described depiction of learning in a PDP system. The internal registration of the fire´s presence, R,[50] has been (via the learning rule) made into a cause of the approach-retreat behavior, M. This is explained by the fact that the system had a need to coordinate its behavior with the presence of fire -- fire can be, given certain system behaviors, highly deleterious to the system. This deleteriousness detection is mediated by the pleasurableness detectors, which were chosen because of their ability to correlate the presence of potentially harmful conditions (brought about by the presence of potentially harmful objects) with a reinforcement signal. R represents fire because R´s causal role lies between "presence of fire" and "fire-approaching-and-retreating behavior". This causal role is the result of changes made in the weights of the network. After learning, an input vector encoding a fire scene produces R, which in turn produces a certain set of behaviors relative to the fire, which tend to increase the probability of the continued functioning of the system.

Let´s look at the same process in light of my possible worlds theory of causation. Suppose that R did not represent fire (either because it represented something else, or because it represented nothing at all). Thus, the production of R would not be correlated with the presence of fire. In this case, R´s causal role would not lie between "presence of fire" and "fire-approaching-and-retreating behavior". (Assume that the general architecture of the system

[50] I have not yet discussed what this internal registration is. That will come presently.

remains unchanged in this counterfactual world.) If R did not represent fire, then R would not govern the particular approach and retreat behaviors that it in fact governs, for the learning of these behaviors was mediated by the heat-producing characteristics of the fire. Without a fire, there would be no heat production; hence, no changes detected in temperature; hence, no positive and negative reinforcement signal (or, at least, not this particular pattern of reinforcement); hence, not this particular change in weights; hence, no learned approaching and retreating behavior. If R were instantiated in the network, the output of the network (and ensuing behavior of the system) would be different. R´s causal role is thus dependent upon its meaning, as it must be if R is to participate in a mental causal law. As I argued in Chapter 2, it is illegitimate to construe the antecedent to the counterfactual conditional testing for causal relevance (i.e., "if the system were not in that mental state") as "stipulate that the system is in the same physical state as the one implementing that mental state in the actual world, but assume that the instantiation of that mental state means something else, and leave everything else in the world (including the past) unchanged." Thus, not only is R meaningful, but its participation in causal laws adverts to its content. Recall Dretske´s soprano example, in which the meaning of the words sung by the soprano are causally irrelevant to the shattering of the glass. This case is different, because the meaning of R *is* causally relevant, for, if R had meant something else, then its effects would have been different.

I mentioned previously that it is currently a debated topic whether Dretske´s account of the relevance of meaning is circular or not. With the general case, as well as with this particular use of Dretske´s account, I don´t think the charge of circularity is correct. Let´s look in more detail at how "causal role" is to be understood in the context of the two sentences:

(S1) R represents fire because R´s causal role lies between "presence of fire" and "fire-approaching-and-retreating behavior".
(S2) R´s causal role is dependent on its meaning fire.

Prior to learning, R was an indicator of fire: a token of type R became instantiated whenever a fire was observed (although, at this early stage, R does not *mean* fire). S2 is describing how R came to have its causal role after learning. During learning, the repeated co-occurrence of the triple:

(1) tokening of R,
(2) particular behavior,
(3) particular reinforcement signal

led to changes in the network such that tokenings of R came to have control over certain behaviors. Had R been such that its meaning (if it were to have one) came to be something other than fire, it would have had a different causal role than it in fact came to have. This is because the changes made in the system relevant to the tokening of the R-behavior sequence during learning were guided by the reinforcement signal received during learning. But the particular reinforcement history would have been different had R (truly) indicated, not fire, but (for example) human face. Thus, S2 is making the counterfactual claim that had R indicated something other than fire (and, hence, had the history of reinforcement vis-a-vis fire been different), then the causal role that R eventually took on would be different. Thus, S2 is true by virtue of the past (actually occurring) tokenings of R in the presence of fire. S1, on the other hand, is explaining why this and all future tokenings (assuming learning had ceased) of R represent fire as opposed to something else: they represent fire because they lie between "presence of fire" and "fire-approaching-and-retreating behavior". S1 is wholly silent on how this causal role for R has been arranged.

While the above fire example was chosen for its relative simplicity, there is no principled reason preventing all representational states in PDP networks from gaining their content in a similar manner. A point that deserves to be emphasized is that learning in PDP networks (and, correspondingly, content acquisition) occurs piecemeal and achieves relative stability only after a long training period. One would assume that the number of learning cycles needed for the formation of a representation of (for example) another cognitive agent would be very large and that a description of its acquisition in terms of environmental feedback would be very complex. Very large and very complex do not, however, imply impossible. Judging PDP as a model of the mind based on this criterion (i.e., amount of exposure to an object needed to acquire a representation of it) does not automatically exclude it, for one-pass learning is the rare exception. When considering representation acquisition in humans, one must also include the learning cycles needed to produce the base of representations from which complex representations are formed -- a process lasting many years and encompassing *very, very* many learning cycles. I can only think of one case of genuinely one-pass learning among humans: namely, the learning of an aversion to a food type after only one exposure to it, when that exposure is followed by severe nausea. However, I am not at all ready to grant that the causal law governing this aversive behavior is mental (i.e., that it adverts to contentful states of the agent).

One advantage of the explanation of the acquisition of intentional states within the PDP framework over that within the traditionalist framework is that one can see in PDP how representational states emerge from an initially unstructured network. The regularities in the environment resulting from the existence and persistence throughout time of objects guide not only the learning of behavioral responses appropriate to the objects, but also the learning of the representations of the objects. Thus, one can see the Humean bent of PDP, which presents a contrast to the more Cartesian bent of traditionalism. In the latter, concepts are innate to the extent that traditionalism provides no explanation for the acquisition of their manner of representation. In contrast, PDP networks must also learn how to represent a concept: it must answer for itself the following questions. Is a concept atomic or complex? If the latter, are there any necessary and/or sufficient conditions associated with it? If so, what are they? Learning within PDP explains how these questions can be answered given merely the system set-up and information from the environment, whereas the process of concept formation within traditionalism either remains unexplained or presupposes the innateness of concepts. This Humean flavor of PDP has led some traditionalists to direct the same arguments against PDP as were aimed against associationism (and, in particular, behaviorism). These come in two sorts: (1) complaints that the emerging model cannot be of the mind because it does not take into account the representational properties of causally efficacious states, and (2) poverty of the stimulus arguments. I have already given reasons for rejecting the first type of attack against PDP: it is clearly concerned with representation, both in explaining its acquisition and its causal relevance. I explicitly addressed this complaint against PDP because it strikes at the root of my contention that PDP can constitute a coherent model of the mind. Poverty of the stimulus arguments, on the other hand, are directed at the empirical adequacy of PDP as a model of the mind, and, as I am not here concerned with arguing either for or against PDP or traditionalism as providing the best model, I shall not pursue this topic further.

I have yet to specify what sorts of states within PDP systems are the bearers of causally-efficacious content. The reasons for my reticence on this point are two-fold: a lack of consistency within the PDP literature and the necessity to tackle only one issue at a time. I hope that the above arguments suffice to convince the reader that PDP networks are capable of possessing representational states. Now all that is left is to pick out which among the candidate states are representational, and participate in causal laws as a function of their content. As to the first reason, I have now reached a point where, if the analysis is to continue, I must disregard as inconsistent many of the stated views on this topic made by PDP researchers.

I have already mentioned the two candidates most often mentioned in the PDP literature for intentional mental statehood:

the weight-state (or parts thereof) and the activation value state (or parts thereof).[51] One condition set down in Chapter 2 for an intentional mental state type is that its tokens must participate in mental causal laws (or, at least, must potentially do so). A mental causal law is one relating an intentional mental state either to another mental state or to some external behavior. How can one understand the causally interacting representational objects in a PDP network? The only regularities of transition that are isolatable in a PDP network are the unit-level rules governing unit output as a function of activation value, and activation value as a function of the connectivity pattern of the network (encoded in the weights) and the state of the local units to which this unit is connected. Is the unit-level the appropriate level to consider in looking for mental states? Perhaps the activation value of a unit represents the presence (or, for real-valued units, the degree of presence) of an object. On first glance, this is the most intuitive interpretation schema for a PDP network. The weights would then correspond to the degree of association of the objects represented by the various units. (This "degree of association" may even constitute a degree of conditional support -- thus, if unit-n is connected to unit-m via a line with a large positive weight, then a large output on unit-n lends a high degree of support that the object represented by unit-m is also present.) This approach to interpretation (hereafter called the

[51] The latter is often used interchangeably with the output of each unit state. In networks with a one-to-one function relating activation value and output of a unit, the two state-types collapse into one. As the network exemplar that I have in mind throughout this chapter is one using back-prop as it is currently construed (ie, in conjunction with a squashing function a la Figure 5c), I confine myself to consideration of only activation value state and weight state, but not output of each unit state.

"local interpretation schema"), while easy to understand, presents some difficulties.

Experiments on PDP networks capable of learning often show that, after training, there is no consistent assignment of individual objects to units: the activation value of the units is not correlated with the presence or absence of any particular object in the domain of commonsense representable objects. Perhaps, even in this case, the local interpretation schema can be salvaged, for, perhaps, the units represent, not the objects picked out by words in our natural language, but objects that are picked out by a very large disjunction. Is there a reason to reject such "objects" as genuine, at least to the extent that representations of them participate in mental causal laws?

For those philosophers (e.g., Grice) who want to use contentfulness of mental states to ground contentfulness of words and expressions in our public language, such a large mismatch between mental representational units and linguistic representational units would be unacceptable. For such philosophers, mental representation is basic, and language has evolved as a means to encapsulate the possible mentally represented items and allow its transmission among minds. But then, our public language should have been capable of easily capturing these mental representational units. This is clearly not the case. (In general, it is only with considerable awkwardness that PDP researchers can encapsulate the content of a unit-level representational state into natural language.) The most obvious response to the argument against viewing the unit activation values as the bearers of causally efficacious mental content of someone keen on defending this view is to point out that there are alternative means to grounding the meaningfulness of words and expressions in our public language. One such alternative (most often associated with Wittgenstein) is to ground meaning in social practice. The possibility of such alternative avenues for grounding meaning of words and expressions takes the punch out of this sort of argument. In order to shore it up, such an opponent to unit-level representation must argue that the Gricean approach is the only contender for grounding linguistic meaning. I, for one, cannot imagine how such an argument would go.

Another line of attack against the unit-level representations as quantified over in mental causal laws view of PDP focusses on the mismatch between mental processing described in terms of unit-level meaning units and mental processing described in "stream of consciousness" reports during (for example) problem solving. An assumption of this argument is that stream of consciousness reports provide an accurate picture of the causal goings-on in the mind of the reporter. There are two possible counters to this argument. The first involves questioning the assumption that stream of consciousness reports bear any relation to the actual mental processing involved in problem solving (or, mentation in general). One sees even within the traditionalist framework (e.g., in the work of Freud) a questioning of the reliability of subjective reports. It is merely one more step down this familiar road to question not only the reliability of the reports, but also the reliability of the vocabulary in which the reports are couched. A second approach to countering this argument against unit-level representation harks back to the equation within PDP (and, within traditionalism as well) of the mind with a certain type of processing. Consciousness in general is left as, at best, a by-product of mental processing. Thus, PDP needn´t take conscious reports as conclusive in this regard. It is perhaps strange that there is such a mismatch between the two vocabularies, but not decisive. (As I am not taking the view of PDP as reconstructed by me to be that unit-level representations are those quantified over in mental laws, both the Gricean complaint and the disparity with stream of consciousness reports complaint constitute mounting evidence in favor of the alternative interpretation schema.)

Neither of the above arguments succeed in knocking out unit-level representation as a contender for mentally causally efficacious items. They at best tend to disconfirm this thesis by pointing out aspects in which its implications are counterintuitive. As the history of other scientific disciplines (particularly in the 20th century) makes abundantly clear, mere counterintuitiveness is not by itself reason to reject a theory. I shall return to a consideration of unit-level representation later. Now, however, I would like to consider another set of candidates for the bearers of causally efficacious representation within PDP: namely, the pattern-level activities. Under the rubric of multi-unit patterns, there are two commonly mentioned possibilities for bearers of meaning: patterns over units´ activation values and patterns over weights. The patterns in question may encompass one unit, multiple units, or all of the units in the network. Recall in Chapter 3 that, strictly speaking, a computational state is the monolithic system state. It is only under the assumption that parts of this monolithic state are causally isolatable that it is legitimate to speak of the representations of these parts as being causally efficacious. A similar issue crops up with respect to the isolatability of parts of the monolithic network state (whether of weights or of activation values) within PDP systems. Under what assumptions is it legitimate to speak of a pattern of activation or pattern of weights as a causally efficacious representational item, when that pattern is only constituted by a proper subset of all of the network´s units? What does "causal isolatability" mean in this context? A pattern consisting of a proper subset of the network´s units is causally isolatable when (1) its representational content is adverted to in a mental causal law, and (2) all other representational states adverted to in that mental causal law interact with that subset of states such that the set of units not a member of the subset are irrelevant to the proper instantiation of the law: relative to all the mental causal laws in which that subset partakes, the non-subset members are irrelevant. Whether there are such causally isolatable subsets is an empirical question which will not be pursued here. When I use the word "pattern", I mean to include any such subsets. If there are none, then "pattern" refers only to monolithic network states. A further note on terminology. The attempt to assign causally relevant meanings to patterns over many units (as opposed to the activation value of a single unit) is referred to in the literature as the "distributed interpretation schema", in that the contents borne by PDP networks are *distributed* over multiple units.

Let´s first examine patterns of weights as potential bearers of content. Can a coherent picture of mental causation emerge from such an assignment? Clearly not, for mental causal laws relate mental states with one another. The format of mental causal laws is: mental-state-1 causes mental-state-2.[52] But patterns of weights do not cause one another. In particular, it is not the case that the instantiation of one pattern of weights is immediately followed by another. So, patterns of weights alone cannot be the sought-after bearers of causally relevant content.

Perhaps, then, sense can be made of patterns of activation values as causally efficacious representations. Here, at least, the overall format of mental causal laws can be applied to transitions in patterns of activation during processing: patterns of activation follow upon one another from one time step to the next. Patterns of activation as mental states also satisfy the format of the mental causal laws relating mental states and behavior, for a particular network output (driving system behavior) is immediately preceded by a pattern of activation. So, at least on this superficial point, patterns of activation are contenders for mental statehood. Closer analysis shows, however, that patterns of activation *alone* cannot be mental states, for the transitions between patterns of activation are law-like only relative to the weight state of the network at the time of the transition. On this point, the traditionalist mindset, with its notion of mental causal laws encoded in a static algorithm governing

[52] Mental causal laws can also relate mental states and behavior. I do not bother considering these, as the patterns of weights can already be excluded as potential causally relevant mental states, because they cannot fit into the framework of mental causal laws relating one mental state with another.

the manner of manipulation of stored representations, must be jettisoned. In a PDP network, there are no explicitly encoded rules governing manipulation of a distinct group of mental states; rather, it is the total pattern of the network (i.e., activation values plus weights) that forces transition to the next pattern. So, the bearers of causally efficacious content within PDP networks are patterns over activation values and weights. I shall look in more detail at the implications of this view for PDP as a model of the mind in the next section.

Before leaving the present section, however, I would like to sum up the most central points of my examination of PDP as it is currently practiced. A useful springboard for such a summarization is a consideration of what the separate words within the name "parallel distributed processing" entail, especially in light of PDP´s use as a model of the mind. The word "processing" distinguishes PDP as a particular way of arranging state transitions -- namely, in a manner that (1) has many sub-processes going on in parallel, and (2) involves somehow the amalgation of each of the individual sub-processes into a larger unit. In relation to PDP as mental model, the thesis is that the mind is likewise a process with just these characteristics. The "parallelism" refers to the fact that many simple units are operating simultaneously, such that each of these simple units has access only to locally available information (i.e., information on the state of those units to which it is connected). The "distributed" nature of PDP networks describes the usual manner of interpretation of representational atoms within the network. This distribution of representation is not so much spatial as it is involving the contribution of many relatively-independent simple units. This distinction is important, because, in one sense, representations in traditionalist systems implemented on a physical computer are distributed -- in space. For example, the medium in which a representation is stored may, and usually does, involve multiple individual locations (whether they be multiple physical locations on a chip, or, in the extreme case, multiple storage devices). What distinguishes the spatial sense of distribution from the one meant within the context of the title PDP is that, in the former, the representation is manipulated *as a unit*, despite its spread-outness.

4.4 PDP as a Model of the Mind

In this section, I examine some of the implications of PDP´s properties, when PDP is construed as a model of the mind. Some of the particular topics I include on this score are: the ramifications of the mental state as patterns over weights and activation values; the (still unresolved) issue of local versus distributed interpretation schemas; and generalization (semantically described) within PDP networks. I end the section and chapter with a description of PDP as an explanatory model of the mind. What ontological commitments does it make? What level of reality is represented by its causally efficacious intentional states? What form will its causal laws take?

I begin by taking a closer look at the ramification of PDP as a model of the mind, under the assumption that the bearers of causally efficacious content are patterns over weights and activation values. Perhaps the simplest way to broach this subject is by a quick contrast of this view of mental activity with that implied by traditionalism. Rumelhart and McClelland provide just such a contrast:

> In most [ie, traditionalist] models, knowledge [ie, a representation] is stored as a static copy of a pattern. Retrieval amounts to finding the pattern in long-term memory and copying it into a buffer or working memory. There is no real difference between the stored representation in long-term memory and the active representation in working memory. In PDP models, though, this is not the case. In these models, the patterns themselves are not stored. Rather, what is stored is the *connection strengths* between units that allow these patterns to be re-created.[53]

While the above-depicted "database" characterization of traditionalism is only one among many (one heavily influenced by the current set of programming languages available to the computer scientist), it points out by overstatement a basic difference between the two modes of processing: in traditionalism, there is a clear distinction between the rules governing manipulation of representations (whether those rules are explicitly or implicitly stored) and the representations thus manipulated. In PDP, on the other hand, the two are not clearly distinguishable: one pattern of activation follows upon another as a function of the weights, which encode the potentiality of the production of patterns of activation. Thus, the particular succession of patterns of activation is determined by the network itself. One cannot therefore view the

[53] *Parallel Distributed Processing*, Vol I, page 31.

network merely as a storage medium for representations, which, as passive storage medium, is operated upon by some external process.

A useful way of viewing the relative contribution of the patterns of activation and the patterns of weights is in terms of explicit versus implicit representations; although, as argued in Section 3, this way of viewing it is only approximately correct, as neither patterns of activation nor patterns of weights are in isolation constitutive of representations. In describing the contents of the mind, we often distinguish between those items that are explicitly represented (usually understood as being available to introspection) versus those that are implicitly represented (only potentially or latently available to introspection). As I have already mentioned, PDP as a model of the mind tends to discount the importance of (conscious) introspection, yet the implicit/explicit distinction in representation has survived within PDP in a slightly altered form. Let´s look again at what happens during network processing. The network has a particular pattern of activation instantiated in the activation values of its units. It also has a particular pattern of weights. The weights determine the line of succession (relative to a sequence of input vectors) of one pattern of activation upon another, and the pattern of activation instantiated picks out where in that line of succession the network is currently located. Thus, the current pattern of activation is explicitly realized in the network, and the pattern of weights encodes the information needed to produce the future patterns of activation: the future patterns of activation are implicit in the weights. It is only a short step to using the word "implicit" to describe the representational status of the weights. One sees this usage quite often in the PDP literature, as in:

> ... almost all knowledge [ie, representation] is *implicit* in the structure of the device [ie, in the connections] that carries out the task, rather than *explicit* in the states of units themselves. Knowledge is not directly accessible to interpretation by some separate processor, but it is built into the processor itself and directly determines the course of processing. It is acquired through tuning of connections as these are used in processing, rather than formulated and stored as declarative facts.[54]

This reworking of the implicit/explicit distinction offers the possibility for a smooth union in the interpretation of the two "directions" of processing with PDP networks: forward processing (i.e., the succession of one pattern of activation upon another) and backward processing (i.e., learning). Backward processing is the tuning of the weights so as to fix the line of succession of patterns of activation relative to a sequence of input vectors -- to bring about the succession of the explicit states that instantiate mental causal laws. Forward processing is then the unfolding of the causal sequence. This way of putting it is still too simple, because the processing involved in learning likewise involves mental causal laws. The picture of mentation that emerges is, at least at this level of description, very different from that of traditionalism. According to PDP, there is less differentiation of the elements of the mental realm into the static, discrete representations and the mental algorithm

[54] Rumelhart and McClelland, *Parallel Distributed Processing*, Vol. I, pp. 75-76.

that operates upon them. Because of the time quantization needed to accommodate the learning rules in current use, there is an element of discreteness to PDP representations, but the distinction between static representations and active algorithm has disappeared.[55]

One question that I left hanging in Section 3 concerned which of the two interpretation schemas was the correct one -- is the local interpretation schema, which isolates the content at the unit level, the correct one for identifying the representational states mentioned in mental causal laws, or is the distributed interpretation schema correct? I should note here that I am using the phrase "local interpretation schema" in a slightly different way than is usually encountered in the literature. (I do this, not for the mere sake of perversity, but because "local" versus "distributed" are, I think, intended as opposites -- whether used by me or within the literature. However, the standard meaning that has evolved for "local" in this context is not precisely the opposite of "distributed". I want to pair these two interpretation schemas off as exact opposites, so I must jettison one of the two standard usages. I have chosen to retain "distributed" in its standard sense and to change that of "local".)

[55] While it certainly lies outside the scope of this work to address the issue of the possibility of learning within continuously processing PDP networks, it is interesting to consider its potential ramifications for PDP. In that case, representational states would not follow upon one another in discrete time steps (as, for example, the integers follow upon one another when counting), but rather would flow continuously -- so that no particular representation could be truly said to follow upon another. In light of PDP´s modelling of the mind, this would translate over into the thesis that for (human) minds, representations are not discrete, isolatable entities.

The most often encountered usage of the epithet "local interpretation schema" assumes that there is some non-disjunctive unit level semantic content. Thus, the class of "acceptable" networks are those for which the unit level content is restricted to representations of objects, properties of objects, or microfeatures. The first two types of content are self-explanatory. "Microfeatures", on the other hand, is used to describe an object or property at a level lower than that at which commonsense objects or properties are described. Thus, if is-a-cup is a property, it may have microfeatures has-a-handle, is-made-of-porcelain, etc. *Micro*featurehood is thus relative to which level is identified as the surface or ground level of description. The restricted class of PDP networks for which a local interpretation schema can be given encompasses the correct mental model, on this view.

As already mentioned in Section 3, most PDP networks whose representational content is determined by training (as opposed to being selected and hand-coded by the researcher) fall outside of this class. The unit level represents a disjunctive object or property. Non-disjunctive objects or properties only emerge in the representational states implemented by multiple units. It is still possible, however, to identify the unit-level representations as those implementing mental causal laws. In this case, the form of mental causal laws would be radically different from what is normally assumed. I use the phrase "local interpretation schema" to encompass both of the above possibilities. Thus, the local interpretation schema is correct if the content adverted to in mental causal laws can be borne only by single units, irrespective of whether that content is disjunctive or not. I also mentioned in Section 3 that, under the rubric "pattern", I wanted to include "patterns" encompassing only a single unit. I remarked that whether such "patterns" are truly causally isolatable is an empirical matter. Allowing the word "pattern" to range over all possible non-empty subsets (both proper and not) of the units in a network is fairly standard in the literature, and the singletons are just as much subsets as are those with more members. The issue boils down to whether multiple unit patterns are even potentially the bearers of content adverted to in mental causal laws: proponents of the local interpretation schema say "no", whereas proponents of the distributed interpretation schema say "yes".

One point in favor of the local interpretation schema is its intuitiveness: units are easily identifiable and labellable. Patterns over many units, on the other hand, are harder to isolate for the purpose of discovering their representational content. Is there anything arguing in favor of the distributed interpretation schema? Before tackling this question, I must make the notion of "distributed" more precise.[56] The word is ambiguous. One of its senses is "spatially extended". Thus, a content is distributed if the physical stuff forming its representation occupies more than a point in space. Clearly this sense is not very useful, for all physical stuff occupies more than a point in space: all stuff has extension. Using this sense, physically realized representations a la traditionalism are likewise

[56] My starting place for this discussion is van Gelder´s paper "On Distributed Representation" in *Philosophy and Connectionist Theory*. While he makes many good points on the way to distinguishing "local" and "distributed", I think his final choice of criteria for distributedness relative to PDP networks misses the mark.

distributed. A second sense of "distributed" (the one preferred by van Gelder as the sense that best fits how the term is and should be understood within PDP) is applicable to a representation if the resources (in this case, units) used to realize it likewise participate in the realization of distinct representations. Thus, a network supports distributed representation if each unit participates in many patterns, each of which is a representation. Alternatively expressed, there is no mutually exclusive partitioning of the set of units separating the overall set into subsets, each of which is responsible for representing a content, such that all representable contents have their own subset. Van Gelder calls this sense of "distributed" "the superposition of representations".

While the distinction superposable/non-superposable is useful, in that superposability of representations is a feature often realized in PDP networks, I don´t think it gets at the heart of the issue, in light of the use of PDP as a model of the mind. Rather, I want to distinguish distributed from non-distributed based upon whether mental causal laws advert to contents borne (at least potentially) by multi-unit patterns. Thus, the important sense in which representations are distributed in PDP networks is that the network models a mental process whose causally interacting items correspond in the network to multi-unit patterns. The superposition of representation may be an additional feature, but it is not the crucial one in distinguishing the interpretation schema as local or distributed.

So now, we can return to the question: is there any reason to reject the local interpretation schema in favor of the distributed one? In Section 3, I mentioned two points of counterintuitiveness associated with the local interpretation schema. There is, in addition, slightly less anecdotal evidence that the (human) mind implements its representations in a distributed fashion. This evidence is not of a sort to yield an out-and-out disconfirmation of local representation; rather, it points to four mental phenomena which follow naturally, without the need for the introduction of special procedures, from distributed representation. The first piece of evidence is the ease with which humans generalize.[57] Stated broadly, generalization is the application of principles learned from experienced examples to novel examples. Usually, though, to say that a system can generalize is to imply that the choice of principle(s) to be applied makes sense -- that the system takes into account the similarities and differences between the previous examples and this novel one, and either chooses the correct among many principles, or adapts a learned principle in light of these similarities and differences. When a representation is spread over many units in a PDP network, there is a natural similarity metric available to compare two representations -- namely, the distance between the two vectors forming the representation. This "similarity check" occurs automatically -- there is no need for an outside agent to assign a similarity metric. To take a particular example, suppose that a network has been trained so that on input

[57] Strictly speaking, it is not distribution per se, but superposition of representations that explains ease of generalization. In rejecting van Gelder´s equation of "distributed" and "superposed", I was not rejecting the thesis that superposition is a property had by most mental representations, but rather the thesis that superposition is the defining feature of distributed representation in general.

I1 it produces output O1, and on input I2 it produces output O2, on input I3, O3, and on I4, O4. Now imagine that the network is presented with a novel input, I5, which, from the perspective of an outside observer, is similar to I1 and I2 in some respects, and similar to I3 and I4 in other respects, but dissimilar to all of the other inputs on which the network was trained. Again, from the point of view of the external observer, the "reasonable" way of handling this novel input is to produce an output, O5, that is similar in some repects to O1 and O2 (to the extent that O1 and O2 are similar) and similar in some respects to O3 and O4 (again, to the extent that O3 and O4 are similar). This high-level description of a "reasonable generalization" is just what a PDP network does. The perceived similarity between two input vectors must somehow be encoded in the actual vectors (else, why would they be called similar?). This similarity is automatically taken advantage of as the processing proceeds and the network output is computed. With a localized interpretation schema, there is *no* automatic generalization. Any generalization that may take place must be guided by a hand-coded procedure and/or a hand-coded similarity metric relating representations. If we take PDP seriously as a model of the mind, generalization within a local interpretation schema requires a homunculus who can look at and appraise the similarity between inputs (and between internal representations). While the idea of such a homunculus is not incoherent, I am working under the assumption that, other things being equal, an interpretation schema that does not require the existence of a homunculus to explain a fact about human mentation (namely, that we can generalize) is preferable to one that does.

A second feature of a distributed interpretation schema that provides evidence that human mental states are implemented in a distributed fashion is a modified version of the "graceful degradation" property of distributed networks. To say that a network´s performance degrades gracefully is to say that no individual piece of the network is so crucial that its loss or damage produces a marked decrease in the performance level of the network as a whole. A further aspect of graceful degradation is the gradual decline in performance with the loss or damage of parts of the system. Clearly, a local interpretation schema does not display graceful degradation: there may be a unit representing a key object or property (i.e., key to the level of performance of the system) such that its loss produces a drastic decline in performance. What about the case for a distributed interpretation schema? Here it is important to be clear on what graceful degradation means when applied to PDP qua mental model. I pause to note one thing in particular that it does *not* mean. (I make this explicit, because the PDP literature is rife with this mistaken understanding of graceful degradation.) It does not mean that units are like neurons, in that (as we know from empirical investigation) neurons die off every day without a noticeable decline in intellectual capacity of the individual. This mixes two distinct understandings of what the PDP project is about -- namely, modelling the mind versus modelling the brain. What I think the feature of graceful degradation within distributed representations has as important implication is that representational content can change in a piecemeal fashion. I am reminded of Stich´s thought experiment involving the woman who, over the course of time, gradually lost the ability to represent President McKinley.[58] Under a local interpretation schema, this is not a possibility: representations are atomic and either all there or all absent, with no inbetween states.

There are two further features of distributed interpretation that provide evidence in favor of it as the schema present in (human) minds. The first of these is the ease of implementing content-addressable memory within a distributed framework. Content-addressable memory is one in which items can be recalled based upon some part of the item. So, for example, when I try to recall a female acquaintance´s name by the "generate-and-test" procedure (i.e., think up a bunch of common female names and ask myself for each one: "is this her name?"), I am taking advantage of content-addressable memory. The "content" in this case is the hypothesized name, which I use to recall each person with whom I am acquainted who has that name, in order to see if the woman in question is among them. Content-addressable memory is easily implemented in a distributed network, but is implementable only with great effort in a local schema. A fourth piece of evidence pointing to a distributed interpretation schema as that used by human minds is the ease with which both humans and distributed networks can create new representations on the fly. I have given less emphasis to these latter two features of distributed representation because they present less conclusive evidence in

[58] See *From Folk Psychology to Cognitive Science*, pp. 54-56.

favor of a distributed interpretation schema as the correct one. I think, though, all in all, that the evidence points against a local interpretation schema as the correct one in constructing a model of the mind. Hence, I shall from here on assume a distributed interpretation schema.

One potential objection to a distributed interpretation schema that I would like to counter goes as follows: doesn´t this schema fall afoul of the locality constraint within PDP networks? In particular, doesn´t it implicitly posit an additional entity who is "looking over" the network to gather together the non-local information regarding which patterns are present? A simple reductio shows the hollowness of this objection. Suppose that instantiation of a mental causal law does require such an external observer to note that a particular contentful state has been tokened. (Assume that this contentful state forms the nomologically sufficient condition for some effect.) There is in this regard no relevant difference between mental causal laws and causal laws simpliciter, so there must likewise be an external observer to note when the antecedent to a causal law is satisfied, in order for that causal process to ensue. But this is clearly false. Therefore, no external observer is required to gather together the non-local information constituting the distributed representation. Therefore, distributed representation per se does not violate the locality constraint on PDP.

I end this section with an examination of several issues which will surface again in Chapter 5 as points of comparison with traditionalism. The first question asks: what ontological commitments are inherent in PDP as a model of the mind? PDP presupposes a physicalist metaphysics, while also maintaining the existence of causally efficacious mental states, identified in terms of their semantic content. Mental causal laws advert to the content of these states.

Looking closer at the form that these mental causal laws will take, we see that the set of possible mental causal laws is constrained by the nature of PDP processing. If we take PDP seriously as offering a mental model, this translates into the statement that the manner in which one representational state in a PDP network can follow upon another reflects the manner in which one mental state can cause another. (Recall that my analysis is not at all concerned with what particular mental causal laws there are, but rather with what form is assumed by, and what restrictions are placed upon mental causal laws in accordance with the framework provided by either traditionalism or PDP.) I will take up this topic in more detail in Chapter 5, where I attempt a point-by-point comparison of the constraints on mental causal laws offered by the two paradigms.

Finally, what level of reality is represented by the causally efficacious mental states according to PDP? As already argued, I think that the most promising interpretation schema for use within PDP is the distributed interpretation schema, according to which the causally efficacious states have non-disjunctive content. While there is no argument within PDP that plays the same role with respect to implying exactly what level of reality is represented by these causally efficacious states as we see in the LOT argument within traditionalism, some general comments on this topic are possible. A review of the literature quickly confirms the view that PDP does not differ much from traditionalism on this score. After training, researchers analyze the network by trying to identify regularities in the succession among internal patterns, and in the relationship between inputs and internal patterns. The labels attached to causally efficacious patterns correspond in most instances to concepts easily expressed in natural language. (Either objects or properties.) There are, however, many renegade trained networks that have yet to succumb to this analysis.[59] For such networks, researchers can identify no consistent mapping between causally efficacious patterns and such easily expressible concepts -- yet, the networks succeed in achieving a high level of performance at the task at hand. I am not quite sure what to make of such networks. One can assume that such a mapping exists, but, because it is so complex, it has not yet been discovered. Contrarily, one can take this as a sign that a full-blown model of the mind may likewise not have mental states whose content is easily expressible in natural language.

[59] The most famous example is the mine identifier system of Gorman and Sejnowski.

CHAPTER 5

ARE THE TRADITIONALIST AND PDP MODELS QUALITATIVELY DISTINCT?

In this concluding chapter, I have set before myself several tasks. First, I develop a general framework for comparing two models of a domain. This framework is general in the sense that it specifies criteria to be used in judging qualitative distinctness, irrespective of the particular domain being modelled. This topic is addressed in the first section.

One often hears PDP referred to as a "new paradigm" for understanding mental phenomena, and the transition within cognitive science (at least, with respect to emphasis in professional meetings and journals) from the traditional to the PDP model of the mind as a "paradigm switch" or "revolution" within the field. The allusions to Kuhn´s theory of scientific change have led me to consider the questions: Can we understand traditionalism and PDP as forming the kernel of disparate paradigms, and are they incommensurable? I examine this issue in the second section. In doing so, I highlight the differences between Kuhn´s "incommensurability" (a concept which is, I think, never fully developed in his *The Structure of Scientific Revolutions*) and my "qualitative distinctness".

One famous exchange in the traditionalism versus PDP literature (see Smolensky´s "On the Proper Treatment of Connectionism" and Fodor and Pylyshyn´s "Connectionism and

Cognitive Architecture: A Critical Analysis") that addresses this topic bears attention. I therefore devote Section 3 to a summarization of the views of these two opposing camps. While they are suggestive, I believe that the arguments put forward by both sets of authors (indeed, by all of the authors who have written on this topic) fall short of the mark. As I have already repeatedly mentioned, the construal of one object (or object-type) as an explanatory model of another is possible only in the context of a theory of causation. When such an articulated theory is absent (whether because it is tacitly assumed or because it is wholly lacking), the analysis of something´s modelhood as well as the comparison of models becomes highly problematic. In each of the above cases, the authors fail to provide the necessary causal theoretic background.

In Section 4 I consider issues relating to computability. This discussion is included to thwart the superficially plausible argument that traditionalism and PDP as models of the mind must be distinct, because the two corresponding abstract machines (namely, the computer and PDP networks) differ in their computational power.

Finally, in Section 5 I give my answer to the question: Are the traditionalist and PDP models of the mind qualitatively distinct? Briefly, my argument takes the following form. While both models describe the mental level, and both make similar ontological commitments, there is no possible isomorphism between the web of causal laws permitted within the constraints of the respective models. Hence, the two *are* qualitatively distinct.

5.1 What is Qualitative Distinctness?

As remarked above, one often meets in the literature surrounding the traditionalism versus PDP debate the assertion that the two camps are proposing distinct theories of the mind. Unfortunately, these assertions are usually left at the level of vague generality, because they are made outside the context of any worked-out explication of what it means for two theories to be distinct. As we shall see in Section 2, even Kuhn fails to give more than the briefest of sketches in describing what criteria distinguish genuinely incommensurable theories from those that merely differ with respect to adopted vocabulary. In this section, I propose my own set of criteria for use in determining whether two theories are qualitatively distinct.

Before beginning that task, I pause to give reasons for my choice of the descriptor "qualitatively distinct". Scientific theories can be distinct in many ways. For example, two theories are distinct if they make differing predictions about future events given the same initial conditions. This divergence in and of itself need not reflect an underlying *qualitative* distinctness between the two theories, for such divergence can result if the two theories merely differ in respect of some value of a parameter. To be more specific, the divergence in prediction that results when two otherwise identical theories of relativistic mechanics differ with respect to their values for the speed of light (say, 2.9x10*17* m/s versus 3.0x10*17* m/s) does not constitute a *qualitative* difference between the two theories. Similarly, two theories that make identical predictions (about both observable and non-observable states of the system), given the same initial conditions are not qualitatively distinct, even though the vocabulary that each employs to identify the objects and states quantified over by its causal laws may differ radically.[1] In the above two respects, my application of qualitative distinctness does not differ much from Kuhn´s incommensurability. For him, two formulations do not constitute incommensurable paradigms either when one is a mere quantitative refinement of the other, or when the two formulations support a ready translation between themselves. I do not, however, adopt his "incommensurability" for several reasons. First, the notion is never clearly defined in his work. If this were my only objection, however, I could view my work as a natural extension and specification of his own. A more important reason for rejecting his "incommensurability" is my desire to distance myself from some of the baggage that comes along with that term. In particular, I (unlike Kuhn) *do* believe that there can be (rationally defensible) reasons for preferring one paradigm over another. Note here my use of the rather weak "can be" over the much stronger -- and, I think, unjustified -- "shall be". I think that for Kuhn, even the "can be" is too strong. While he explicitly rejected the accusation that his view

[1] This statement will get me into trouble with anyone who rejects the possibility of a theory-neutral language of description. However, as will become clear later in this section, I mean here to exclude from the extension of the set of obviously qualitatively distinct pairs of theories only those pairs permitting the most superficial mappings between their respective terminologies.

of science turns it into a "subjective" and "irrational"[2] enterprise,[3] still, a consequence of his view is that any attempt at arguing for one against another of two incommensurable paradigms will be necessarily circular.[4] I shall have much more to say on Kuhn´s "incommensurability" in the next section.

On my use of the term, qualitative distinctness takes in two aspects: respective ontological commitment and respective decomposition of phenomena into causal sequences. In posing the question "what ontological commitments are made by a particular theory?", I am presupposing that that theory takes a realist stance towards the objects and states quantified over in its causal laws. So, this question is transformed into: "what things must exist on the assumption that this theory correctly subdivides the world (or, at least a level of causal interaction within the world -- more on this later) into its causally efficacious parts?" Answering this question is in general a very difficult task, for several different reasons. The most obvious is that theories do not wear their ontological commitments on their shirt-sleeves: rarely does a researcher or theoretician give explicit declarations regarding what assumptions are and are not being made within a scientific theory. As Newton-Smith remarks, this tendency toward silence on the part of scientists is not a new phenomenon:

> In examining scientific theories for ontological commitment, it will not usually be such a trivial matter.

[2] I think he meant "arational".

[3] See especially his Postscript to the second edition of *The Structure of Scientific Revolutions*, pp. 191-198.

[4] *The Structure of Scientific Revolutions*, page 94.

> For instance, it remains as controversial today as it was at the time for Leibniz and Newton whether theories of time carry a commitment to the existence of moments of time over and above collections of events.[5]

In performing an analysis of the ontological commitments made by a particular theory, therefore, one must do some interpolation.

A second source of difficulty in teasing out ontological commitments (particularly relevant when the aim of this analysis is an inter-theory comparison) is that objects and/or their states that are hypothesized within one theory to be causally efficacious entities qua singletons may appear in the other theory only as one part of a unit. In the latter case, only the unit (i.e., that singleton entity plus the other singleton entities with which it is conjoined) is causally efficacious. In this case, would one say that both theories are committed to the existence of that entity? I think not. My reason for denying this is that the latter theory does not recognize *the singleton* as alone causally efficacious, even though the terminology standardly used by practitioners of that theory have a word that picks out that singleton. (This is particularly prevalent when the theory that has the singleton being causally efficacious in isolation of other facts about the world temporally precedes the theory that has the singleton being one part of the true causally efficacious entity.) The classic example of the difficulty is provided by a consideration of whether Newtonian (classical) and Einsteinian (relativistic) mechanics are both committed to *mass* as a causally efficacious

[5] Newton-Smith, *The Rationality of Science*, page 38.

property.[6] A problem arises because the most intuitive construal of "mass" as used by Newtonians equates to "rest mass" in the terminology of Einsteinians, but rest mass is not alone the causally efficacious property within that latter theory. Rather, it is the conjunction of the rest mass of an object and a measure of the velocity of that object relative to the speed of light that is causally efficacious. Setting aside for the moment the fact that the phrase "rest mass" was never used within Newtonian mechanics, would we still want to say that both theories are committed to the reality of rest mass as a causally efficacious property? As noted above, my answer is "no": Newtonian mechanics is committed to the reality of rest mass, but Einsteinian mechanics is not.

A third source of difficulty in comparing the ontological commitments of two theories is that, while the same word is employed within both theories to pick out an object, many of the causal interactions in which the object can participate according to one theory are not recognized by the other, and vice versa. A case in point is determining the import of the sentence: "this theory is committed to the existence of light" when made with reference to a corpuscular versus a wave theory of light. Clearly, many of the properties possessible by light in the one theory are not recognized by the other. Does this mean that the two theories are ontologically committed to different things -- namely, light-qua-particle for the corpuscular theory and light-qua-wave for the wave theory? I think so, for the causal efficacy of the two "types" of light differ. For the

[6] Put in terms of the causal efficacy of objects and states, this is the same as asking whether both theories recognize "having-mass-x" as a causally relevant state of an object.

wave theory (but not for the corpuscular theory), light must exist as something having a particular wavelength: the state of having-wavelength-x is a state of a beam of light that must exist if the wave theory correctly describes the world. In a sense, the above example shows the difficulty of teasing apart the ontological commitment aspect of a theory from the set of causal laws propounded by a theory: the wave theory is committed to light-qua-wave because its causal laws mention possible states of light (e.g., its wavelength) that presuppose that it is a wave.

Choosing an example closer to the theme of this work, consider whether the ontological commitments made by Freud´s psychological theory and by folk psychology prior to Freud are the same. What must exist if the former is a true depiction of the world? Clearly, mental agents with various attitudes (beliefs, desires, etc.) towards representations must exist. The causally efficacious items are the conjunction of representational content plus attitude. (Thus, the belief that I will receive a raise has a distinct causal efficacy from the desire that I will receive a raise, which in turn has a distinct causal efficacy from the desire that I eat French fries.) This much is uncontroversially shared with folk psychology. A possible point of divergence crops up in considering whether the fact that some of these causally efficacious states are unconscious for the Freudian theorist adds something new to the ontology of Freudian theory not found in folk psychology. (Note: Although I am certainly no expert on Freudian theory, I am assuming that the unconscious itself is only metaphorically causally efficacious within that theory: what are causally efficacious are beliefs, desires, etc., some of which are unconscious. For a fuller discussion, see the section "Freud and the Unconscious", in *The Construction of Reality* by Arbib and Hesse, pp. 114-117.) This boils down to the question: does Freudian theory recognize unconscious-beliefs, unconscious-desires, etc., as distinct entities (vis-a-vis causal relevance) from conscious-beliefs, conscious-desires, etc.? I don´t think a definitive answer can be given, because Freudian theory is not a monolithic theory, but rather a group of schools of thought. The version of Freudian theory presented by the school most at home with cognitive psychologists does not give unconscious mental states a distinct status. The case is perhaps otherwise for other schools within the broad Freudian tradition -- I just don´t know. The question concerning distinctness of ontological commitment is much more clear-cut when the two rival theories are folk psychology and behaviorism. The latter makes no place for the causal relevance of beliefs, desires, or any other representational states; hence, the two theories make differing ontological commitments.

A second aspect of qualitative distinctness involves sameness of causal interaction across the two theories. This aspect is in a sense secondary to that of ontological commitment, for it presupposes sameness of ontological commitment. Where two theories have well worked-out sets of causal laws, this comparison is (conceptually, at least) straightforward. First, pair off the objects and states referred to in one theory with the corresponding objects and states in the other. (Again, I reemphasize that this step is predicated on the existence of a correlation between the causally efficacious objects and states posited by the two theories.) One can think of causation as a relation over sets of states. So, for example, the causal law that says *a&b&c* causes *d&e* relates the set containing *a*, *b*, and *c* to the set containing *d* and *e*. Now, the two theories are distinct with respect to causal interaction when the relation thus defined by the one listing of causal laws is not isomorphic to the other listing under the mapping equating the causally efficacious objects and states in the one theory with the corresponding objects and states in the other.

This version of distinctness is, however, not exactly what we are looking for, as it is too quick to label two theories as distinct. For example, it labels as distinct two versions of relativistic mechanics that differ only in their respective approximations of the speed of light. This, I think, slices the world of scientific theories too finely. A coarser slicing (one more befitting the epithet "qualitative") requires some additional distinctions within each respective theory. In particular, each theory would group its causally efficacious states into relatively quantitatively similar sets. The two theories are *qualitatively* distinct with respect to causal interaction when the relation defined by the listing of causal laws is not isomorphic to the other listing under the mapping using the coarser grained sets of quantitatively similar objects and states. There are two points to note on this refined version of distinctness. First, it fails to distinguish between two theories that differ only in non-qualitative ways. This is more than an empty truism, for the ultimate determiner of what differences are quantitative and what differences are qualitative is relative to the two theories being compared. By specifying the quantitatively similar groupings, each theory implicitly says: "any theory that keeps within these parameters will be viewed by me as the qualitatively same theory." Secondly, this understanding of qualitative distinctness can distinguish theories that differ in quantitative ways, where those quantitative differences constitute qualitative differences. For example, a relativistic theory of mechanics that allows particles (whether massive or not) to travel above its approximation of the speed of light may be judged as qualitatively distinct with respect to causal interaction from one that disallows such fast moving particles relative to its approximation of the speed of light. Thus, the "mere quantitative" difference of 1 m/s, when that difference occurs at the cusp separating sub-light from supra-light speeds, may constitute a qualitative difference.

The above characterization of my method for determining qualitative distinctness with respect to causal interaction requires a general remark. For those theories wherein the cardinality of the set of causally efficacious objects and states is equal to that of the set of real numbers, the problem of the inability to enumerate all of the causal laws crops up: if you can´t enumerate all of the causal laws, how can you possibly compare the one listing with the other? My response is to shrug my shoulders and note that my method, while conceptually straightforward, presents some difficulties in implementation.

Considering my method in light of the task at hand (namely, comparing the theories presented by traditionalism and PDP, to be discussed in Section 5) points to another potential problem: both theories are too young to have a well worked-out set of causal laws.

In particular, the model driving the theory in each case is merely such as to place constraints on the possible causal laws, it does not dictate particular causal laws. Thus, in applying my method, I will have to examine, not whether the two causal relations permit an isomorphism, but whether the constraints placed on the possible causal relations leave open the possibility of an isomorphism. If not, the two theories are qualitatively distinct. This lack of conclusiveness (in that the failure to discover the impossibility of a possible isomorphism does not entail that the two theories are not qualitatively distinct -- failure to prove that *p* does not imply that *not-p*) is tolerable, for my over-arching goal - the question that has spurred my interest in this subject - involves whether the accusation that PDP is just the same old thing (i.e., traditionalism) with some updated vocabulary is true. At a minimum, the analysis to be performed in Section 5 should settle that question.

Before leaving this section, I pause to give a high-level summary of my method for determining qualitative distinctness, and to characterize the two (grossly described) "flavors" in which qualitative distinctness comes. The first stage of testing for qualitative distinctness involves asking the question: "do the two theories make the same ontological commitments?" If the answer is "no", then the matter is settled: the two theories *are* qualitatively distinct. If, however, the answer is "yes", then one must continue the analysis to include a consideration of the causal interactions posited by the respective theories. When the two are qualitatively distinct with respect to causal interaction, then they are qualitatively distinct, simpliciter. Otherwise, they are qualitatively indistinct.

Described at a high level of abstraction, qualitative distinctness comes in two "flavors": (1) same level but no or little overlap, and (2) different levels. The qualitative distinctness of two theories in the latter flavor is perhaps easiest to see, for the level of reality (viewing the world as a quasi-hierarchy of causally interacting objects and states) encompassed by each differs. Consider current quantum mechanics (or, more precisely, consider one version of current quantum mechanics) and modern cellular biology. These two scientific theories are clearly qualitatively distinct, because the ontological commitments are so radically different in the two respective theories. Modern cellular biological laws quantify over cells, membranes, and their states. Thus, modern cellular biology is committed to the existence of these objects and states. However, there is no mention whatsoever of these objects and states within modern quantum mechanics. In general, disparate scientific disciplines have qualitatively distinct theories; this distinctness is reflective of differences with respect to ontological commitment within the two theories.

What, though, of the case in which it is not obvious that the two theories are describing different levels of reality, either because the two do indeed describe the same level, or because there is disagreement among the proponents of one or both theories with respect to the level of reality being described? (This latter possibility is particularly relevant to the issue at hand in light of the lack of consensus within PDP regarding what their model is a model of.) Returning to the previous example of the corpuscular versus wave theory of light, we see a case in which the two theories do describe the same level of reality (namely, the behavior of light), yet they share little in common in terms of either ontological commitments or causal interactions. The classic example of the overthrow of one theory by another within a scientific discipline fits this mold. Those proponents of PDP who are fond of describing their model of the mind as a new paradigm usually have such an understanding of the relationship between traditionalism and PDP.

5.2 Kuhn´s Theory and the Relationship between Traditionalism and PDP

Such references to the terminology in Thomas Kuhn´s *The Structure of Scientific Revolutions* leads me naturally to ask several questions. Is the usage appropriate: is psychology in the midst of a revolution pitting traditionalism against PDP? Also, is my "qualitative distinctness" just his "incommensurability"? If not, what are the points of divergence between the two?

Kuhn lays out the typical development of a scientific revolution as seen in historical case studies of transitions that, in retrospect, are clear instances of paradigm changes. First, a particular discipline is united around a single paradigm in the process of doing normal science. Practitioners in the field are occupied with fleshing out some aspects of the paradigm not yet fully concrete and solving puzzles (i.e., results not predicted by the paradigm as it currently stands) within the context of the paradigm. If a sufficient number of puzzles prove to be recalcitrant with respect to being explained with the (perhaps slightly modified version of the) paradigm, a crisis situation develops. (Puzzles not thus explainable are called "anomalies".) There is growing discontent among some of the members of that discipline, particularly the younger ones, who have less stake in the maintenance of the old paradigm, leading eventually to the feeling among some members that the "existing paradigm has ceased to function adequately in the exploration of an aspect of nature to which that paradigm itself had previously led the way."[7] Eventually the discontented group congeal around a rival paradigm, and the battle for allegiance within the previously unified group commences. The allegiance centers around three sorts of commitments that a scientist derives from her paradigm:

> Less local and temporary, though still not unchanging characteristics of science, are the higher level, quasi-metaphysical commitments [which follow from the acceptance of a paradigm] that historical study so regularly displays. ... That nest of commitments proved to be both metaphysical and methodological. As metaphysical, it told scientists what sorts of entities the universe did and did not contain. ... As methodological, it told them what ultimate laws and fundamental explanations must be like. ... More importantly still [a paradigm] told scientists what many of their research problems should be.[8]

The two rival paradigms are not only incompatible (they must differ on some of their predictions, in that the new paradigm is put forward in response to and as supplying an explanation of the

[7] *The Structure of Scientific Revolutions*, page 92.
[8] *The Structure of Scientific Revolutions*, page 41.

results that were anomalous within the context of the old paradigm), but may also be incommensurable. Kuhn breaks incommensurability into its three aspects:

> (i) "[T]he proponents of competing paradigms will often disagree about the list of problems that any candidate for paradigm must resolve. Their standards or their definitions of science are not the same."[9]
> (ii) "Since new paradigms are born from old ones, they ordinarily incorporate much of the vocabulary and apparatus, both conceptual and manipulative, that the traditional paradigm had previously employed. But they seldom employ these borrowed elements in quite the traditional way. Within the new paradigm, old terms, concepts, and experiments fall into new relationships one with the other. The inevitable result is what we must call ... a misunderstanding between the two competing schools."[10]
> (iii) "[Most fundamental:] the proponents of competing paradigms practice their trades in different worlds."[11]

This incommensurability between paradigms makes rational discourse concerning the relative merits of the two paradigms impossible. Members of the opposing camps often find themselves "talking past" one another, because the meanings of the terms that they use are paradigm-relative, and because "they cannot ... resort to a neutral language which both use in the same way."[12] Arguments for or against a particular paradigm must be persuasive (rather than

[9] *The Structure of Scientific Revolutions*, page 148.
[10] *The Structure of Scientific Revolutions*, page 149.
[11] *The Structure of Scientific Revolutions*, page 150.
[12] *The Structure of Scientific Revolutions*, page 201.

rational) in nature. As enough of the members of the field come to adopt the challenger paradigm (either because people *are* actually persuaded, or because the scientists who remain committed to the older paradigm die off), the challenger attains the role of accepted paradigm and normal science commences within that field again, albeit around a new paradigm.

Kuhn describes the process by which an individual scientist becomes committed to the new paradigm as akin to the gestalt switch that occurs in the oft-cited duck/rabbit picture: it is a quantum experience not further decomposable into substages. Kuhn describes the transition from a field in crisis to the field again unified around a new paradigm as:

> a reconstruction of the field from new fundamentals, a reconstruction that changes some of the field´s most elementary theoretical generalizations as well as many of its paradigm methods and applications. ... When the transition is complete, the profession will have changed its view of the field, its methods, and its goals. One perceptive historian, viewing a classic case of a science´s reorientation by paradigm change, recently described it as "picking up the other end of the stick," a process that involves "handling the same bundle of data as before, but placing them in a new system of relations with one another by giving them a different frame."[13]

With this description of the Kuhnian view of scientific revolution in hand, we can consider the question: is the current state within cognitive science one of crisis/revolution, with

[13] *The Structure of Scientific Revolutions,* page 84-85. Quote from H. Butterfield´s *The Origins of Modern Science, 1300-1800,* pp. 1-7.

traditionalism as the old paradigm and PDP as the challenger? My approach in the next several pages will be to give a depiction of the emergence and development of PDP within cognitive science that is as sympathetic as possible to the view that takes PDP as a new paradigm within that field. (As already hinted at previously, this view predominates among PDP researchers themselves, whereas the prevailing view among adherers to traditionalism is that cognitive science is not in a crisis situation, and the puzzles that are not yet explainable within the framework of traditionalism will, with further research, eventually succumb.)

Traditionalism came to dominate psychology after the overthrow of behaviorism. It was the accepted paradigm, and the normal scientific phase of research within psychology during the last 30 years or so assumed it, as can be seen in researchers´ acceptance of its (1) ontological commitments, (2) methodological presuppositions (in the form that causal laws and causal explanations within psychology would take) and (3) depiction of the sorts of phenomena that a theory within psychology should be able to explain. Thus, the work of psychologists during this time period consisted in fleshing out the particulars within the traditionalist framework (e.g., performing experiments to determine what the individual mental causal laws are) and making minor adjustments within the framework in order to solve the outstanding psychological puzzles.

Many puzzles did indeed prove to be explainable within traditionalism; however, many remained (and continue to remain) recalcitrant. Some of the anomalous mental phenomena have already been described in Chapter 4. One such example is the inability of traditionalism to explain how content-addressable memory is achieved in the mind, given empirical data on access time required to recover a particular item. Some researchers within psychology (particularly the newer ones to the field, who either had just completed their professional training or had just gained an interest in psychology after work in some other field -- most often, neuroscience or computer science) began placing more emphasis on the failures of traditionalism (i.e., the anomalies) than on its successes. The "youth factor" among converts to PDP is quite apparent, and is explained by the fact that part of the reluctance of the older members of the psychological community to give up on traditionalism relates to their professional commitment to that paradigm. They continue to see only puzzles to be solved within the framework of traditionalism, whereas the younger generation is much more willing to brand the "as yet not explained" mental phenomena as "anomalies" and to reject traditionalism as itself inadequate. These discontents have by and large congealed around PDP as a rival for the allegiance of psychologists. At a minimum, they contend, PDP can explain what has heretofore remained unexplained within traditionalism:

> [PDP holds] out the hope of offering computationally sufficient and psychologically accurate mechanistic accounts of the phenomena of human cognition which have eluded successful explication in conventional computational formalisms [ie, traditionalism].[14]

[14] Rumelhart and McClelland, *Parallel Distributed Processing*, page 11.

One can see in some of the debates that have occurred between adherents of traditionalism and adherents of PDP the tell-tale signs of incommensurability between the two theories. The clearest example of this is the denial that the sorts of problems that PDP systems are particularly good at solving are even within the purview of psychology. Thus, we see Fodor and Pylyshyn (the truest of believers in traditionalism) unintentionally illustrating this very aspect of the crisis situation within psychology today in their denial of the successes of PDP qua model of the mind:

> We have, in short, no objection at all to [PDP] networks as potential implementation models, nor do we suppose that any of the arguments we´ve given are incompatible with this proposal. The trouble is, however, that if connectionists do want their models to be construed this way, then they will have to radically alter their practice. For, it seems utterly clear that most of the connectionist models that have been proposed must be construed as theories of cognition, not as theories of implementation. This follows from the fact that it is intrinsic to these theories to ascribe representational content to the units (and/or aggregates) that they postulate. And, as we remarked at the beginning, a theory of the relations among representational states is ipso facto a theory at the level of cognition, not at the level of implementation. It has been the burden of our argument that when construed as a cognitive theory, rather than as an implementation theory, connectionism appears to have fatal limitations. *The problem with connectionist models is that all the reasons for thinking that they might be true are reasons for thinking that they couldn´t be psychology.*[15]

[15] "Connectionism and Cognitive Archtitecture: A Critical Analysis", page 66, italics added.

Similarly, some of the terminology often seen within traditionalism has been retained within PDP, but with a different meaning. A case in point is the word "representation". The reader likely noted this in Chapter 4. While the broad definition of "representation" as "one thing that stands in for another" is retained, many of the particular defining features of representations change from traditionalism to PDP. This, in turn, produces radical changes in the very notion of what processing representations means. For example, RumelharT and McClelland note that for traditionalism:

> [t]here is no real difference between the stored representation in long-term memory and the active representation in working memory. In PDP models, though, this is not the case. In these models, the patterns themselves are not stored. Rather, what is stored is the *connection strengths* between units that allow these patterns to be re-created.[16]

The implication of this view of representation for cognitive processing is that:

> [u]sing knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear; it is part and parcel of the processing itself.[17]

[16] *Parallel Distributed Processing*, page 31.

[17] *Parallel Distributed Processing*, page 32.

Also, the relevance of some previous experiments within psychology has been reevaluated within PDP. Again, Rumelhart and McClelland make this point explicitly.

> [T]hese same mechanisms [associated with PDP] exhibit emergent properties which lead to novel interpretations of phenomena which have traditionally been interpretted in other ways.[18]

One way of describing the advent of PDP is as provider of a whole new way of picking out and slicing up (into its causally efficacious parts) the mental world, such that the very notion of "mental" has been transformed. Two quotes, again from Rumelhart and McClelland, bear proof of this transformation.

> [PDP has] radically altered the way we think about the time-course of processing, the nature of representation, and the mechanisms of learning.[19]

And

> This is a profound difference between our approach and other more conventional approaches, for it means that almost all knowledge is *implicit* in the structure of the device that carries out the task rather than *explicit* in the states of units themselves. Knowledge is not directly accessible to interpretation by some separate processor, but it is built into the processor itself and directly determines the course of processing. It is acquired through tuning of

[18] *Parallel Distributed Processing*, page 13.
[19] *Parallel Distributed Processing*, page 13.

> connections as these are used in processing, rather than formulated and stored as declarative facts.[20]

One interesting aspect of the traditionalism versus PDP debate is that, while some participants in the debate (most noteably, proponents of PDP) relish seeing their theory as a challenger in a scientific revolution sweeping over psychology, and often use Kuhn´s terminology in describing their theory as a new *paradigm*, they just as often fail to take note of some of Kuhn´s other remarks concerning the evolution of science. In particular, they fail to notice several key aspects of his view of scientific revolution with respect to the nature of discourse between the proponents of the two clashing theories. Thus, one often finds in the PDP literature an attempt at arguing that PDP is (objectively considered) the better of the two theories at explaining psychological phenomena. This flies in the face of Kuhn´s assertion that rational discourse on the relative merits of two incommensurable paradigms is impossible: to the extent that argumentation for or against a paradigm is possible, it will be persuasive (rather than rational) in nature. Thus, the adoption of a Kuhnian construal of their situation undercuts a second important working premise of the most vocal PDP enthusiasts: namely, that their theory is the best.

This sketch of the history of the rise of PDP-as-challenger-paradigm leaves us still in the midst of a crisis within psychology. Certainly it is premature to say either that PDP has become the new paradigm around which future normal science within psychology

[20] *Parallel Distributed Processing*, pp. 75-76.

will work or that a crisis-like mentality within psychology has been eased by the solution of some potentially anomalous counterexamples to traditionalism. Thus, it is still too early to say whether the final stage of the revolution (i.e., the congregation of the members of the psychological community around PDP, as a result of either persuasion or a dying-off of the traditionalist generation) will be reached. Clearly, it is the view of some PDP adherents that this paradigm change will eventually occur (or, at least, that it is a goal of the movement), as is illustrated by the following:

> We wish to replace the "computer metaphor" as a model of the mind with the "brain metaphor" as model of mind.[21]

With respect to the "gestalt switch" between paradigms predicted by Kuhn, not all first-hand accounts lend support to this aspect of Kuhn´s theory. For example, Rumelhart and McClelland describe their piecemeal acceptance of PDP as a model of the mind.

> The idea began to seem more and more attractive to us as the contrast between our convictions about basic characteristics of human perception, memory, language, and thought and the accepted formal tools for capturing mental processes became more apparent.[22]

Interestingly, the gestalt switch aspect of a paradigm change at the level of the individual researcher is highly contested within the

[21] *Parallel Distributed Processing*, page 75.
[22] *Parallel Distributed Processing*, page ix.

literature that has grown out of *The Structure of Scientific Revolutions.* Some authors maintain that, as historical fact, it is mistaken. For example:

> The shift in world view associated with paradigm changes are likened to the sort of gestalt switch one may have when, having first seen the notorious duck-rabbit as a duck, one suddenly sees it as a rabbit. By and large this analogy is absurdly far-fetched. For few of us had anything like this dramatic shift of attitude when, having learned Newtonian mechanics in school, we came slowly and perhaps painfully to appreciate the greater virtues of Einsteinian mechanics.[23]

Newton-Smith continues on to argue that such a gestalt switch phenomenon, if true, would undercut some of Kuhn´s own theses. Even Kuhn, in some of his examples,[24] contradicts his prediction of a gestalt type phenomenon on the part of scientists. Thus, the lack of concurrence within the PDP literature on this point is not surprising.

Hopefully, the above description of Kuhn´s "incommensurability", both in general and as potentially applicable to the situation relating traditionalism and PDP as competing theories, has given the reader a better understanding of how a Kuhnian may try to interpret the import of "qualitative distinctness", particularly as applied to traditionalism and PDP. I would like to contrast that with my own interpretation of "qualitative distinctness", both as an aid to seeing the differences between my

[23] W. H. Newton-Smith, *The Rationality of Science*, page 118.

[24] See especially his description of the Priestley/Lavoisier debate. *The Structure of Scientific Revolutions*, pages 54-56.

view and Kuhn´s, and, more importantly, as an aid to gaining a better understanding of my "qualitative distinctness", simpliciter. I begin by noting that Kuhn never gives any hard criteria (or even hints) for deciding when two theories are incommensurable. All of his case studies describing incommensurability have been historical ones in which, in retrospect, it is clear that a major shift within a scientific discipline has occurred. It is not sufficient to leave the matter at the level of generality at which he speaks. While his characterization of incommensurability is certainly fruitful, for the purpose at hand I need concrete criteria for comparison of two competing theories. Lacking such criteria, it is not even an option on my part to compare traditionalism and PDP for possible incommensurability.

It may even be argued that my setting up of criteria for cross-theoretic comparison refutes the claim that they are indeed incommensurable.[25] However this may be, it is clear that my views

[25] I do not want to turn this section into a commentary on *The Structure of Scientific Revolutions*, but a few words on one of the more contentious aspects of that work are in order. The most straightforward reading of the first edition would have Kuhn disallowing the very possibility of cross-theoretic comparisons, when those two theories are incommensurable. (I henceforth call this the "strong incommensurability thesis".) According to this view, there can never be any such comparison, because there is no neutral language within which to perform the comparison. Thus, my posing of the question: "does traditionalism make the same ontological commitments as PDP?" in the process of deciding qualitative distinctness is illegitimate, for it assumes the existence of something (namely, a neutral language) which does not exist. In his Postscript (added to the second edition as a response to many criticisms of the first edition), however, he seems to take back this strong incommensurability thesis and replace it with something else. What that something else *is* is not quite clear, but his statements regarding the possibility of inter-theoretic translation (see page 202.) show that it cannot be the strong incommensurability thesis. I think the most generous reading is that incommensurability a la the second edition is merely the claim that arguments in favor of one theory against another will inevitably be merely persuasive in nature, as they will use as premises comparison measures based on aesthetic considerations (eg, simplicity).

(both those stated in this chapter and in the remainder of this work) are at odds with those of Kuhn. In particular, I interpret simplicity and strength criteria as not "merely aesthetic", but as providing the basis for the truth or falsity of causal laws. My combination of a Lewisian version of lawhood with a realist construal of causation (admittedly unorthodox, but not, I think, inconsistent) is at odds with Kuhn´s relativism, which explains our differing views on the possibility of rational arguments for one theory against another. As I have already often mentioned, I am not in the least here concerned with a comparison of the relative merits of traditionalism and PDP, so I shall not develop this point of divergence further.

5.3 Some Answers Given by Others

Most often, the subject under discussion in the traditionalism versus PDP debate is not whether the two are qualitatively distinct, but rather which of the two provides the best explanation for mental phenomena. Such arguments obviously presuppose *some* sort of distinctness between the two theories; else, how could one be better than the other? Whether the distinctness presupposed is my qualitative distinctness is another matter. In this section, I shall examine what is perhaps the most famous (or, at least, the most often cited) exchange in the debate between Fodor and Pylyshyn (representing traditionalism) and Smolensky (representing PDP) as to the relative merits of the two theories. My focus will not be so much empirical evidence cited for or against either theory, but rather the assumptions (both implicit and explicit) in the argumentation of each set of authors with respect to the qualitative distinctness of traditionalism and PDP. I shall also mention some of the other positions vis-a-vis qualitative distinctness found in the literature.

I begin with Smolensky´s treatment of the relationship between traditionalism and PDP. Smolensky states right from the start that the two theories are distinct:

> A set of hypotheses is formulated for a connectionist approach to cognitive modeling. These hypotheses are shown to be incompatible with the hypotheses underlying traditional cognitive models.[26]

And:

> ...the level of cognitive analysis adopted by the subsymbolic paradigm [ie, PDP] for formulating connectionist models is lower than the level traditionally adopted by the symbolic paradigm.[27]

What remains to be done is to understand the specific points of departure between the two theories, and to examine whether his notion of distinctness is the same as or entails qualitative distinctness.

The remark in the second quote concerning a difference in levels is telling, for, as noted in Section 1, a difference in levels is evidence for a difference in ontological commitment with respect to the causally relevant entities. The issue is not so clear, however, for

[26] "On the Proper Treatment of Connectionism", page 1.
[27] "On the Proper Treatment of Connectionism", page 3.

Smolensky is not always so consistent in identifying the PDP model of the mind with the unit-level (as opposed to the pattern-level) of analysis of PDP systems. I think, though, that on either of the two identifications, his view results in a qualitative distinctness between traditionalism and PDP. If we take the second quotation as representative, and consider PDP systems at the unit-level of analysis as the model of the mind to be associated with the PDP paradigm, then the two paradigms are qualitatively distinct, for they postulate distinct causally efficacious entities. This follows from three premises that he makes, either explicitly of implicitly:

> (1) In PDP systems, the only exceptionless laws are to be found at the level of changes of states in individual units.
> (2) The representational content of individual units is *not* that of concepts.
> (3) Traditionalism is committed to the causal efficaciousness of entities that bear content at the conceptual level.

He sums up the conclusion of this line of reasoning:

> Does the complete formal account of cognition lie at the conceptual level? The position taken by the subsymbolic paradigm is: No -- it lies at the subconceptual level.[28]

There are, however, passages in which he implies that it is on the level of analysis of patterns of activation within PDP networks that the official PDP model of the mind is to be found. This "tension"

[28] "On the Proper Treatment of Connectionism", page 7.

is his thinking has a clear source; namely, the fear that the unit-level of analysis of PDP networks is not a model of the *mind*, for the laws governing transitions among unit states do not advert to meaning. He states this himself:

> subsymbols [the units of representation of individual units] are not operated upon by symbol manipulation: they participate in numerical -- not symbolic -- computation.[29]

If one takes this quotation in its strongest interpretation (i.e., that unit level laws are *only* syntactic -- they do not admit of a construal as adverting to the meaning of unit states) then this fear is justified. Thus, he is on occasion driven to considering the level of analysis of patterns of activation (which he allows represent concepts) as the model of the mind supplied by PDP. Even on this interpretation, PDP is qualitatively distinct from traditionalism. Even though the two make the same ontological commitments vis-a-vis what the mentally causally efficacious objects and states are (namely, symbols with conceptual level content), the laws describing state transitions will be different. According to traditionalism, the mental causal laws are precisely formalizable and computable; whereas, according to Smolensky:

> [Typically, interactions at the level of patterns of activity, which, under this intrepretation, would be the mental causally relevant entities,] can be computed only approximately. In other words, there will generally be no precisely valid, complete,

[29] "On the Proper Treatment of Connectionism", page 3.

> computable formal principles at the conceptual level; such principles exist only at the level of individual units -- the subconceptual level.[30]

There is a third construal of PDP that Smolensky never considers. He never examines whether sense can be made of the unit-level interaction as adverting to content; rather, as stated above, he assumes that unit-level laws governing transitions between activation values are *only* syntactic (i.e., they admit of no semantic counterpart). I think, though, that his limiting of content to the pattern level is unnecessarily restrictive, and shows a common misunderstanding in the role played by meaning in structuring causal laws.

Smolensky, along with many others, makes the assumption that, if causal laws adverting only to syntactic features of a system are available to explain all syntactic transitions within the system, then semantic considerations (for example, the semantic properties possessed by that same system) must be causally inert. Thus, according to this way of thinking, because a complete description of unit-level activity can be given in syntactic terms alone, no semantic laws at the unit-level exist. My view, as made clear in Chapter 2, is that his assumption results from an incorrect construal of the counterfactual testing for causal relevance of semantic properties. Smolensky (and others) who right from the start deny causal relevance to semantic properties at the unit-level, misinterpret the antecedent to the counterfactual:

[30] "On the Proper Treatment of Connectionism", page 6.

> If the system were not in a state with this particular (unit-level) meaning, then this other state would not follow.

as:

> Hold everything else constant (e.g., the system architecture, the items presented to the system during the learning phase, and the weights learned during the previous cycles of the system), but stipulate that the unit state has a different meaning ...

rather than (my preferred interpretation of it):

> Hold the system architecture constant, but allow the past to vary so that the system is not in a state with this (unit-level) meaning ...

The weights acquired during the learning cycle govern transitions between activation values, but the learning cycle (in particular, the inputs received from the environment and the changes in weights that result from application of the learning rules as a function of the reinforcement signal), and, hence, the weights, would have been different had the environment been different. I am taking seriously the theory of representation presented in Chapter 4, whereby the (Type III) representational content of a state is determined by the causal role that it acquires in mediating the causal sequence between the presence of an object and the production of behavior appropriate to that object. Had the object during the learning phase been something other than it was, the causal sequence connecting registration of the presence of that object and object-appropriate behavior would have been different; hence, the system that resulted would be *syntactically* different (because the weights would be different). So, we would not expect the behavior in this counterfactual world to remain as it was in the actual world.

Smolensky´s quick dismissal of a unit-level construal of PDP as provider of a model of the mind is thus seen to be premature. There are, as stated in Chapter 4, other reasons for rejecting the unit-level analysis of PDP as a model of the mind; although, these reasons are based more on empirical considerations than on an analysis of what content-adverting causation could possible mean.

The local interpretation schema view of PDP is clearly qualitatively distinct from traditionalism, as the presumed causally relevant entity on this view of the mind is a singleton (subconcept) that is not recognized as alone causally relevant within traditionalism. Rather, traditionalism recognizes only conceptual-level states as efficacious.

The most consistent position that Smolensky could take (although he nowhere does so explicitly) is that PDP as model of the mind presupposes a distributed interpretation schema -- thus, the ontological commitments of PDP are the same as those of traditionalism. However, PDP is not in general a mere implementation of traditionalism (a charge that we shall see made by Fodor and Pylyshyn), for it is the rare case in which PDP systems admit of exceptionless transitions between patterns of activation (and, on a semantic level, between traditionalism-like representational states). The second half of this position reinforces the relevance of the unit-level analysis to PDP as a model of the mind, in that, qua explanatory model, the actual causal laws governing mental state transitions must be not merely simulated but implemented, along with their concommitant ceteris paribus clauses. This genuine implementation is possible only if the model includes the unit-level -- not as itself providing the mentally causally relevant representations (a la the local interpretation schema), but as implementing the true mental causal laws.[31]

The second of the two works that have set the tone in the traditionalism versus PDP debate is Fodor and Pylyshyn´s "Connectionism and Cognitive Architecture: A Critical Analysis". As already mentioned, these authors are also of the opinion that traditionalism and PDP are distinct, as is demonstrated by the following quote from the introductory section of that paper.

> When taken as a way of modeling cognitive architecture, Connectionism really does represent an approach that is quite different from that of the Classical cognitive science [traditionalism] that it seeks to replace.[32]

The major burden that Fodor and Pylyshyn assume in this paper is to show that, if PDP is a (true) model of anything, it is something other than the mind. One can see this in a previously-quoted passage. Their reasons for arguing in favor of this PDP-as-

[31] For a more thorough discussion of this topic, see my unpublished draft "On the Necessity of Including the Implementation Level in a Model of the Mind".

[32] "Connectionism and Cognitive Architecture: A Critical Analysis", page 4.

implementation view are that mental activity displays certain regularities (summarized in that paper, but more fully developed in Fodor´s *The Language of Thought*) which PDP, when interpreted as a model of the mind, cannot explain. In Section 4 of this chapter, I shall look at parts of the LOT argument in some detail, in considering whether it is laying down principles of mental processing (i.e., necessary conditions), or merely citing empirical evidence (i.e., the instances of mental processing that we have heretofore encountered possess these properties, but there might be, in theory, a being lacking one or several of these properties who nevertheless has a mind).[33]

The LOT argument describes certain conditions had by mental causal sequences. So, for example, the systematicity of thought condition stipulates that the possession of certain mental states implies the ability to possess certain other mental states. The causal structure of the mind ensures this systematicity. In particular, the components of certain mentally causally efficacious representations are themselves causally efficacious, and the causal role of the complex representations are somehow a function of the causal role of the constituents. The effect of this combinatorial structure of certain representations is the above-mentioned systematicity. Fodor and Pylyshyn then ask the question: "are the causally efficacious states postulated by PDP similarly structured, so as to ensure systematicity (and the other conditions of the LOT argument)?" They

[33] I think it is safe to assume that all readers of this work are already sufficiently familiar with the oft-reproduced LOT argument that I can omit a full summarization of it here. In the process of arguing that it is citing empirical evidence relevant to, but *not* laying down necessary conditions for, mental processing in Section 4, I shall be summarizing parts of it.

answer: "no": such regularities within PDP systems, on those occasions when they do obtain, would be merely accidental, not the result of an underlying combinatorially-structured syntax and semantics. Does this mean that traditionalism (which does possess this combinatorial structure) and PDP are qualitatively distinct? The answer is, I think, not so clear. Fodor and Pylyshyn´s preferred view would answer "yes", for the two models imply divergent mental causal laws (and, hence, divergent causal sequences).[34]

There is also a way of interpreting the LOT argument that yields a less unequivocal answer. If it really is a principle of mental phenomena that the LOT conditions hold, then perhaps a PDP system

[34] Fodor and Pylyshyn give a somewhat confused account on this matter, for they want to distinguish the *merely* causal relations among representations from their structural relations, as in:

"Connectionist theories acknowledge *only causal connectedness* as a primitive relation among nodes; when you know how activation and inhibition flow among them, you know everything there is to know about how the nodes in a network are related. By contrast, Classical theories acknowledge not only causal relations among the semantically evaluable objects that they posit, but also a range of structural relations, of which constituency is paradigmatic." ("Connectionism and Cognitive Architecture: A Critical Analysis", page 12.)

However, it is not clear how such structural relations can make a difference if they are not in some way realized in the causal relations among representations. The whole point of the LOT argument is that certain observable mental phenomena are explained by these structural relations; but, in order to be observable mental phenomena, these relations must make a difference in the causal relations among the representations. It is not clear, therefore, what the structural relations are above and beyond restrictions on the mental causal laws. Indeed, in many of their examples illustrating the difference between traditionalist and connectionist models, they describe the ramifications of combinatorial structure solely in terms of restrictions on the set of mental causal laws, as in:

"Now consider a Classical machine. This machine has a tape on which it writes expressions. Among the expressions that can appear on this tape are: "A", "B", "A&B", "C", "D", "C&D", "A&C&D" ... etc. The machine´s causal constitution is as follows: whenever a token of the form P&Q appears on the tape, the machine writes a token of the form P. An inference from A&B to A thus corresponds to a tokening of type "A&B" on the tape causing a tokening of type "A"." ("Connectionism and Cognitive Architecture: A Critical Analysis", pp. 15-16.)

will be forced to develop a structure consistent with these conditions during the learning process. Fodor and Pylyshyn have no argument (indeed, none is possible) that PDP systems are disbarred from evolving during learning into systems displaying all of the LOT conditions. Furthermore, if these conditions constitute a part of the set of conditions operative in the development of a Type III (learned) representational system, then the fact that the PDP system displays these regularities would not be mere coincidence. Fodor and Pylyshyn never consider this possibility: they assume that such conditions, in order to be non-contingent, must be "built-into" the structure of the model, rather than resulting from the learning process. And, they argue, any PDP system with such a "built-in" combinatorial structure is nothing above and beyond an implementation of a traditionalist model. I shall return to this topic in Section 4, when I examine the nature of the LOT conditions in more detail.

A second burden of Fodor and Pylyshyn´s paper is to argue that PDP is best understood as a (non-mental) model of the implementation of the mind: to the extent that PDP has been successful in explaining certain phenomena, its success can only be granted on the assumption that it is *not* a model of the mind, but rather a model of the subprocesses (none of which involve mental causal laws) that implement the mind. As I have argued previously, when two models explain distinct levels of reality (in this case, the traditionalist model explains the mental level and the PDP model the implementing level, according to Fodor and Pylyshyn), they are qualitatively distinct. The ontological commitments of traditionalism include states picked out by their representational content, whereas (on this view) the ontological commitments of PDP include no such things.

There is a third alternative position on the relationship between traditionalism and PDP. It states that the mind cannot be wholly explained by either of the two models; rather, PDP explains one subset of mental phenomena, traditionalism explains a distinct subset, and there is no reduction between the two subsets. Both Smolensky and Fodor and Pylyshyn suggest this as a possibility, but, in each case, the subset of mental phenomena that "the other" model explains is vanishingly small. For example, Fodor and Pylyshyn say:

> It could still be that [PDP] networks sustain *some* cognitive processes. A good bet might be that they sustain such processes as can be analyzed as the drawing of statistical inferences. ... Since we doubt that much of cognitive processing does consist of analyzing statistical relations, this would be quite a modest estimate of the prospects for network theory compared to what the Connectionists themselves have been offering.[35]

Smolensky is perhaps a bit more generous in allowing that the traditionalist model explains mental phenomena dealing with novice problem solving (i.e., that characterized as the conscious following of explicit rules). However, neither set of authors wants to grant much. More ecumenically-minded authors are Robert van Gulick[36] and

[35] "Connectionism and Cognitive Architecture: A Critical Analysis", page 68.

[36] See, for example, his commentary on Smolensky´s article "On the Proper Treatment of Connectionism", pp. 57-58.

Woodfield and Morton[37], who suggest the possibility that fairly large expanses of mental phenomena are explained by traditionalism and PDP, respectively. (Smolensky refers to this version of ecumenicalism as "cohabitation".)

The question then follows, are the two models qualitatively distinct on this view? Note that this is similar to a comparison of theories that we have already met in this chapter: namely, that between the corpuscular and wave theories of light.[38] There was a time earlier in this century when physicists interpretted the two theories not as competing, but as cohabitating. The corpuscular theory´s purview included optical phenomena in which the energy packet nature of light and the states mentioned in quantifying this nature were causally efficacious, whereas the wave theory´s purview included those phenomena in which the wave properties were causally efficacious. Thus, while both theories attempted to explain the same level of reality, they were not competing to explain the same phenomena. This is just the view (albeit, with respect to a different domain) held by the cohabitation proponents in the traditionalism versus PDP debate. As already argued, the corpuscular and wave theories are qualitatively distinct. An analogous argument can be given for the qualitative distinctness of

[37] Also part of commentary on Smolensky´s article, page 58.

[38] Usually, when one speaks of the corpuscular versus wave theories, one means the two competing paradigms in optics in the early 19th century: the Newtonian theory and the (newer) wave theory (not clearly identifiable with a single name). In the above passage, however, I mean the cohabitating theories within optics just before the development of the photon-as-quantum-mechanical-entity theory of light. During this period, it was common to consider neither corpuscular nor wave theory as alone encompassing all optical phenomena; rather, some phenomena were explainable by means of the corpuscular theory, and other phenomena by the wave theory.

traditionalism and PDP on the cohabitation view: by the very supposition that the two theories cover distinct ranges of phenomena, there must be some causally relevant property or properties that split the overall set of cognitive phenomena into two subsets, such that those phenomena explainable by traditionalism have that property or properties and those explainable by PDP do not, and vice versa. Thus, the ontological commitments of the two theories differ. The end effect of the differences are clearly identifiable when considering whether there is a possible isomorphism between causal laws.[39] If the strict exclusivity of theory range is maintained (as, in a consistent cohabitation view, it must be), there can be no isomorphism; for, by the very supposition of distinct ranges, no isomorphism of causal sequences is possible. The effects mentioned in the one theory would never be mentioned in the other. To cite a concrete example, it is common among proponents of the cohabitation view to put memory storage and access phenomena within the range of PDP, and outside the range of traditionalism. Thus, while traditionalist mental causal laws may make use of remembered items, the actual retrieval of those items in the mind is not within traditionalism´s purview. So, traditionalism would postulate no causal law, the "effect-side" of which is the recalling-of-x. This could, however, appear on the "effect side" of a PDP causal law. Where a correlate of an effect

[39] As noted, sameness of ontological commitment is conceptually prior to isomorphism of causal laws; however, as the ontological commitment of a theory becomes clear only on an analysis of the objects and states quantified over in its causal laws, the two characteristics are not independent.

postulated by one theory is wholly absent from the other, there can be no isomorphism.

This concludes this discussion of the views concerning the qualitative distinctness between traditionalism and PDP found in the literature. We have seen the two flavors of qualitative distinctness exemplified. Both Smolensky and Fodor and Pylyshyn admit readings in which they are describing different levels of reality (for Smolensky, this is the less preferred interpretation of his view, whereas for Fodor and Pylyshyn, it is the more preferred). The other readings of Smolensky, Fodor and Pylyshyn, as well as the cohabitationists represent the other flavor of qualitative distinctness, for which the two theories describe the same level, but have little or no overlap. In Section 5, I will put my neck on the line and give my answer to this question.

5.4 Issues Concerning Computability and Computation (Plus, the LOT Argument)

Before tackling this issue, though, I feel obliged to discuss a topic that has been oft discussed in the literature, the treatment of which I believe shows a widespread misunderstanding of the nature of computation -- and, in particular, what distinguishes computational processes from non-computational processes. The form in which one most often meets this topic in the traditionalism versus PDP debate is:

> My preferred model [proponents on both sides of this debate make use of the argument form] is

> superior as a model of the mind, because it can compute function-x; whereas, your model cannot.[40]

Thus, a traditionalist might mention some Turing-computable function and hint that no PDP network could realize it. As follows from what I said in Chapter 4, Section 1, such an argument is guaranteed to be unsound, for it has been known since the 1940s that a PDP network can (in theory) be constructed to instantiate any Turing-computable function. One also sees the reverse: PDP proponents saying that their model is superior because it can compute functions that no traditionalist-type machine could compute. This error is particularly glaring, as the set of computable functions is *determined by* the set of Turing-computable functions. (This assumes a construal of Church´s Thesis as a quasi-analytic statement -- to be computable is to be Turing-computable. The reader is free to disagree with this construal. This does not, however, ameliorate the errorfulness of referring to the processes within PDP systems as computational.)

Both of these lines of argument show the same misunderstanding, in that both presuppose that what PDP systems do is to compute. This is, however, not the case; rather, PDP systems instantiate functions. The misunderstanding arises because, given the current state of technology, PDP systems must be simulated on digital computers; so, it is perhaps natural to assume that PDP systems compute their functions. (Interestingly, though, no one

[40] Obviously, this is premised on function-x´s being a cognitively relevant function -- ie, a function realized in cognitive agents. Arguments for this suppressed premise are, however, usually lacking.

would make the analogous mistake of saying that a real spring/mass system whose state transitions are being simulated on a digital computer therefore computes its state transitions.) Looking back to Chapter 3 and the definition of a computational process (i.e., a computational process is one in which (1) the representational states being manipulated are explicitly stored, and (2) the program that refers to and transforms these states corresponds to formal rules governing the manner of manipulation), one sees that the second condition is not satisfied. In PDP systems, there is no distinct program governing the manipulation of representational states; rather, the representational states (i.e., activation values + weights) "include" the "program".

One very interesting thesis mentioned by proponents of PDP is that their systems can instantiate functions that are not computable. Whether any particular one of these non-computable functions is of significance to PDP as a model of the mind is still an open question. The only attempt I´ve seen in the literature to isolate a *particular, clearly cognitive* phenomenon[41] that does not correspond to a

[41] The emphasis is meant to make clear that I wish to exclude Rumelhart and McClelland´s "On Learning the Past Tenses of English Verbs", which is sometimes cited as a provider of an argument that there is a cognitive function that is not computable. But, as the authors themselves state: " ... we suggest that the mechanisms that process language and make judgments of grammaticality are constructed in such a way that *their performance is characterizable by rules*, but that the rules themselves are not written in explicit form anywhere in the mechanism." (Page 217, italics added.) I also wish to exclude the various writings of Dreyfus and Dreyfus, for they nowhere cite particular psychological evidence that some cognitive phenomenon does not correspond to a computable function. In any event, they are not proponents of PDP as I have laid it out in Chapter 4, for they explicitly deny causal relevance to meaning. Also excluded are the recent writings of Roger Penrose, whose arguments, I admit, I apparently do not understand. And, like Dreyfus and Dreyfus, he is in any case no supporter of PDP as a model of the mind.

computable function is due to Cummins and Schwarz.[42] However, their example of such a cognitive phenomenon (behavior relating to clothing) remains purely anecdotal.

Perhaps, even under the assumption that no individual non-Turing computable function is cognitively necessary, the mere fact that PDP does not engage in computation is sufficient to pronounce the two theories qualitatively distinct. This is, I think, the most ambitious interpretation of Fodor and Pylyshyn´s intent in their charge against PDP, that it is not getting it (i.e., modelling the mind) right because it does not presuppose that mental processes are computational in nature. Their LOT argument can be summarized as follows: Human mental phenomena display certain pervasive properties (i.e., productivity, systematicity of representation, compositionality of representation, and systematicity of inference). These properties are explainable only on the assumption that mental processing consists of rule-governed manipulation on representations in a combinatorially structured language of thought -- ie, that mental processing consists of computation. They couch this argument in several different forms. One version is:

> But we are *not* claiming that you can´t reconcile a Connectionist architecture with an adequate theory of mental representation (specifically with a combinatorial syntax and semantics for mental representations). On the contrary, of course you can: All that´s required is that you use your network to implement a Turing machine, and specify a combinatorial structure for its computational language. What it appears that you can´t do,

[42] See "Connectionism, Computation, and Cognition", in *Connectionism and the Philosophy of Mind*, edited by Horgan and Tienson.

> however, is have both a combinatorial representational system and a Connectionist architecture at the cognitive level.[43]

Thus, they attempt a transcendental proof (in the Kantian sense) of the computational nature of the mind. If their proof goes through, then the non-computationality of PDP processes would make it qualitatively distinct from traditionalism. In this case, it would be a condition on any explanatory model of the mind that it instantiate all mental functions computationally: mental causal laws would quantify necessarily only over computational states.

While the thought of proving such a strong result might bring glee to Fodor and Pylyshyn, their actual supporting argument for the LOT falls very far short of establishing any such thing. As I read it, all that they have done is to have listed several properties possessed by mental phenomena (although, as even they would admit, not universally so). One way of explaining these properties is by positing a certain type of mental architecture (namely, traditionalism). Thus, the whole LOT argument argument is a sort of "inference to the best explanation". They tend to waffle on this point. There are passages, such as: "The traditional argument has been that these features of cognition [systematicity, etc.] are, on the one hand, pervasive and, on the other hand, explicable *only* on the assumption that mental representations have internal structure,"[44] as well as passages suggesting a more modest proposal, such as: "But

[43] "Connectionism and Cognitive Architecture: A Critical Analysis", page 28.

[44] "Connectionism and Cognitive Architecture: A Critical Analysis", page 33, italics added.

if this explanation [linking systematicity with the existence of combinatorial structure] is right (and there don´t seem to be any others on offer), then mental representations have internal structure and there is a language of thought."[45] On this more reasonable interpretation of the LOT argument, the computational nature of mental states is not causally relevant, so no qualitative distinctness based on differences in ontological commitment can be traced back to the non-computational nature of the causally efficacious states in PDP systems.

Perhaps, though, one could argue that there could be no possible isomorphism between the causal laws in the two theories because, whereas the traditionalist laws would satisfy those properties (systematicity, etc.) identified in the LOT argument, the PDP laws would not. This seems to be, however, a bit premature without an accompanying argument that the PDP laws as resulting from the learning process would not possess just these properties.[46] Fodor and Pylyshyn consider this possibility as a way of reconciling PDP to traditionalism, but then reject it:

> It´s possible to imagine a Connectionist being prepared to admit that while systematicity doesn´t *follow from* -- and hence is not explained by -- Connectionist architecture, it is nonetheless *compatible* with that architecture. It is, after all, perfectly possible to follow a policy of building

[45] "Connectionism and Cognitive Architecture: A Critical Analysis", pages 39-40. Note in particular the weakness of claim implicit in this passage.

[46] Yet another topic for future research is the ability of multi-layered PDP networks to develop a connectivity pattern (perhaps generalizing from presented examples in which both P&Q and P are true, to the universally quantified *x&y->x*) that produces a system meeting the conditions identified in the LOT argument.

> networks that have *a*R*b* nodes only if they have *b*R*a* nodes ... etc. There is therefore nothing to stop a Connectionist from stipulating -- as an independent postulate of his theory of mind -- that all biologically instantiated networks, are, de facto, systematic. ... [However], it´s not enough for a Connectionist to agree that all minds are systematic; he must also explain *how nature contrives to produce only systematic minds.* Presumably there would have to be some sort of mechanism, over and above the ones that Connectionism per se posits, the functioning of which insures the systematicity of biologically instantiated networks. ... There are, however, no proposals for such a mechanism. Or, rather, there is just one: ... Classical architecture.[47]

This is the closest that Fodor and Pylyshyn ever come to arguing that PDP will certainly produce a set of causal laws non-isomorphic to those of traditionalism because its laws will not necessarily reflect the above-mentioned conditions. However, recall that, for the case at hand, one cannot speak of the specific mental laws associated with either traditionalism or PDP because they have yet to be formulated. Hence, the test for isomorphism, simpliciter, must be weakened to a test for possible isomorphism: are the two sets of laws possibly isomorphic? Lacking a proof that the mental causal laws that will result in PDP systems necessarily display non-systematicity, Fodor and Pylyshyn´s LOT argument does not supply evidence that traditionalism and PDP are qualitatively distinct.

[47] "Connectionism and Cognitive Architecture: A Critical Analysis", page 50. Italics as in original.

5.5 Are Traditionalism and PDP Qualitatively Distinct?

I have repeatedly stated and reiterate here (more for my own benefit than for that of the reader -- I must constantly remind myself of my goal so as not to wander too far afield) that I am performing a conceptual analysis and comparison, not a comparison of the empirical adequacy of either traditionalism or PDP as a model of the mind. My job is done as soon as the issue of qualitative distinctness is settled; the relative merits of the two theories is a topic for another day. So, finally, we reach the point where I give *my* answer to this question.

My starting assumption is that PDP is put forward as a model of the mind. While many researchers (predominantly neurobiologists) use PDP-type networks to model neural processes, to assume this view decides the issue of qualitative distinctness leaving little room for a philosophically-interesting discussion. Even on the other view, though, there are several ways of interpreting the relationship between the processes modelled by traditionalism and those modelled by PDP. The outline of this section is as follows. I describe each of these (three) alternatives, and consider the question: are the two models qualitatively distinct under this interpretation? As the first alternative has already been discussed in Section 3, and the second alternative is, in a sense, a special case of the first and third alternative, I shall give a rather abbreviated treatment to these. I concentrate on the third alternative, which is, I think, the alternative that correctly depicts the relationship between traditionalism and PDP. I argue that, under this interpretation, the two models *are* qualitatively distinct.

We have already met the first interpretation of the relationship between traditionalism and PDP in Section 3 of this chapter on the possible readings of Smolensky. According to this view, PDP is an implementation of traditionalism (so, some reductive or at least supervenient relation is assumed to hold), yet both are models of the mind, because both posit mental causal laws (i.e., causal laws that pick out states based upon their representational content). This corresponds to the reading of Smolensky that identifies the mental causal laws with unit level state transitions. At first, the very idea of two models explaining the same causal realm (namely, that realm whose laws advert to content), one of which is an implementation of the other, may seem ludicrous -- after all, there can be at most one level describing mental phenomena. Hence, two models assumed to describe distinct levels cannot both be models of the mind. The summary dismissal of this view is, however, premature (at least, it is premature to dismiss it as being incoherent). Perhaps the mental realm is causally "fat", in that mental causal phenomena are distributed over two distinct levels. If the reader is having difficulty making sense of this as a genuine possibility, consider an analogous relationship -- that between Newtonian mechanics and quantum mechanics restricted to the domain of (relatively) slow-moving, large objects. It is not incoherent to maintain that both describe physical causal phenomena, yet that quantum mechanics implements Newtonian mechanics. By restricting consideration to large objects -- for which quantum effects are not discernible -- the domain of mismatch, and, hence, of non-reducibility, is excluded. Perhaps, then, the mental realm is like the domain of classical physical objects -- explainable by two distinct theories, one of which implements the other. (This obviously skirts the issue of whether both theories do in fact correctly describe this domain.) In any event, this demonstrates that this possible relationship between traditionalism and PDP is not incoherent.

As already argued, the two models would be qualitatively distinct on this view. Also as mentioned previously, I do not find this alternative to be the one that best depicts the relationship between traditionalism and PDP, for it identifies the model of the mind offered by PDP as the unit level of description of PDP systems.

A second possible relationship allows only partial reduction or supervenience between the two models. Thus, for a restricted domain within the overall mental realm, PDP implements traditionalism; but, for another domain, no reduction or supervenience of mental states posited by traditionalism to states posited by PDP is possible. For this alternative, the qualitative distinctness of the two models with respect to the domain of reducibility or supervenience decides the issue of qualitative distinctness simpliciter: they are qualitatively distinct. I do not reject this possibility outright, but I note that the most philosophically interesting question relates to the qualitative distinctness of the two models when restricted to the domain for which traditionalism is not implemented by PDP. This is an issue that I consider with the third alternative interpretation of the relationship between traditionalism and PDP.

According to this third alternative, there is no possible reduction or supervenience relations linking traditionalism and PDP: there is no sense in which the mental realm is causally fat, and the two models explain distinct levels of phenomena within that realm. Looking back to Chapter 2 and my preferred view of causation presented there, the mental facts are predicted by the traditionalist mental causal laws plus initial conditions, and the mental facts are predicted by the PDP mental causal laws plus initial conditions, and there are no bridge statements (either universally quantified or particular) linking the states quantified over within the two models. Thus, traditionalism and PDP are models competing to explain the same level. As mentioned in Section 1 of this chapter, competing models are distinct, but not necessarily qualitatively distinct. To settle that question we must examine the two aspects of qualitative distinctness as applied to this interpretation of their relationship.

The first question to ask is: do they make the same ontological commitments? In the final sections of Chapters 3 and 4, I laid out the ontological commitments associated with traditionalism and PDP, respectively. First, the very broad considerations. Both theories are based on a physicalist metaphysics, and both assume that there are causally efficacious mental states picked out by their content. These mental states are explicitly instantiated in physical states, presumably in the physical states of the brain.

Since mental causal laws advert to content, one aspect of the ontological commitment of a theory of the mind is the level of reality represented by these causally efficacious mental states. I think that a good case can be made that the LOT argument imposes on traditionalism constraints that imply that the level of reality represented by mental states corresponds to word-concepts and propositions easily expressible by words and sentences in our public language. Similarly, the general considerations favoring a distributed interpretation schema impose on PDP constraints that likewise imply this level of reality as that represented by the causally efficacious mental states. Thus, the ontological commitments of the two theories are the same.

The question of ontological commitment is, however, only one of two questions to be addressed in deciding the issue of qualitative distinctness. The more complex of the two questions is: Are the two causal relations on the set of efficacious states associated with traditionalism and PDP respectively possibly isomorphic? Here it is important to recall what it means when a proponent of one or the other of the two models pronounces: "this is an explanatory model of the mind." Mental state transitions are regulated by the mental causal laws. In particular, the transitions of the states of a traditionalist (or PDP) system identified as the mentally causally efficacious states are regulated by the mental causal laws. These state transitions (picked out by content in a semantic description of system behavior) are also describable in syntactic terms.[48] Indeed, it

[48] I feel obliged to reiterate a point made repeatedly in this work, and justified in Chapter 2 -- my account of mental causation does not leave content causally impotent, even though the behavior of a system (whether artificial or biological in origin) that admits of a semantic description also admits of a syntactic description.

is the syntactic description of state transitions that defines the corresponding abstract machine.

According to traditionalism, it is the computational states that are the causally efficacious mental states. According to (my reworking of) PDP, it is the activation value plus weight vector states. So, traditionalism and PDP are qualitatively distinct only if there is no possible mapping between computational states and activation value plus weight vector states, such that the state transitions within the respective systems are the same. Is there possibly such a mapping? Given what I have said in Chapters 3 and 4, there cannot be such a mapping. Consider the evolution of the representational states within a PDP system. When a state changes, it must be either that one (or more) unit activation values changed, or that one (or more) weights changed, or that one (or more) of each changed. Let´s consider the first of the three cases separately. When only a unit activation value (or multiple unit activation values) has changed, there is no change in the line of succession of activation values -- all that has happened is that the next overall activation value has replaced the current one. A line of succession (determined by a fixed pattern of weights) defines a system dynamics not necessarily dissimilar from the succession of computational states within a traditionalist system. (Keep in mind that my task is to identify a dissimilarity between the state transitions of traditionalist and PDP representation states that is guaranteed to occur.)

The situation is otherwise when the change of state within a PDP system is owing to a change in weight (or weights). In this case, the line of succession of future states is altered. Recall from Chapter 4 the role that weights play within PDP systems: they implicitly encode the sequence of patterns of activation that will be produced on a particular input. Whereas a change in activation value merely changes the position of the system within a sequence of activation values, a change in weight changes the whole sequence. There is no comparable feature within a traditionalist system.

Weight change within PDP systems using the back-prop technique is the result of back-propagation of an error signal during learning. Perhaps, an opponent of the qualitative distinctness of traditionalism and PDP may argue, a similar effect can be found during the execution of a learning cycle within a traditionalist system. Certainly traditionalist systems can learn (an active area of research within traditionalist-based AI is the development of learning algorithms); however, the computational assumption at the heart of traditionalism limits learning within traditionalism to changes in the manipulated data structures. Changes in the algorithm (or program) take one outside the scope of computational processing. Again, this imagined opponent may wish to divorce traditionalism from computationalism. This move, however, would prove the undoing of traditionalism as a model of the mind, for the representational content of traditionalist states depends crucially on those states being computational states. Expressed compactly: No computationalist assumption, no representational content. No representational content, no subsumption under mental causal laws. No subsumption under mental causal laws, no model of the mind.

Traditionalism without the computational assumption is not a possible contender for a model of the mind.

Again, this imagined opponent may try another tack. Perhaps, the mode of learning within PDP is a detail not relevant to PDP´s being a model of the mind -- we should consider PDP sans learning when asking the question: are traditionalism and PDP qualitatively distinct? This tack is, like the previous one, doomed to failure, for PDP sans learning is not a possible candidate for a model of the mind. In accordance with Chapter 4, the representational content of PDP states is determined by the causal role of those states; but, the causal role for Type III representational systems (with "genuine intentional mental states") is determined via learning. Unlike traditionalism, PDP has no alternative means (i.e., computational statehood) for fixing meaning, so learning is an integral part of its being a possible model of the mind.

Looking at the foregoing argument in overview, we see that there is no possible isomorphism associating the transitions between mentally causally efficacious states within traditionalism and PDP. Various attempts at modifying the features of either theory so as to allow a possible isomorphism result in a product that is not a possible model of the mind. Therefore, taking the two theories seriously as models of the mind implies their qualitative distinctness.

BIBLIOGRAPHY

Anderson, J., (1983). *The Architecture of Cognition*, Cambridge, MA: Harvard Press.

Arbib, M., and Hesse, M., (1986). *The Construction of Reality*, Cambridge: Cambridge University Press.

Baker, L. R., (1991). "Dretske on the Explanatory Role of Belief", *Philosophical Studies*, Vol. 63, pp. 99-111.

Ballard, D., (1986). "Cortical Connections and Parallel Processing: Structure and Function", *Behavior and Brain Science*, Vol. 9, No. 1, pp. 67-120.

Banerji, R. and Mesarovic, M. (editors), (1970). *Theoretical Approaches to Non-Numerical Problem Solving*, Berlin: Springer-Verlag.

Bechtel, W., (1988). "Compatibility of Connectionist and Rule-based Systems", *Philosophical Psychology*, Vol. 1, pp. 5-16.

Boden, M. (editor), (1990). *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Boden, M., (1988). *Computer Models of the Mind*, Cambridge: Cambridge University Press.

Boyd, R., Gasper, P., and Trout, J. D. (editors), (1991). *The Philosophy of Science*, Cambridge, MA: MIT Press.

Churchland, P. S., (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*, Cambridge, MA: MIT Press.

Davidson, D., (1980). *Essays on Actions and Events*, Oxford: Clarendon Press.

Dennett, D., (1987). *The Intentional Stance*, Cambridge, MA: MIT Press.

Dretske, F., (1991). "How Beliefs Explain: Reply to Baker", *Philosophical Studies*, Vol. 63, pp. 113-117.

Dretske, F., (1988). *Explaining Behavior*, Cambridge, MA: MIT Press.

Dreyfus, H., (1985). "From Micro-worlds to Knowledge Representation: AI at an Impasse", *Readings in Knowledge Representation*, R. Brachman and H. Levesque (editors), San Mateo, CA: Morgan-Kaufmann, pp. 71-94.

Fodor, J., (1990). *A Theory of Content,* Cambridge, MA: MIT Press.

Fodor, J., (1987). *Psychosemantics*, Cambridge, MA: MIT Press.

Fodor, J., (1983). *The Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J., (1975). *The Language of Thought*, Cambridge, MA: Harvard University Press.

Fodor, J. and Pylyshyn, Z., (1988). "Connectionism and Cognitive Architecture: A Critical Analysis". *Cognition*, Vol. 28, pp. 3-71.

Gorman, R. and Sejnowski, T., (1988). “Learned Classification of Sonar Targets Using a Massively-parallel Network”, *IEEE Transactions: Acoustics, Speech, and Signal Processing.*

Haugeland, J., (1985). *Artificial Intelligence: The Very Idea*, Cambridge, MA: MIT Press.

Hawthorne, J., (1989). "On the Compatibility of Connectionist and Classical Models", *Philosophical Psychology*, Vol. 2, No. 1, pp. 5-15.

Hebb, D., (1949). *The Organization of Behavior*, NY, NY: Wiley.

Heil, J., (1992). *The Nature of True Minds*, Cambridge: Cambridge University Press.

Heil, J. and Mele, A. (editors), (1993). *Mental Causation*, Oxford: Clarendon Press.

Hempel, C., (1965). *Aspects of Scientific Explanation*, NY, NY: Free Press.

Hertz, J., Krogh, A., and Palmer, R., (1991). *Introduction to the Theory of Neural Computation*, Redwood, CA: Addison-Wesley.

Horgan, T. and Tienson, J. (editors), (1991). *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer.

Kuhn, T., (1962/1970). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Lewis, D., (1983). *Philosophical Papers*, Vol. 1, Oxford: Oxford University Press.

Lewis, D., (1973). *Counterfactuals*, Cambridge, MA: Harvard University Press.

Loar, B., (1981). *Mind and Meaning*, Cambridge: Cambridge University Press.

Malcolm, N., (1968). "The Conceivability of Mechanism", *Philosophical Review*, Vol. 77, pp. 45-72.

Marr, D., (1982). *Vision*, San Fransisco, CA: Freeman Press.

McCulloch, W., and Pitts, W., (1943). "A Logical Calculus of the Ideas Immanent in Neural Nets", *Bulletin of Mathematical Biophysics*, Vol. 5, pp. 115-137.

Millikan, R., (1984). *Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press.

Minsky, M., and Papert, S., (1969 and 1988). *Perceptrons: Expanded Edition*, Cambridge, MA: MIT Press.

Nadel, L., Cooper, L., Culicover, P., and Harnish, R. (editors), (1989). *Neural Connections, Mental Computation*, Cambridge, MA: MIT Press.

Newell, A., (1990). *Unified Theories of Consciousness*, Cambridge, MA: Harvard University Press.

Newell, A. and Simon, H., (1972). *Human Probelm Solving*, Englewood Cliffs, NJ: Prentice Hall.

Newton-Smith, W. H., (1981). *The Rationality of Science*, Boston, MA: Routledge & Kegan Paul.

O´Brien, G., (1991). "Is Connectionism Commonsense?" *Philosophical Psychology,* Vol. 4, No. 2, pp. 165-178.

Pinker, S., and Price, A., (1988). “On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition”, *Cognition*, Vol. 28, pp. 73-193.

Poeppel, E. (editor), (1989). *Gehirn und Bewusstsein,* Weinheim, VCH Verlag.

Poeppel, E., (1985). *Grenzen des Bewusstseins,* Stuttgart: Deutsche Verlags-Anstalt.

Putnam, H., (1975). *Mind, Language and Reality,* Cambridge: Cambridge University Press.

Pylyshyn, Z., (1984). *Computation and Cognition,* Cambridge MA: MIT Press.

Ramsey, W., Stich, S., and Garon, J., (1990). "Connectionism, Eliminativism, and the Future of Folk Psychology", *Philosophical Perspectives,* Vol. 4, pp. 499-533.

Ramsey, W., Stich, S., and Rumelhart, D. (editors), (1991). *Philosophy and Connectionist Theory,* Hillsdale, NJ: Erlbaum.

Rosenblatt, F., (1962). *Principles of Neurodynamics,* NY, NY: Spartan Press.

Rumelhart, D., and McClelland, J., (1986a). *Parallel Distributed Processing,* Vol. 1 and 2, Cambridge, MA: MIT Press.

Rumelhart, D., and McClelland, J., (1986b). "On Learning the Past Tenses of English Verbs", in Rumelhart and McClelland (1986a), pp. 216-268.

Searle, J., (1984). *Minds, Brains and Science,* Cambridge, MA: Harvard University Press.

Shepard, R., and Cooper, L., (1982). *Mental Images and Their Transformations,* Cambridge, MA: MIT Press.

Smolensky, P., (1988). "On the Proper Treatment of Connectionism", *Behavioral and Brain Sciences,* Vol. 11, No. 3, pp 1-74.

Stich, S., (1991). "Causal Holism and Commonsense Psychology: A Reply to O´Brien", *Philosophical Psychology,* Vol. 4, No. 2, pp. 179-181.

Stich, S., (1983). *From Folk Psychology to Cognitive Science,* Cambridge, MA: MIT Press.