ON SOME ISSUES CONCERNING SYMBOLS AND THE STUDY OF COGNITION

by

JAY AARON LEBED

B.A., Philosophy, Swarthmore College
(1982)

SUBMITTED TO THE DEPARTMENT OF
LINGUISTICS AND PHILOSOPHY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1988

Signature of Author.....
Department of Linguistics and Philosophy
April 28, 1988

Certified by .....
Ned Block
Professor, Philosophy
Thesis Supervisor

Accepted by .....
Paul Horwich, Chairman
Committee on Graduate Studies
Department of Philosophy

TABLE OF CONTENTS

ON SOME ISSUES CONCERNING
SYMBOLS AND THE STUDY OF COGNITION

by

JAY LEBED

Submitted to the Department of Linguistics and Philosophy
on April 20, 1988 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in
Philosophy

ABSTRACT

This thesis concerns two issues of current interest in the foundations of cognitive science. The two issues are treated in three independent essays.

The first two essays concern a recent attack on the Representational Theory of Mind (RTM). According to RTM, propositional attitude states, states such as believing that p, desiring that q, etc. are to be analyzed as relations to mental symbols. Like the sentences of a language, these mental symbols are held to have semantic properties as well as formal, or syntactic, properties. Stephen Stich has argued that, while cognitive scientists should continue to posit mental symbols, there is no need to regard the symbols as having any semantic or representational properties. He thus proposes to replace RTM with the Syntactic Theory of Mind (STM).

In the first essay, I consider the question of whether, in fact, theories in cognitive science must appeal to the semantic properties of mental symbols. I argue that if homuncular functionalism, a popular account of the structure of explanations in cognitive science, is correct, then explanations of cognitive capacities will appeal in an essential way to the representational properties of mental symbols. Modifying such explanations to eliminate this appeal results, I argue, in a loss of generality and a loss of simplicity.

The second essay considers STM itself. I note that one of the potentially attractive features of STM is that it seems to avoid what Hilary Putnam has called the collateral information problem (the CI problem), the problem of having to distinguish the acquisition of beliefs that change the content of a mental symbol from the acquisition of mere facts, or collateral information. I argue that the principle of STM which Stich uses to escape the CI problem is itself untenable. In particular, the principle has the consequence that STM-based theories are unable to distinguish between pairs of beliefs that have the same logical form. I conclude that STM must abandon the principle which has this consequence, leaving it vulnerable to the CI problem. The attractiveness of STM is thereby diminished.

The final essay in the thesis concerns the use of connectionist networks in cognitive science. These networks are composed of large numbers of densely connected simple processing units that operate in parallel. Connectionist models are often portrayed as radical alternatives

to models based on conventional, symbol-processing computer architectures. It has been suggested in response that connectionist models are most plausible as low-level accounts, or implementations, of symbol-processing models. I argue for this latter view. In particular, I show that some who have urged the more radical understanding of connectionist models have based their views on very particular kinds of connectionist networks, which turn out to be overly simple. More adequate networks, I argue, are more likely to turn out to be implementations of conventional, symbol-processing models.

Thesis Supervisor: Dr. Ned Block
Title: Professor of Philosophy

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to a number of people for their contributions to this thesis.

Ned Block exhibited a great deal of patience in reading many versions of this thesis, some pairs of which were not easy to distinguish. His suggestions for improving the arguments in the thesis were of tremendous help. He also provided much needed encouragement at points when things seemed to be progressing very slowly if at all.

Although I did not ask Jim Higginbotham to join my thesis committee until late in the process, he read each version I gave him with great care. His important insights led me to rework and, I believe, improve much of the tnesis.

My views on a number of issues in tne philosopy of psychology, including those examined here, were formulated to a large extent in numerous discussions with Ron McClamrock and Greg Smith. They deserve a substantial share of the credit for what is good in this thesis.

Discussing the arguments in tnis thesis with fellow students has been enjoyable and enormously helpful in clarifying my thinking. A discussion with Michael Antony in the hallways of Building 20 was particularly beneficial. Tne questions he raised led to some important revisions in Essay I.

Writing a thesis is a long and difficult process. Having the support, encouragement and understanding of others engaged in the same process was, I believe, absolutely essential to the completion of this one. In tnis regard, I am greatly indebted to David Craig, and especially to Marcia Lind and Catherine Womack. Although he completed his doctoral thesis many years ago, Jay Keyser also provided encouragement at some times when it was greatly needed.

Most of all, I am grateful for the constant love, faith and support of Karin Kahn.

Homuncular Functionalism and Syntactic Theories of Mind

Since the publication of The Language of Thought in 1975, Jerry Fodor and like-minded philosophers of psychology have been developing a view of the mind based on the oft-drawn analogy between thought and language. For Fodor and other proponents of the Representational Theory of Mind (RTM), the central element of the analogy is the claim that having a belief or a thought is like having a sentence inscribed in one's brain.[1] In fact, having a belief or a thought is, on this view, having a sentence inscribed somewhere in one's head. Its just that the sentence is a sentence in a special language--the language of thought (LOT)--and, the sentence is not inscribed in ink or lead, but in some yet to be understood manner. If this analogy between language and thought is a good one, then we should expect sentences in the language of thought to have both a syntactic form and a meaning. Proponents of the Representational Theory of Mind believe that they do. The meaning of language of thought sentences plays an important role according to RTM in determining the nature of a person's beliefs. Specifically, when you have a belief, and, therefore, a sentence of LOT inscribed in your head, it is the meaning or content of the sentence that determines which belief you have. To have the belief that it's raining, for example, is to have a sentence of LOT inscribed in one's head whose content is that it's raining.

---

1. Although Fodor doesn't label it the "Representational Theory of Mind" until Fodor, 1980, the locus classicus for the view is probably Fodor, 1978.

One of the challenges facing RTM is to say what it is to have an LOT sentence of a given syntactic form inscribed or encoded somehow in your head. The proponent of RTM receives some comfort here from the analogy, which she embraces, between minds and computers. For, it is reasonably clear that we can invent formal languages--programming languages, e.g.--and store sentences of such languages in a computer's memory. Another challenge facing RTM is to say what determines the meaning or content of an LOT sentence with a particular syntactic form. This challenge is widely thought to be considerably more vexing.

In a recent book, From Folk Psychology to Cognitive Science: The Case Against Belief, Stephen Stich goes so far as to argue that the challenge of saying what determines the meaning or content of a particular sentence in the language of thought presents insurmountable obstacles. As a result, he urges a more limited analogy between language and thought. Stich agrees that the psychological states corresponding to what we call beliefs are like sentences written in the head. And, like sentences in a natural language, sentences in the head have a syntactic form. But, unlike sentences in a natural language, they do not, on Stich's view, have a particular meaning or semantic content. And, the same goes for beliefs and other psychological states which, according to RTM, inherit their content from the content of sentences in the language of thought. The psychological states of interest to cognitive scientists, Stich argues, should be ascribed syntactic, or formal, properties, but the question of what the states mean, or what their semantic content is, should be discarded. Stich urges that the Representational Theory of Mind be replaced by the Syntactic Theory of Mind (STM).

If questions about the meaning or semantic content of psychological states can simply be discarded, then the semantic content of these states must not, or at least need not, play any important role in cognitive science. Hence, the plausibility of STM, and any other view which urges cognitive science to dispense with the notion of the meaning or content of psychological states, depends crucially on a claim about explanations in cognitive science. Specifically Stich must endorse the following thesis[2]:

(T) Appeal to what a psychological state means, represents, is about, or refers to plays no essential role in explanations in cognitive science.[3]

(By "essential" I mean a role that could not equally well be filled by appeal to the syntactic form of the psychological state.) The thesis above is the subject of this essay. I shall call it the thesis of the dispensability of meaning (in cognitive science), and the opposing thesis, the one I shall defend, the thesis of the indispensability of meaning (in cognitive science). In arguing for the indispensability of meaning my aim is to undermine the Syntactic Theory of Mind and any other view which urges that cognitive science "go syntactic."

---

2. Well, not quite. Stich might want to limit his claim to those states which are analogous to beliefs, admitting that appeal might be made to the meaning or content of other psychological states. One would think, however, that if claims about meaning were to be essential for any cognitive states at all, they would be essential for belief-like states. For this reason, I won't worry about distinguishing between belief-like states (or other propositional-attitude-like states) and other cognitive states in arguing for the indispensability of meaning.

3. My talk of "what a psychological state means" is really just shorthand for what the mental representation or symbol whose tokening the state consists of means.

Arguments for the indispensability of meaning have usually emphasized the importance of generalizations that quantify over mental contents.[4] It is claimed, for example, that there are certain true important generalizations relating cognitive states and behavior, which could not be stated except by appeal to the content of those states. Hence, the argument goes, a theory which failed to ascribe content to these states, would miss true important generalizations.[5] In this essay, I give a rather different argument for the indispensability of meaning. The argument begins by assuming a widely accepted view of explanation in cognitive science, one which Bill Lycan has dubbed homuncular functionalism (homunctionalism for short).[6] Versions of the view can be found in Fodor, 1968; Dennett, 1978; Haugeland, 1978; Lycan, 1981; and Cummins, 1975 and 1983. Homunctionalism specifies a certain form for explanations in cognitive science. I will argue that psychological explanations which have that form commonly depend on the ascription of meaning or content to psychological states referred to in the explanation. Hence, the homunctionalist conception of explanation in cognitive science is incompatible with the thesis of the dispensability of meaning.

---

4. And, these are the only arguments Stich responds to in defending the thesis of the dispensability of meaning. See, in particular, Stich, 1983, pp. 170-83.

5. This argument is given in detail in Pylyshyn, 1981 (p. 161). See also Pylyshyn 1984, pp. 31-32, Fodor, 1986, p. 3, and Fodor, 1987, pp. 4-8 in regard to the emphasis on generalizations that quantify over mental contents.

6. Cummins also suggests that the importance of appeals to meaning is a consequence of the homunctionalist framework (Cummins, 1983, p. 42). But, his remarks fail to give the dispensabilist position serious consideration.

1. Homunctionalism and the Dispensability of Meaning.

1.1

The metaphor which underlies homunctionalism was suggested in a passage in Fodor (1968) in which he discusses the question of how we tie our shoes:

> We might thus consider expanding the population in one's head to include subordinate little men who superintend the execution of the "elementary" behaviors involved in complex sequences like grasping a shoelace. When the little man reads 'take the left free end of the shoelace in the left hand', we imagine him ringing up the shop foreman in charge of grasping shoelaces. The shop foreman goes about supervising that activity in a way that is, in essence, a microcosm of supervising tying one's shoe. Indeed the shop foreman might be imagined to superintend a detail of wage slaves, whose functions include: searching inputs for traces of shoelace, flexing and contracting fingers on the left hand, etc. (Fodor, 1968, p. 628.)

The homunctionalist view which this metaphor engenders is well known, so I won't bother presenting all the details. The basic idea is this. The aim of a theory in cognitive science is often to explain how it is that an organism or a system has a certain capacity--the capacity to tie its shoelaces, the capacity to understand language, the capacity to remember things, the capacity to recognize visually presented objects, the capacity to construct and subsequently execute plans, etc. In order to explain the organism's having the target capacity, the psychologist begins by dividing the capacity into a number of component subcapacities, and describes the manner in which they must interact in order to produce the target capacity (in Fodor's metaphor, each of the subcapacities would be assigned to a little man, or to a team of little men). This reduces the cognitive scientist's job to that of explaining the organism's possessing the component subcapacities and explaining how the execution of the subcapacities interact in the appropriate way. To do this the psychologist

simply repeats the process: she divides the subcapacities into their components, etc. The process is repeated until she reaches capacities which the organism or system has as a direct consequence of the physiological or physical characteristics of its parts--those parts presumably being something like neurons or groups of neurons in the case of organisms, and flip-flops or registers in the case of computers.

When this picture is spelled out in full detail, the story becomes rather complicated. For my purposes, however, the critical features of homunctionalism are captured by three claims. First, in attempting to explain the fact that an organism (O) has a capacity (C), cognitive scientists often divide the capacity into component subcapacities S_1, ..., S_n and specify the manner in which the subcapacities must interact to produce C. Second, this process can be iterated many times; and, oftentimes, if it is iterated enough, one reaches subcapacities the organism's possession of which can be explained by direct appeal to its physical constitution. Third, when the process can thus be completed, the following becomes a partial explanation of O's having capacity C[7]:

S: (i) O has subcapacities S_1, ..., S_n.

(ii) Subcapacities S_1, ..., S_n interact in O in the following way:
...

(iii) For the following reasons, any system that has capacities S_1, ..., S_n, and in which these capacities interact in the manner specified in (ii) will have the capacity C:
...

S can be extended to include the decomposition of subcapacities S_1, ..., S_n. In that case, the schema is repeated n times, one for each of the

---

7. Compare Haugeland, 1978, pp. 246-49 and Cummins, 1983, chapter 2.

subcapacities. And, of course, one can extend the explanation still further to include decompositions of the second-level subcapacities; etc.

One of the central claims of homunctionalism is that the schema above (interpreted in accordance with the discussion below) gives the form of typical explanations in cognitive science of an organism or system's having a particular cognitive capacity. I'll call explanations of this form homunctional explanations.

Three aspects of schema S require further comment. First, I want to say something about why I am focusing on partial rather than complete homunctional explanations. Second, I want to comment on the gaps in the schema that appear in steps (ii) and (iii). Finally, I want to discuss cases in which there is a discrepancy between a target capacity and what the system actually achieves.

A complete homunctional explanation would culminate, as noted earlier, with a level of subcapacities which the system or organism has as a direct consequence of the physical or physiological features of its physical or physiological components. There are three reasons for considering partial homunctional explanations to be the norm in cognitive science and complete homunctional explanations to be the exception. First of all, in a very complex system, such as the brain, a complete homunctional account may simply not be available. Still, partial homunctional explanations can be proposed and tested. Secondly, complete homunctional explanations may present an overwhelming amount of detail, not all of which is of interest. Finally, the closer a particular account is to a complete homunctional explanation, the less general it will be. Consider, for example, an explanation of how an automobile works. A partial homunctional account

which discusses the various systems involved--the electrical system, the braking system, etc.--and their major components may be almost completely general, applying to all or almost all automobiles. As the details of the explanation are filled out, the explanation will inevitably apply to fewer and fewer kinds of automobiles, though perhaps still to several models. A complete homunctional explanation, finally, would involve a detailed discussion of the actual physical components being used, and would, most likely, apply to only a single model. Obviously, other things being equal, the more general an explanation, the better. One of the great virtues of homunctionalism is that it allows one to add incremental levels of detail until the depth of understanding that is gained thereby is outweighed by the accompanying loss of generality and decrease in ease of comprehension.

Two gaps occur in schema S, one in (ii) and one in (iii). The gap in (ii) is to be filled by some description of how the subcapacities interact in O. Typically, the description will be in the form of a computer flow chart. The gap in (iii) is completed with whatever support is needed for the claim with which (iii) begins. Sometimes one can see just from inspecting the description in (ii) of how the subcapacities are to interact, that any system which has those subcapacities interacting in that way, will have the target capacity. In that case, the gap in (iii) need not be filled. Suppose, for example that we are explaining how an assembly line assembles a jar of Heinz catsup. There are just three subcapacities: the capacity to pour approximately 22 oz. of Heinz catsup into a jar, the capacity to move the jar, and, the capacity to take a cap and screw it onto the top of a jar. The manner of interaction is simply that the first capacity is executed first, the second capacity second and the third capacity third. It is apparent that any system which has these three

subcapacities, and in which the subcapacities interact in that manner will have the capacity to assemble a jar of Heinz catsup given an empty jar and a cap.

In other cases the gap in (iii) will be filled by a rather extensive discussion. Consider, for example, an account of how a car has the capacity to travel at high speeds on a flat surface. Much of the burden of the explanation may fall on step (iii). Suppose that the following subcapacities are described in step (i): the capacity of a certain component of the system periodically to deliver a spark to a certain chamber; the capacity of another component periodically to deliver a certain amount of air/fuel mixture to the same chamber; and the capacity of a third component to transform explosions in that chamber into a certain sort of energy. The discussion in (iii) will have to indicate why the sparks in the chamber will cause explosions given the periodic delivery of the air/fuel mixture. If the account is to convince someone ignorant of such matters, a fair amount of the theory of combustion will have to be discussed. Because the gap in (iii) may be filled by a rather extensive discussion and because it is rather open ended, I will refer to this part of the explanation as <u>the story</u>. If we say that the specification of the subcapacities of O and the manner in which they interact ((i) and (ii)) tell us how O is organized, then we can summarize by saying that, according to homunctionalism, an explanation of O's having capacity C will be a description of how O is organized together with a story about why system's organized in that way have capacity C.

I want to make one last point before going on to discuss the relevance of homunctionalism to the thesis of the dispensability of meaning. Suppose that the target capacity of a particular homunctional explanation is the

capacity to compute some function r. And, suppose that the system in question does not really compute r. That is, for one or more stimuli s, the system's response to s is something other than r(s). Suppose that r' is the function that the system actually computes. So, for some s, $r'(s) \neq r(s)$. If the function r in question maps retinal stimuli to descriptions of the shapes of objects in the visual field, then the stimuli for which $r(s) \neq r'(s)$ are optical illusions.

Now, we could change our target capacity to be the capacity to compute the function r'. But, sometimes it will be more useful to leave the target capacity as it is. We then proceed to analyze the capacity, aiming to explain how it is achieved (to the extent that it is) by the system. At some point in the analysis, we will analyze some capacity C into subcapacities S_1, ..., S_n interacting in some manner, such that a system with those subcapacities interacting in that way will not really have the capacity C, but something a little short of C instead. (If we don't encounter this situation, then we must not be analyzing the target capacity into subcapacities that the system really has.) Such steps in the decomposition, where there is some "slack" between what the system is supposed to achieve and what it actually achieves, allow us to pinpoint the source of the difference between the function that the system is trying to compute--r--and the function it actually computes--r'. For such decompositional steps, we will have to modify step (iii) of the schema above slightly. The story will have to show us not why any system with the posited subcapacities organized in the right way has the capacity in question, but (a) why any system with the posited subcapacities organized in the right way has something close to the capacity in question, and (b) for which stimulus conditions such a system fails. If we later want to

account for the stimuli s for which the system does not generate the response r(s), we simply go through the homunctional account and look for all the places where there is slack between a capacity and the subcapacities it is analyzed into and note the stimulus conditions for which the system fails to achieve that capacity. Note that we wouldn't be able to pinpoint the source of the system's errors in this way if we had reformulated the target capacity and regarded the system as attempting to compute r'.

1.2

Frequently the subcapacities alluded to in an homunctional explanation of a cognitive capacity involve the manipulation of some sort of structured mental entity. The structured entity may be a mental image, a prototype for a concept, a sentence in some sort of language of thought, an entry in a mental lexicon, a parsing tree, a frame, a 2-1/2 D sketch, or any of the other kinds of structured mental entities that cognitive scientists posit. The generic term for these entities is "mental representation," and that is what I'll call them.[8]

Let me pause for a moment here to get clear about the way I am going to talk about capacities. Suppose that Bob Hope's favorite function is the one that takes x into x^2. Then, if there is an ordinary way of counting capacities, perhaps the capacity to determine the square of a given number

---

8. I use the term 'mental representation' with some trepidation since it suggests that I have begged the central question of this essay--whether these entities must be regarded as having representational properties. I use it because the term is ubiquitous in cognitive science, and the alternatives (such as 'structured mental entity') strike me as exceedingly awkward. It will be apparent, I hope, that my use of the term (if not its connotations) never presupposes that these objects must be regarded as representing the world in some particular way.

would ordinarily be considered to be identical to the capacity to determine the result of applying Bob Hope's favorite function to a given real number. In other words, it may be that ordinarily we count capacities "extensionally." When, however, capacities are adverted to in explanations, counting capacities extensionally won't do. We can explain a machine's capacity to compute the area of a circle given the length of its radius, in terms of its capacity to compute the square of a given number, its capacity to multiply a given number by pi, etc., when the capacities are described in the explanation in just those words. We cannot, on the other hand, explain the machine's capacity to compute the area of a circle in terms of its capacity to compute the result of applying Bob Hope's favorite function to a given number, its capacity to multiply a given number by pi, etc., when the capacities are described in the explanation in just those words. In other words, it is really capacities under a description, and not capacities themselves, that figure in homunctional explanations. Of course, it is cumbersome always to speak of the capacity to X under that description. So, I will normally just refer to "the capacity to X" even though it is the capacity under that description that I will usually have in mind.

Now suppose that a given capacity in an homunctional explanation involves the manipulation in some manner of a particular mental representation. And suppose that the description of the capacity adverts to what the mental representation means or is about. Suppose, for example, that everyone maintains in a particular place in memory a list of the names of their close friends, relatives, colleagues, etc. Call this the P-list. We might speak of the capacity to discriminate those items on the P-list that refer to men from those items that refer to women. This description

of the capacity obviously presupposes that the mental representations--the items on the P-list--have a particular reference. Unless the items on the P-list refer to people, the description of the capacity doesn't make any sense, and, therefore, couldn't figure in a good explanation of any cognitive ability. Now let's suppose that each item consists not just of a name, but of a name together with another symbol--a letter, if you will. The letter is an 'M' if the item contains the name of a man and a 'W' if it contains the name of a woman. So, someone might have a P-list that begins

Nancy Reagan W
Ed Meese M
George Bush M
Elizabeth Dole W
....

The capacity to discriminate those items of the P-list that refer to men from those items that refer to women amounts to nothing but the capacity to discriminate those items ending in an 'M' from those items ending in a 'W.' And, this latter description of the capacity presupposes nothing about what the items on the list refer to.

We see, then, that a capacity under one description--e.g., "the capacity to discriminate items on the P-list referring to men from those that refer to women"--may presuppose something about the meaning of a mental representation even though the capacity under another description--e.g., "the capacity to discriminate items on the P-list ending with an 'M' from those ending with a 'W'"--presupposes nothing about the meaning of the mental representation.

When a capacity (under a description) presupposes something about the meaning of a certain mental representation, I will say that the capacity (under that description) is dependent on the meaning of the mental representation, or, more simply, that the capacity (under that description)

is meaning dependent. Oftentimes, when a meaning dependent capacity occurs in an explanation, it is possible to redescribe the capacity so that the capacity (under the new description) is no longer meaning dependent. When this is done, I will say that the capacity has been syntacticized. When a capacity is syntacticized, the redescription of the capacity may render the explanation impotent, but sometimes it won't. When a meaning dependent capacity occurs in an explanation and it is possible to syntacticize the capacity (and redescribe some of the surrounding capacities if necessary) so that the effectiveness of the explanation is not substantially impaired, I will say that the capacity (under the original description), as it figures in the explanation, is inessentially meaning dependent. When, on the other hand, a meaning dependent capacity occurs in an explanation and it is not possible to syntacticize the capacity without impairing the effectiveness of the explanation (even when some of the related capacities are also redescribed if this is necessary), I will say that the capacity as it figures in the explanation is essentially meaning dependent.

1.3

Recall the thesis of the dispensability of meaning in cognitive science:

(T) Appeal to what a psychological state means, represents, is about, or refers to plays no essential role in explanations in cognitive science.

When an homunctional explanation in cognitive science involves an essentially meaning dependent capacity, the explanation appeals, at least implicitly, to the meaning of a psychological state in ascribing the capacity to the system in question. And, such appeal is, of course,

essential. Since essential appeal to the meaning of psychological states is ruled out by (T), the following is an immediate corollary of (T):

(T') Homunctional explanations in cognitive science do not involve essentially meaning dependent capacities.

My aim in this essay is to argue against the thesis of the dispensability of meaning in cognitive science by arguing against this corollary of the thesis. In particular, I claim that homunctional explanations in cognitive science commonly involve essentially meaning dependent capacities.

As we shall see, it is usually, perhaps always, possible to syntacticize a meaning dependent capacity. But, I will argue, doing so will often substantially reduce the effectiveness of the explanation. In particular, the explanation will sometimes suffer a substantial loss of generality and will sometimes suffer an even greater loss of simplicity. In the next section I discuss an example of how syntacticizing a meaning dependent capacity can result in a less general explanation. In sections 3, 4 and 5, I turn to a more detailed example to show how syntacticizing a meaning dependent capacity can result in a dramatic reduction in simplicity. In section 6 I suggest that these syntacticizing explanations is also likely to produce a loss of insight. Finally, in section 7, the last before the conclusion, I argue that the examples discussed in sections 2 through 5 are by no means unique, thus completing the argument that homunctional explanations in cognitive science involving essentially meaning dependent capacities are common.

2. Meaning Dependence and Considerations of Generality

2.1

Consider an explanation of how we add pairs of whole numbers.[9] The way we do it, of course, is to write one of the numbers underneath the other, aligning the last digits of each number. We then proceed from right to left, focusing on one column at a time. We compute the sum of the two digits in the column and the carry, if there is one. Etc. Now, in order to make the processes involved in this example purely psychological, imagine that we typically perform such operations in our head, and not on paper.

One of the capacities involved in an homunctional explanation of our ability to add pairs of numbers in this way will be the capacity to determine the sum of three digits (the first two of which will be less than 10 and the third of which will be a 1 or a 0). Consider the question of whether this capacity is meaning dependent. It is helpful to think ourselves as receiving three expressions, say '4' and '3,' and '0,' as input, and having to produce a third expression, '7' in this case, as output. Now, given the characterization of the capacity as the capacity to produce a sum, it is clearly meaning dependent. A numeral can not be thought of as the sum of two other numerals unless the numerals are understood to represent numbers. Hence, given the current characterization of the capacity, the output (as well as the inputs) will have to be taken to represent a number.

---

9. The example was suggested to me by James Higginbotham.

But, it is easy enough for the dispensabilist to syntacticize the capacity. It can be recast as the capacity to produce the expression '2' if the first two inputs are the expression '1,' and the third the expression '0'; to produce the expression '3' if one of the first two inputs is the expression '2,' the other the expression '1,' and the third the expression '0'; etc. This move is somewhat problematic in a number of ways that will not be of immediate concern. First, the complete characterization of the capacity in question will be rather lengthy, and would, therefore, make the homunctional explanation more cumbersome and less easily understood. Second, the question arises whether the syntacticized capacity can play the same role in the homunctional explanation as the original explanation did. It is easy enough to see how the capacity to produce the sum of three digits fits into the capacity to add any pair of numbers; it is not nearly so obvious that one can see how (without noting that it is equivalent to the original capacity) the syntacticized capacity fits into the capacity to sum any pair of numbers. These two concerns will come up again later. But, for now I want to set them aside.

Instead, I want to raise another difficulty with the syntactic characterization. To get at the problem, I want to begin with an argument against the syntactic characterization that won't quite work. Let us grant, the argument begins, that the characterization produces a perfectly good homunctional explanation of our ability to add any pair of numbers. Still, the explanation only works for those of us who use Arabic numerals. The explanation will simply not apply to people or systems who use a different set of numerals--'a,' 'b,' ..., 'j,' for example, instead of '0,' '1,' ..., '9'--even if the system they use is functionally equivalent to

the Arabic numeral system. Such persons or systems may not even have, for example, the syntactic capacity described above: given the expressions '1,' '1,' and '0,' they may have no idea of what to do, even if they know they are supposed to be adding. The same is not true of the original capacity--the capacity to compute the sum of any three digits. Since it does not refer to specific numerals at all, it can be ascribed to any person or system using the familiar algorithm. The explanation containing the original capacity, therefore, is more general than the dispensabilist's explanation; it applies equally well to me and to a person unfamiliar with Arabic numerals who adds, using an equivalent numeral system, in the same way I do. And, this is as it should be, since, by hypothesis, we add in the same way.

The argument based on differences between Arabic numerals and other possible numerals rests on a crucial assumption. Specifically, it is assumed that the identity of the syntactic expressions being used must be derived from their physical manifestations. We needn't make this assumption. Stich, for example, proposes to individuate syntactic items functionally.[10] And, since the alternate numeral system and Arabic numeral systems are, by hypothesis, isomorphic, there is reason to think that the mental counterparts in me and the person using the alternate system will have equivalent functional roles. In that case, the relevant syntactic items would be identical, and the same syntactic characterization of the capacity to compute the sum of three digits could be applied to each of us.

Consider, however, a case involving me and someone who computes in base-12. Base-12 has, of course, 12 numerals. The syntactically

---

10. See Stich, 1983, pp. 150-53.

characterized capacity, referring as it does to only 10 numerals, could not possibly apply to addition in base-12. A syntactic characterization of the capacity to compute the sum of three digits in base-12 would have to refer to 12 distinct syntactic items. Hence, an homunctional explanation with purely syntactic characterizations of the sort anticipated could apply to addition base-10 or to addition in base-12, but not to both. Distinct explanations would be required. Still, as many children learn in elementary school, it is perfectly possible to use the familiar algorithm to add in base-12. But, the sense in which the algorithm is the same is only captured if the steps in the algorithm are expressed in a way that refers to what the numerals being manipulated mean. Only if, for example, one of the steps is computing the sum of the two digits in a column, plus the carry if there is one. Indeed, our original homunctional explanation of adding pairs of numbers can easily be stated so as to apply to addition in any base.[11]

We see, then, that if the homunctional explanation of adding pairs of numbers is to be as general as possible, we cannot replace the capacity to compute the sum of three digits with a syntacticization. Any such recharacterization has the result of narrowing the explanation's range of application. We are relying here on a very important feature of homunctional explanations. As has already been noted, homunctional explanations have a tree-like structure. One starts with a target capacity, analyzes it into a number of subcapacities, analyzes the subcapacities in turn, etc. The tree ends when one reaches capacities

---

11. The most substantial modification would involve places where the explanation refers to a sum being greater than 10. "10" would have to be replaced by "the number of the base-system being used." The explanation remains essentially the same.

which are explained directly by the physical constitution of the system. But, to give an explanation of the system's having the target capacity, it isn't necessary to present the entire tree. Describing the tree to any particular depth gives some sort of explanation of the system's ability: the greater the depth the more complete the explanation.

In the case just described, we have two systems, me and the person computing in base-12, that have the same ability. The complete homunctional explanations of the abilities are distinct. But, the differences only show up at a certain depth in the associated trees. Hence, if the homunctional explanations are cut off above that depth, the same explanation applies to both systems. Thus, we are able to see the sense in which the two systems achieve the target capacity in the same way, and also see the sense in which the target capacities are achieved in distinct ways. If the dispensabilist is forced to give entirely different explanations for two such systems, she will fail to be able to fully account for the similarities between the two systems. She will thereby fail to take advantage of a powerful aspect of homunctionalism.

2.2

The reader may have noticed a difficulty with the discussion above. The problem concerns the question of whether very simple machines have intentionality. One can imagine a very simple machine wired up to implement the addition algorithm. The machine need not be this complex, but let's suppose that it is a standard digital computer. Since the computer implements the addition algorithm, it ought to come under the purview of the homunctional explanation referred to in the previous discussion. And, I have argued, some states of such systems must be

regarded as having a semantic value. In particular some states of such systems must be regarded as representing numbers. And, this is, of course, to impute intentionality to the states of this very simple computer. Do I really want to say that the states of this very simple computer have meaning? No, I don't.[12]

What, then, do I want to say about the simple machine? To begin, notice that if we were literally to regard the machine as adding, we would have to assign meaning to its states. Syntactic objects can be added only when they represent numbers. The obvious move is to say not that the machine adds, but that it is interpretable as adding.[13] By saying that it is interpretable as adding, I mean that if we associate the inputs and outputs to the machine with numbers in the obvious way, and regard the machine as computing a function of its inputs, the function it computes is the addition function. It may be interpretable as doing other things as well, but that's fine. Rigorizing this notion of interpretability so that it applies to more complex examples would be a substantial task. But, at least in simple cases such as the one at hand, everyone understands intuitively what it means to claim that a machine is interpretable as performing some task.

---

12. Perhaps they have some sort of indirect meaning or intentionality if someone has programmed the machine to add. Then, for that person the inputs to the machine really do represent numbers. To avoid this side issue, the person issuing the challenge in question could suppose that the machine appears spontaneously as the miraculous result of the random motion of the particles that come to compose it.

13. The move here is essentially the one that Cummins makes in chapter 3 of Cummins, 1983. The ensuing discussion, including the use of "*-notatation" follows that discussion.

Let us say that a system *Xs just in case it performs operations that are interpretable as Xing. So, our machine doesn't add, but it does *add. And, what we want to explain is not its ability to add, but its ability to *add. We can transform the homunctional explanation of how we add into an explanation of the machine's ability to *add by putting asterisks in front of all the subcapacities that are meaning dependent. So, part of the explanation of the machine's ability to *add is its having the capacity to *(compute the sum of three digits). Of course, the two homunctional explanations are virtually identical. And, we could have the machine's ability and my ability come under the very same explanation by adding *'s and having *Xing mean either Xing or doing something interpretable as Xing. Similar remarks apply to all systems that use the adding algorithm. They will either add or *add, and their ability to do so will be accounted for by the original homunctional explanation or the *explanation, which, again could be turned into the same explanation. Thus, the homunctional explanation allows us to see what is common to all systems that use the algorithm.

I have now shown how we can bring the capacities of the simple computer under the purview of the original homunctional explanation without attributing intentionality to it. But, in doing so, it might be argued that I have played right into the dispensabilist's hand. After all, why can't the dispensabilist apply the analysis of the machine's ability given above to <u>us</u>. In other words, why can't the dispensabilist say that we don't really add, we *add. And, our capacity to do this doesn't involve computing the sum of three digits, but *(computing the sum of three digits). And, instead of regarding a certain psychological state as a token of a symbol that represents a number, we can simply regard the symbol

as being interpretable as a number. Thus, the dispensabilist could say, the homunctional explanation doesn't really require that we regard any of our psychological states as being tokens of interpreted expressions, only that we regard them as being interpretable in a certain way.[14]

The dispensabilist can indeed use this tack to avoid actually ascribing semantic properties to psychological states without abandoning homunctional explanations of cognitive capacities. But, such a move does not really save the thesis of the dispensability of meaning in cognitive science. The point of dispensabilism is not to allow the cognitive scientist to associate meanings with cognitive states without saying that the states are themselves meaningful. Rather, the point is to set the stage for getting rid of meaning from cognitive science altogether. My claim here can be stated in terms of the role of the word "essential" in the thesis of the dispensability of meaning. The thesis says that appeal to what a psychological state means "plays no essential role in explanations in cognitive science." The word "essential" could be interpreted so that the role couldn't be played by appeal to a meaning that is merely associated with the psychological state. But, this interpretation must be rejected, since the thesis would not then support the elimination of semantic notions from cognitive science altogether.

In light of the above remarks, we should refine the notion of meaning dependence. Before we said that a capacity (under a description) is meaning dependent if it presupposes something about the interpretation of a

---

14. Of course, since we want to explain our ability to add and not our ability to do something interpretable as adding, something we do--verbalizing the answer, for example--will have to be interpreted. But, of course, that is a social, not a psychological, act.

certain mental representation. Instead, we could say that a capacity (under a description) is meaning dependent if it presupposes something about the interpretation or interpretability of a certain mental representation. We could again say that when a meaning dependent capacity occurs in an explanation, and it is not possible to redescribe the capacity so that it is no longer meaning dependent and the explanation remains effective, the capacity is essentially meaning dependent. Given the above interpretation of the thesis of the dispensability of meaning, the thesis would still imply that meaning dependent capacities do not appear in homunctional explanations in cognitive science. Although the refinement is necessary to complete our solution to the problem at hand, I will simply observe that it can be so refined, and proceed to use the original version for the sake of simplicity.

The worry that the discussion of the example of adding pairs of whole numbers somehow commits me to saying that the states of something as simple as a 99 cent calculator have intentionality has now been put to rest. The conclusion of section 2.1--that syntacticizing the capacity to compute the sum of three digits results in a dramatically less general explanation of the capacity to add pairs of whole numbers--stands.

3. Newell and Simon on Problem Solving

In this section and the next two I want to present in detail another example of an explanation containing an essentially meaning dependent capacity. In this case, the example will hinge on considerations other than those of generality.

The example I will present is based on the work of Alan Newell and Herbert Simon on problem solving. Their views are presented in a number of places but most notably in their book Human Problem Solving (Newell and Simon, 1972), and I will be drawing from the account of problem solving presented there. I have chosen to draw from this work for two reasons. First, among the virtues of Newell and Simon's account is that it is very concrete and clearly laid out. This clarity and concreteness extends, I believe, to the homunctional explanation I will construct, and to my characterization of the representations whose semantic properties I claim the explanation to be dependent on. This clarity prevents, as far as possible, the judgment as to whether such semantic properties are being appealed to from being clouded by vagueness and sketchiness in the explanation or in the characterization of the representations. The second reason for choosing Newell and Simon's treatment of problem solving as the basis for my example lies in its centrality in the field of cognitive science. The study of problem solving lies at or near the heart of the study of cognition and Human Problem Solving is the classic treatment of the subject. Because of this centrality, one would expect that if explanations engendered by Newell and Simon's work are essentially meaning dependent, many other cases of essentially meaning dependent explanations in cognitive science could be found. And, if the dispensabilist wants to suggest that the explanation I cite is exceptional in this regard, the burden of proof falls on her. I want to emphasize that my reasons for choosing Newell and Simon's account of problem solving as a basis for my example have nothing at all to do with the question of whether their view is correct. All I claim for now is that a plausible theory in cognitive science, one which is quite characteristic of the field as a whole, yields

an essentially meaning dependent explanation. This should be enough to cast grave doubt on the dispensabilist thesis.

I turn now to the account of problem solving presented in Human Problem Solving. The first thing to note is that Newell and Simon worked with a rather constrained notion of a problem. Three classes of problems were studied in detail: symbolic logic problems, chess problems and cryptarithmetic problems. The authors are hesitant to say just how wide the scope of the theory is, writing, "It is clearly broader than the three tasks. On the other hand, the evidence at hand is ill-suited to define the limits of its scope." (Newell and Simon 1972, p. 791.) I will assume that these three sorts of problems typify a broader class of problems, and that the theory aims to be a theory of how people solve problems of the broader class. I'll use the term 'puzzle' to refer to problems falling into this broader class. So, I will take Newell and Simon's theory to be an account of the human capacity to solve puzzles.

The guiding notion of Newell and Simon's account of puzzle solving is the notion of a problem space. In fact, I'll sometimes refer to their view as the problem space account of puzzle solving. A problem space is constructed from the following four elements:

1. A set of knowledge states, K, each of which encodes a state of knowledge about the problem being solved;

2. A set of operators, O, each of which maps knowledge states into knowledge states;

3. A subset, S, of K, each element of which constitutes a solution to the problem;

4. An initial knowledge state, u_0.

The easiest way to see what all of this means is to look at a sample problem space, examining each of the four elements of the sample problem space in turn. Consider the following cryptarithmetic puzzle: DONALD + GERALD = ROBERT. A cryptarithmetic puzzle consists of a simple arithmetical equation in which the digits have been systematically replaced by letters so that the equation contains words or names instead of numbers. The letters can always be replaced by digits so that: (a) the same letters are replaced by the same digits (so, if the O in DONALD is replaced by a 2, then the O in ROBERT is replaced by a 2 as well); (b) two letters (types) are not replaced by the same digit; and, (c) the result of making these substitutions is a correct equation, e.g., 526485 + 197485 = 723970. The puzzle is solved when such a replacement scheme, an example of which is depicted below in Figure 1, has been discovered.

Puzzle:

```
  DONALD
+GERALD
------
 ROBERT
```

Solution:

```
D = 5; O = 2; N = 6;          526485
A = 4; L = 8; G = 1;   ---->  +197485
E = 9; R = 7; B = 3;          ------
T = 0                         723970
```

Figure 1

Let's see what a problem space for the puzzle, DONALD + GERALD = ROBERT might look like.

K, the set of knowledge states.

Knowledge states are quasi-linguistic entities. The set of all possible knowledge states for a given problem space is called the knowledge set. K*, the knowledge set for the puzzle DONALD + GERALD = ROBERT, will include knowledge states like (E=9,c5=1,A=4,L>5,T even). This knowledge state indicates that the letter E is to be replaced by the digit 9, the carry in the fifth column from the right in the target sum is a 1, the letter A is to be replaced by the digit 4, the letter L is to be replaced by a digit greater than 5, and T is to be replaced by an even digit. The knowledge set is specified in the same way that the elements of a formal language are--by giving a recursive syntax for the language, or in this case, the set of expressions. Newell and Simon use a notation often used for the grammars of programming languages to describe the recursive syntax. Below is a recursive syntax for K* presented in this notation.

```
<digit> ::= 0|1|2|3|4|5|6|7|8|9
<letter> ::= A|B|C|D|E|G|L|N|O|R|T
<carry> ::= c1|c2|c3|c4|c5|c6
<variable> ::= <letter>|<carry>
<expression> ::= <variable><relation><digit>|<variable> <parity>|
                 <variable>=<digit-set>
<relation> ::= =|>|<
<parity> ::= even|odd
<digit-set> ::= <digit>|<digit> V <digit-set>
<knowledge-state> ::= O|<expression>|<expression>,<knowledge-state>
```

The only metasymbols in this notation are '::=', '|', '<', '>', and 'O' (the symbol for the null string). '::=' indicates that the extension of a class of expressions is being described. On the left of '::=' is the name of the class of expressions whose extension is being described. These names, whether they occur on the left or right of the '::=' sign, are flanked by the signs '<' and '>.' '|' is a sign of disjunction. So, the first line indicates that a symbol is in the extension of the class of expressions, <digit>, if it is '0' or if it is '1' or if it is '2', etc. The disjuncts on the right of the '::=' sign may be individual expressions, as in the first line, classes of expressions, or concatenations involving individual expressions, classes of expressions, or both. These are to be read in the obvious way. So, in the fifth line above, which assigns the extension of the class <expression>, the last disjunct indicates that if an expression in the class <variable> is concatenated with the expression '=' and then with an expression in the class <digit-set>, the result will be a member of the class <expression>.

Note that K* will include some absurd knowledge states such as (E=9,E=8). This is to be expected: there are many contradictions among the grammatical sentences of English.

O, the set of operators.

The idea of a problem space is that puzzle solving consists in moving step by step from the initial knowledge state (often the null string) to a knowledge state which constitutes a solution to the puzzle. The individual steps from one knowledge state to another involve the application of procedures which transform one knowledge state into another. These procedures are called operators. The problem space in question contains four operators. Among them are PC (process-column) and GN (generate-numbers). PC has one parameter--the number of the column to be processed. The purpose of PC is to take the information contained in the knowledge state pertaining to the values of the three letters and two carries associated with the relevant column, and deduce new information about the three letters and two carries, if possible. So, to take a simple example, consider the operator PC(4) applied to the knowledge state (N=6,B=3).[15]

---

15. I will use expressions such as 'PC(4)' to refer, e.g., to the specific procedure gotten from the general procedure PC by setting the value of the procedure's parameter at 4. I will, for the time being, engage in a systematic ambiguity, using the term 'operator' sometimes to refer to specific procedures, such as PC(4), and sometimes to general procedures such as PC.

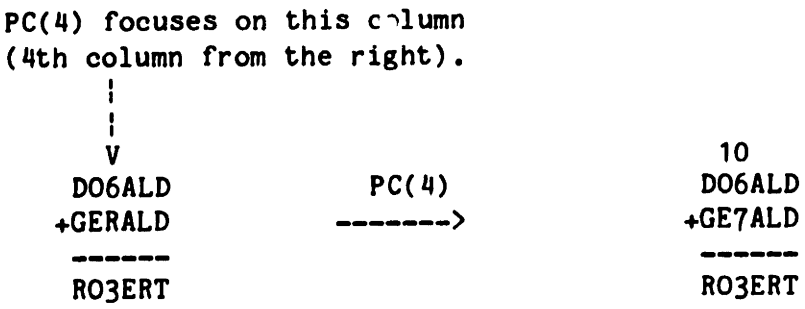

Figure 2

As illustrated above in Figure 2, PC(4) applied to (N=6,B=3) yields (N=6,B=3,R=7,c4=0,c5=1). Given the 6 and the 3 in the fourth column, R must either be 6 (in which case the carry from the 3rd column would have to be a 1) or 7. Since N has already been assigned 6, R can't be 6 as well. So, R=7. The carry from the 3rd column must be 0 (0+6+7 modulo 10 = 3). So, c4=0. And, since 0+6+7 is greater than 10, we have a carry of 1 in the fifth column: c5=1.

Like PC, GN (generate numbers) has just one parameter--the letter for which possible numbers are to be generated. It takes all current information about the letter, any current assignments of digits to other letters, and generates an explicit list of numbers that could be assigned to that letter. So, for example GN(R) applied to (R odd,R>4,E=9) would combine the information that R is odd, that R is greater than 4, and that E has been assigned the digit 9, to determine that the numbers that could be assigned to R are 5 and 7. In other words, GN(R) applied to (R odd,R>4,E=9) yields (R=5v7,E=9).

Complete rigor would demand a complete specification of the functions associated with these operators. Providing such a specification would be a

formidable task, and in practice, an informal definition of the operators of the sort given above is usually considered sufficient.

S, the set of solution states.

S is just the subset of K consisting of all knowledge states that represent a solution to the puzzle being solved. Of course, in settling on a solution set, one doesn't actually produce a list of solutions. If you could do that, the puzzle would already be solved. Rather, one must have a characterization of the solutions, so that given a candidate solution, you can test it to see if it really is a solution or not. For the puzzle at hand, producing a rigorous characterization would be a substantial task. The idea, of course, is that if you use the solution to substitute into the equation DONALD + GERALD = ROBERT, the resulting numerical equation is correct.

u_0, the initial knowledge state.

The final component of a problem space is u_0, the initial knowledge state. Frequently, as with the puzzle at hand, the initial knowledge state will be the null string. Sometimes, to make a cryptarithmetic puzzle easier, Newell and Simon include one piece of information with the initial characterization of the puzzle. For example, with DONALD + GERALD = ROBERT, they specify that D is to be equal to 5. u_0 for the puzzle DONALD + GERALD = ROBERT, D=5 is, of course, (D=5).

We have now looked at each of the component of the problem space for a sample puzzle. Formally, the problem space itself is a graph. The nodes of the graph are the elements of K. Nodes are connected directionally. Node q is connected to node p (but not necessarily vice-versa) if there is

an operator in O which maps p to q. A path through the problem space is defined as a sequence of nodes, the first of which is u_0, and the last of which is a member of S, such that for any node q in the sequence, if p is the node immediately preceding q in the sequence, then q is connected to p. Newell and Simon's account of problem solving is frequently summed up by saying that problem solving consists of search in a problem space. Formally, this means that it consists of the application of a sequence of operators in O to elements in K (the first of which is u_0) so that the sequence of knowledge states passed through is a path through the problem space.

Below I present an example of a path through the problem space just described. On the right of each knowledge state in the sequence I indicate the operator (and arguments) used to get from that knowledge state from the next one.

(D=5)	PC(1)
(D=5,T=0,C1=1)	PC(5)
(D=5,T=0,C1=1,E=9,C4=1,C5=1)	PC(3)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1)	PC(4)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R>6)	PC(2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R>6, R odd)	GN(R)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7)	PC(2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8)	PC(6)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1)	GN(N)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=2v3v6)	TD(N,2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=3v6)	AV(N)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6)	PC(4)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6,B=3)	GN(O)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6,B=3,O=2)	

I leave it as an exercise for the reader to determine the nature of the operators TD (test-a-digit) and AV (assign-value), which are applied in steps 10 and 11 respectively, but which are not described above.

I have now sketched the notion of a problem space. For Newell and Simon, this conception was just the first step toward a theory of problem solving. They probably considered the theory of the implementation of search in a problem space to be the core of the project. It was here that they were led to the tremendously influential idea of a "production system." For my purposes, however, that aspect of the theory which goes beyond the conception of a problem space will not be important. What has been presented so far will be sufficient to allow one to see how the first few steps in an homunctional explanation of the capacity to solve puzzles might go, and that is all I shall need.

4. A System that has the Capacity to Solve Puzzles

4.1

I now want to show how to decompose the capacity to solve puzzles into seven subcapacities in accordance with the hypothesis that puzzle solving consists of search in a problem space. The idea that puzzle solving is search in a problem space immediately divides the process of puzzle solving into two parts: picking a problem space to use to solve the puzzle, and conducting a search in the space selected. Since a problem space has four components, picking a problem space will involve four subcapacities, one for each component. Let's call these subcapacities KSET, OPSET, SOLSET, and ALPHA. KSET is the capacity to select intelligently a knowledge set for the puzzle to be solved; OPSET is the capacity to select intelligently a set of operators for the puzzle to be solved; SOLSET is the capacity to produce a characterization of the solution set--those elements of the knowledge set that are to count as solutions to the puzzle; and, ALPHA is

the capacity to correctly pick out the initial knowledge state for the puzzle. Clearly these capacities cannot be executed independently of each other. Each of the latter three capacities depends on knowing what knowledge set has been chosen. More will be said about this in discussing the manner in which all seven subcapacities are to interact.

Now consider the subcapacities needed to conduct a search in the problem space thus selected. Conducting such a search consists of constructing a path through the problem space and then recognizing when the path has led to a solution state. Taking the individual steps in the path consists of selecting an operator to apply to the knowledge state associated with the current node in the path, and then applying the operator to produce the knowledge state associated with the next node. As long as each operator is selected wisely, the individual steps thus chosen will ultimately constitute a path through the problem space. Conducting the search thus requires three subcapacities: the capacity to choose, at any given step in the search, an appropriate operator to apply to the current knowledge state; the capacity to apply an operator that has been selected to the current knowledge state and thereby produce a new knowledge state; and, the capacity to determine, at any step along the way, whether the current knowledge state is a member of the solution set. Let's call these capacities PICKOP, EXEC and TEST respectively.

The capacity to solve puzzles has now been decomposed into seven subcapacities. If the problem space account of puzzle solving is correct, then any system or organism that has these seven subcapacities, and in which they interact in the requisite way, will have the capacity to solve puzzles. And, its having that capacity will be susceptible to an homunctional explanation based on this decomposition. If Newell and Simon

are correct in claiming that people solve puzzles via search in a problem space, then this decomposition could be used to construct an homunctional explanation of the human capacity to solve puzzles. I could consider the question of whether such an explanation would involve essentially meaning dependent capacities. Instead of considering human puzzle solving abilities, however, I want to consider the puzzle solving capacities of a hypothetical cybernetic system. I do this because it will be possible to describe the operations of this system of digital computers more vividly than it would be possible to describe hypothetical human cognitive processes. Talk of cognitive processes being applied to a mental representation is vague and metaphorical in a way that talk of a computer executing a sequence of instructions which modify the contents of a certain register is not. And, the explanation of the puzzle solving capacities of our system, Solver, will suit my polemical needs as well as an explanation of human puzzle solving abilities would. Newell and Simon's work was explicitly based on the supposition that digital computers, programmed appropriately, could be endowed with the same puzzle solving abilities that we have, and that the explanations of our abilities and their abilities in this regard should be fundamentally identical. This assumption means that any lessons we learn about the explanation of the puzzle solving abilities of our cybernetic system should apply to humans as well.

So, consider a system--Solver--of seven digital computers together with a keyboard and a CRT, which system has the capacity to solve puzzles in virtue of having the seven subcapacities described above. There is one computer in Solver for each of the seven subcapacities. I will use the name of a given subcapacity to refer to the computer that has that subcapacity, using boldface to refer to the machines and regular capital

letters to refer to the subcapacity. So, Solver consists of a terminal, a CRT and **KSET**, **OPSET**, **SOLSET**, **ALPHA**, **PICKOP**, **EXEC**, and **TEST**.

Each of the machines in Solver is connected by an output-input link to one or more of the other machines in Solver. In other words, in virtue of an electronic connection between the two, the output of any given machine will serve as the input for some other machine. It is these output-input links that guarantee that the seven subcapacities interact, in Solver, in what I referred to above as "the requisite way." These output-input links can best be described by tracing the activity of the system as it solves a puzzle. First, a description of the puzzle to be solved is entered into the terminal. This description becomes the input to **KSET**. **KSET** produces a description of an appropriate knowledge set for the puzzle and delivers this description, along with the description of the puzzle to **OPSET**, **SOLSET**, and **ALPHA**. OPSET produces a description of the set of operators to be applied to elements of the knowledge set. This description, together with the description of the knowledge set, and the description of the puzzle, are delivered to **PICKOP**. The description of the set of operators and the description of the knowledge set are also delivered to **EXEC**. Meanwhile, SOLSET produces a criterion that solutions to the puzzle must meet and delivers it, along with the description of the knowledge set, to **TEST**. And, **ALPHA** selects the initial knowledge state and delivers it to **PICKOP**. **KSET**, **OPSET**, **SOLSET**, and **ALPHA** have now selected the problem space. It is up to **PICKOP**, **EXEC** and **TEST** to conduct a search in the problem space. The search consists in the execution of the following loop: **PICKOP** selects an appropriate operator to apply to the current knowledge state. The first time around, this is the knowledge state delivered by **ALPHA**. On subsequent occasions, the current knowledge state comes from

TEST. **PICKOP** delivers the operator it has selected to **EXEC**, which then applies that operator to the current knowledge state, thereby producing a new knowledge state, which it passes on to **TEST.** **TEST**, having received a characterization of the solution set from **SOLSET**, has constructed an algorithm for determining whether a given knowledge state is a solution state. **TEST** applies this algorithm to the new knowledge state. If the algorithm determines that the new knowledge state is a solution state, the loop is completed and the new knowledge state is displayed on the CRT.[16] Otherwise, **TEST** delivers the new knowledge state, which is now the current knowledge state, to **PICKOP**, and the loop begins again. The loop is repeated until a solution state is produced. At that point, the search is completed and the result displayed. Solver is depicted below in Figure 3. The thick arrows indicate connections involved in the loop:

---

16. One might want to translate the knowledge state into some other form before it is displayed and thereby presented as the solution to the puzzle. In fact, in a sense it partly begs the question against the dispensabilist to treat the knowledge state as a solution to the puzzle since the dispensabilist will generally regard knowledge states as uninterpreted syntactic objects. And, an uninterpreted syntactic object obviously cannot be regarded as a solution to a puzzle. To accommodate this point, I could have added another component to Solver that translates the solution state into English. But, such an added component would have remained idle in the rest of the discussion, and I have therefore not included it.

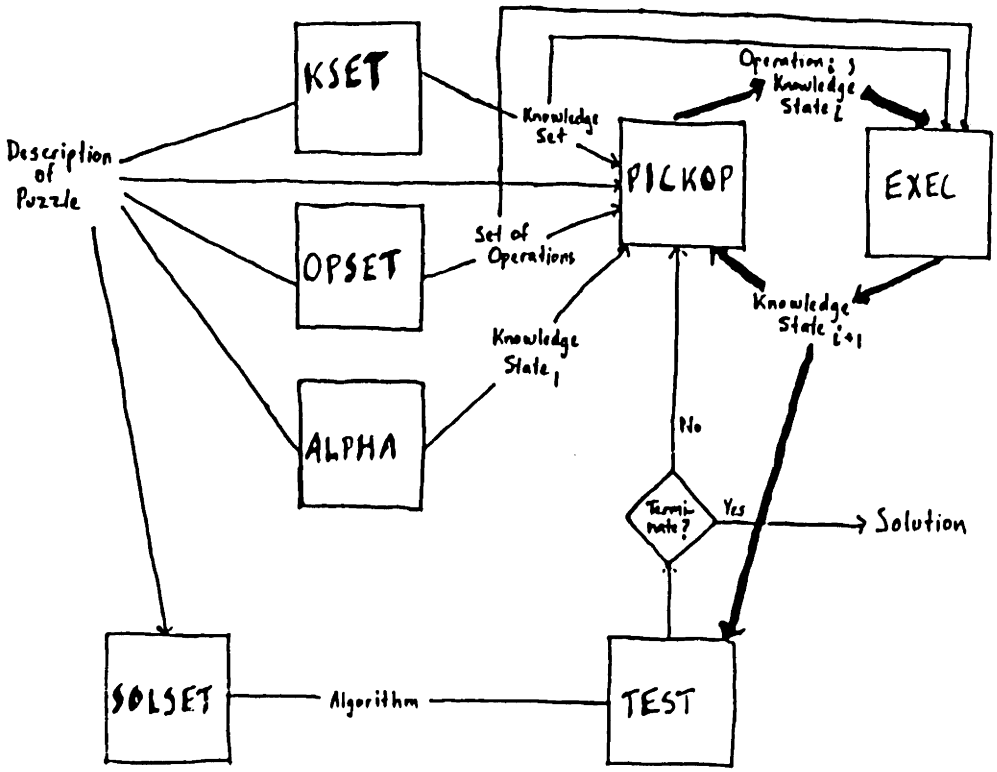

Figure 3

4.2

Here is an homunctional explanation of Solver's capacity to solve puzzles:

E: (i) Solver has the subcapacities KSET, OPSET, SOLSET, ALPHA, PICKOP, EXEC, and TEST.

(ii) In Solver, the subcapacities KSET, OPSET, SOLSET, ALPHA, PICKOP, EXEC, and TEST interact in the manner pictured in Figure 3.

(iii) Any system that has the subcapacities KSET, OPSET, SOLSET, ALPHA, PICKOP, EXEC, and TEST and in which those capacities interact in the manner pictured in Figure 3 will have the capacity to solve puzzles.

Step (iii) must be supplemented by what I called in section I "the story." The story associated with a decomposition of a capacity is simply whatever is needed to make it apparent why a system with the relevant

subcapacities interacting in the relevant manner will have the target capacity. In section 1, I discussed the example of the homunctional explanation of a car's capacity to move at high speeds on a flat surface. In that case, the story needed to include enough of the theory of combustion to explain why, if a spark is delivered periodically to a certain chamber and a certain air/fuel mixture is continually delivered to the same chamber, there will periodically be an explosion producing a certain amount of energy in that chamber. To see that a given system would have the capacity to move at high speeds on a flat surface, one needed a bit more than a characterization of its subcapacities and a description of the manner in which they interact; one needed to know something about combustion.

Something similar happens in E. Having read section 2, you know what search in a problem space is. Furthermore, having seen the example presented in that section, you see how search in a problem space can be used to solve puzzles. Finally, it is a pretty straightforward matter to see that the hypothetical system described in step (iii) will conduct a search in a problem space. This will follow almost straightaway from the definitions of the relevant capacities. Furthermore, the system will not pick just any problem space, nor will it conduct just any search in the space it has selected. We have required that the elements of the problem space are appropriate for the puzzle being solved. And, we have required that the operators applied at each step in the search also be appropriate given the current knowledge state. In sum, we have guaranteed that the system will conduct an intelligent search in an intelligently chosen space.

So, you see that the system in step (iii) will conduct an intelligent search in an intelligently chosen problem space, and you see how an

intelligently conducted search can be used to solve a puzzle. Do you then see that the system in step (iii) has the capacity to solve puzzles? Well, not quite. Roughly, you need to know that the example in section 2 is not just a fluke. And, you need to know this in two ways. First, you need to know that the system will be able to discover the path in the problem space described in the example. Or, if not that path through that problem space, then some other path through some other problem space. Secondly, you need to know that this will be true not just of that one cryptarithmetic puzzle, but of puzzles in general. The way to see all of this is to see a large number of examples worked out. By looking at a number of puzzles, and for each puzzle seeing that many problem spaces can be used and for each of these problem spaces, that paths through the problem space are easily discovered, one eventually becomes convinced that any system capable of conducting an intelligent search through an intelligently conducted problem space will have the ability to solve puzzles. So, the story associated with E must consist in this sort of a presentation of a wide variety of examples. (It is no coincidence that in their book, a central aim of which was to make the problem space account of puzzle solving plausible, Newell and Simon work through a wide variety of examples in great detail.) Once E has been thus supplemented, it will constitute an homunctional explanation of Solver's capacity to solve puzzles. In brief, E will show that Solver is organized in such a manner that it intelligently conducts search in a problem space, and that any system thus organized will have the capacity to solve puzzles.

Of course, E, supplemented in the manner described above, does not show that Solver can solve all puzzles. It only shows, roughly, that Solver can solve an awful lot of puzzles and many different kinds of

puzzles. But, this is all I mean when I say that Solver has the ability to solve puzzles, just as that is all I mean when I say that I have the ability to solve puzzles.

5. Meaning Dependence and Considerations of Simplicity

5.1

I now want to argue that the capacity PICKOP in E is essentially meaning dependent. PICKOP, attributed in E to **PICKOP**, is the capacity to select an appropriate operator given the following input: a description of the puzzle to be solved, the problem space to be used, and the current knowledge state. In regarding **PICKOP** (or any other system) as having this capacity, then, we picture it as receiving these three inputs, and having the capacity to produce an output that is appropriate relative to these inputs. To make the picture simpler, we can assume that **PICKOP** has already received the description of the puzzle to be solved and the problem space to be used. (This is an accurate description of **PICKOP** during the actual search; during the actual search, it just needs to be given the current knowledge state and it will then pick an operator to apply to it.) Then, we can say that a system has the capacity PICKOP just in case when it is given a knowledge state, it produces an appropriate operator to apply to that knowledge state. The situation is depicted below in Figure 4.

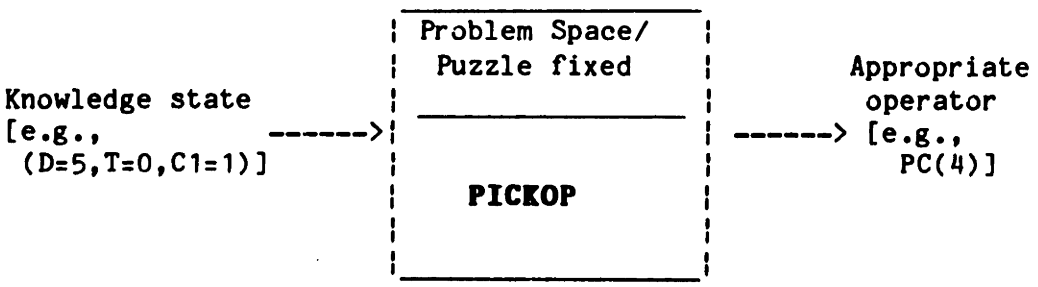

Figure 4

The capacity PICKOP will be meaning dependent just in case regarding an organism or system as having the capacity depends on a semantic characterization of some state or states of the system. Shortly, I will argue that regarding a system as having this capacity depends on a semantic characterization of both the input to and output of the system, and conclude that the capacity PICKOP in E is meaning dependent.[17] In so arguing, I am denying that someone could reasonably hold of a particular system: (a) that it has the capacity described above; and (b) that its inputs and outputs are syntactic objects which are never regarded as having an interpretation.

Given the way I have been using the term 'operator,' my claim might seem trivially true. For I have been using the term to refer to procedures for transforming knowledge states. Clearly, a syntactic object is not a procedure for transforming knowledge states (though it may designate or

---

17. Implicit here is the assumption, which I won't bother to defend, that the input and output of the system on a particular occasion are states of the system. More precisely, the assumption is that the input (or output) to the system having a particular value is a state of the system.

represent one). If operators are procedures, it is not possible to regard a system whose sole output is an uninterpreted syntactic object[18] as selecting an operator at all, much less an appropriate one. But, we needn't stick to this use of the term 'operator.' Associated with every operator/procedure in a problem space--PC(4) or TD(G,3), e.g.--is a syntactic item --'PC(4)' or 'TD(G,3),' e.g.--and it is perfectly natural to use the term 'operator' to refer to these syntactic items. In order to be as fair to the dispensabilist as possible, I will use the term 'operator' in this way. And, when I want to talk about the procedure itself, I'll use that term. One more terminological clarification is in order before we return to the question of whether the capacity PICKOP is meaning dependent. I have sometimes used the term 'operator' to refer to general procedures for transforming knowledge states--PC or processing a column, for example--and sometimes to refer to the specific procedures which result when the parameter(s) for the general procedure has been fixed--PC(4) or processing the fourth column, for example. From now on, I will usually be talking about these specific procedures. Unless it is clear that I have the general procedure in mind, the term 'procedure' will refer to these specific procedures, and the term 'operator' will refer to the syntactic items that designate them.

5.2

By an appropriate operator, as I have said before, I mean an operator that given the current state of the puzzle, as encoded in the knowledge state, could reasonably be expected to lead toward progress in the solution

---

18. When I use the term "uninterpreted syntactic object," I mean a syntactic item that is not assigned an interpretation for <u>any</u> purpose.

of the puzzle. Whether an operator is appropriate or not apparently depends on the interpretation of both the operator and the knowledge state which is the input to **PICKOP** (or whatever system is in question). To see this, suppose that Solver is solving the puzzle DONALD + GERALD = ROBERT. And, suppose that **KSET** and **OPSET** have selected a knowledge set and set of operators/procedures respectively and delivered these to **PICKOP**. Finally, suppose that **PICKOP** is given the input '(D=5,T=0,C1=1,E=9),' and produces the string of characters 'PC(4)' as output. Has **PICKOP** selected an appropriate operator? We simply don't know enough to say. To know whether the operator is appropriate, we need to know what procedure the operator 'PC(4)' designates. And, we need to know what partial solution to the puzzle '(D=5,T=0,C1=1,E=9)' represents. Only then can we determine whether or not the operator that has been selected is appropriate.

Suppose that the knowledge state '(D=5,T=0,C1=1,E=9)' is to be interpreted as before. That is, suppose that it encodes the following partial solution to the puzzle:

```
    1
  5ONAL5
+ G9RAL5
  ------
  ROB9RO
```

Now look at the third column from the right. The E in ROBERT has already been assigned the digit 9. Hence, there is only one letter in the third column which hasn't been assigned a digit. In such a situation, it shouldn't be hard to narrow down the digits which can be assigned to that letter. (In this particular case it is easy to narrow it down to one possibility: A must be assigned the digit 4.) In the problem space described earlier, there was a procedure that would focus on a particular column in the puzzle, and determine any new information which can be

inferred just from what is already known about that column. This procedure--focusing on a particular column and determining the consequences of what is already known about that column--was called processing a column. Given the fact that there is only one unknown letter in the third column from the right, and nothing is known about it, the procedure of processing the third column would be certain to generate new information. Hence, in the present situation, an operator that designates the procedure of processing the third column from the right would be an appropriate operator. Now look at the fourth column from the right. Since it contains three unknown letters, the procedure of processing the fourth column from the right would be fruitless in the current situation. An operator that designates the procedure of processing the fourth column from the right would not be an appropriate operator. What does all of this tell us about the operator 'PC(4)'? Well, that depends on what procedure the expression designates. If 'PC' designates the procedure of processing a column, and the '4' indicates that the fourth column from the right is to be processed, then 'PC(4)' is not an appropriate operator. If, on the other hand, the '4' indicates that the fourth column from the left is to be processed, then, of course, this being the same as the third column from the right, 'PC(4)' will be an appropriate operator. Finally, if 'PC' designates some other procedure, then until we know more about that procedure, we can't make any determination about the appropriateness of 'PC(4).'

The appropriateness of the operator 'PC(4)' will vary with the interpretation of the knowledge state as well. We could imagine that the knowledge state '(D=5,T=0,C1=1,E=9)' is to be interpreted differently. Perhaps the letter 'E' as it occurs in this knowledge state is not used to encode what is known about the 'E' in the puzzle. Perhaps it is used to

encode what is known about the 'G' in the puzzle. So, the knowledge state '(D=5,T=0,C1=1,E=9)' might encode the following partial solution to the puzzle:

```
   1
  5ONAL5
+ 9ERAL5
 ------
  ROBERO
```

In this situation, processing the fourth column from the left would be fruitless, since we have three unknown variables (the two letters and the carry). So, assuming that the operator 'PC(4)' designates the procedure of processing the fourth column from the left, the operator will be appropriate on the earlier interpretation of the knowledge state but not on this interpretation.

Given a particular knowledge state and a particular operator, whether the operator is appropriate or not depends on how the knowledge state and the operator are to be interpreted. I do not mean by this that there is no fact of the matter as to whether a given operator is appropriate relative to a given knowledge state, a given puzzle, and a given problem space. A complete characterization of a problem space must tell us how to interpret knowledge states and operators. It was implicit in my discussion of the problem space in section 2 that 'PC(4)' designates the fourth column from the right and that knowledge states are to be interpreted in the standard way. Hence, given that problem space, the operator 'PC(4)' is not appropriate relative to the knowledge state '(D=5,T=0,C1=1,E=9).' The point is not that such facts are not determinate. Rather, it is that they are determinate only because it is determinate <u>how these expressions are to be interpreted</u>.

Consideration of the examples adduced in the paragraphs above simply dramatizes what we should have known already: given any operator and any knowledge state, the appropriateness of the operator relative to the knowledge state is determined not by the syntactic form of these items but rather by the nature of the procedure designated by the operator and the nature of the partial solution to the puzzle represented by the knowledge state. The notion of appropriateness appealed to in section 4 applies directly to a procedure for transforming partial solutions to a puzzle, which procedure is to be applied to some particular partial solution. The notion of appropriateness applies only indirectly to syntactic objects such as operators and knowledge states, and only then when they designate procedures and partial solutions respectively. The notion of appropriateness does not apply to an operator at all if the operator and associated knowledge state are regarded as syntactic objects that are never assigned an interpretation. If we think of a system as receiving a knowledge state as input and producing an operator as output, but insist on regarding these as uninterpreted syntactic objects, we could never regard the system as producing an operator that is appropriate relative to the knowledge state it received as input. The capacity to select an appropriate operator, then, is quite clearly meaning dependent.

5.3

Faced with the meaning dependence of PICKOP, the dispensabilist must somehow syntacticize the capacity. She might attempt to follow the example of the recharacterization of the capacity to compute the sum of three digits (where the third is either 1 or 0), discussed in the previous section. There the idea was essentially to list all the possibilities. The capacity to add three digits became the capacity to produce the

expression '2' if the first two inputs are the expression '1,' and the third the expression '0'; to produce the expression '3' if one of the first two inputs is the expression '2,' the other the expression '1,' and the third the expression '0'; etc.

In principle, the dispensabilist can do something similar for the capacity to select an appropriate operator. Let's continue to assume that the puzzle being solved and the problem space being used to solve it are fixed. The knowledge set described in the problem space (call it P) contains only a finite number of knowledge states. Hence, only a finite number of strings of characters are knowledge states in P. Similarly, the set of operators/procedures in P determines only a finite number of strings of characters to be operators. Now, consider the cartesian product of these two sets of strings. That is, consider the set of all ordered pairs of strings such that the first element in the pair is a knowledge state in P and the second element is an operator in P. Call this set S_P. For any given pair in this set, we can consider the question of whether the second element would be an appropriate operator if the first were the current knowledge state. Whether it is will obviously depend on what procedure the operator designates and what partial solution to the puzzle the knowledge state represents. But, of course, the problem space P determines all of this. Hence, we may consider the subset of S_P consisting of all pairs such that the second element of the pair is an appropriate operator relative to the first pair. Call this set S^*_P. Now suppose that there were some way of characterizing this set without in any way appealing to the semantic interpretation of the relevant strings of characters. Suppose, for example, that the set could be characterized extensionally--by listing all its members. Then, the dispensabilist could recharacterize the capacity

PICKOP (at least as it applies to the particular puzzle and problem space in question) as the capacity to select an operator O, whenever it is given a knowledge state K, such that the ordered pair (O,K) is an element of the set S^*_p.

Of course, listing all the elements of S^*_p would be an immense task: the number of elements is enormous. Consider, for example, the number of knowledge states in the problem space we have been using. Even if we employ certain restrictions--counting strings with the same elements in a different order, like '(D=5,T=0)' and '(T=0,D=5)', as the same string, for example--to keep the number down, the number of knowledge states must be well over 10^{10}.

But, let us suppose that such a list could be produced and, therefore, that a purely syntactic characterization of the capacity PICKOP given, that is that it could be recharacterized as the capacity to select one of the expressions '___', '___', or '___' if the current knowledge state is '_____'; to select one of the expressions '___', '___', or '___' if the current knowledge state is '_____'; etc. Let us call this syntacticized capacity PICKOP'. The enormous size of the list necessary to describe PICKOP' will, I claim, destroy the effectiveness of the explanation.

The situation is analogous to the central example in Putnam's "Reductionism and the Nature of Psychology." Putnam considers two possible explanations of why a square peg whose sides measure slightly less than 1" won't pass through a round hole of diameter 1" in a board, but will pass through a square hole whose sides measure exactly 1" in the same board. The first explanation "is that the peg is approximately rigid under transportation and the board is approximately rigid. The peg goes through

the hole that is large enough and not through the hole that is too small." (Putnam, 1973, p. 131.) The second possible explanation considers the peg and the board as clouds of (Newtonian) elementary particles. These characterizations include the precise position and velocity of each particle in the board and the peg at some particular time. It is then supposed that some immense calculation is performed which shows that the collection of particles that the peg comprises may pass through the spatiotemporal region that the square hole comprises but will not pass through the spatiotemporal region that the round hole comprises.

Putnam rejects the second proposal as an explanation on the grounds that "whatever the pragmatic constraints on explanation may or may not be, one constraint is surely this: The relevant features of a situation should be brought out by an explanation and not buried in a mass of irrelevant information." (Putnam, 1973, p. 132.)

An explanation of Solver's ability to solve puzzles (or, more precisely, DONALD + GERALD = ROBERT, since we have been holding the puzzle fixed) that uses the capacity PICKOP' can be rejected on the same grounds. The explanation fails to bring out the contribution that exercise of the capacity PICKOP' makes in producing a solution to the puzzle. It hides this contribution in a mass of irrelevant information. The mass of irrelevant information makes the explanation literally incomprehensible. Given the limits of human patience one could not really digest or comprehend the entire explanation: it would be too long. Considerations of comprehensibility are particularly relevant to homunctional explanations. One of the great virtues of homunctional explanations is that they can make very complex systems comprehensible by organizing them into layers of interacting parts. Even though one can't understand the

operation of the system all at once, at a low enough level, the operation of the components can be comprehended directly, and comprehensibility then filters back up.

We needn't even go so far as Putnam does and claim that the inclusion of enormous amounts of irrelevant information prevents the account that uses the capacity PICKOP' from being an explanation. We need only claim that it severely weakens the explanation. Because it does, the dispensabilist will have to find some other way of generating a syntactic characterization of PICKOP, or concede that PICKOP is meaning dependent.

5.4

Another tack the dispensabilist might take in recharacterizing PICKOP would be to replace the notion of an appropriate operator with the notion of a "syntactically effective" operator. Suppose that the notion of progress in the search for a solution to the puzzle could be defined syntactically. Then, the dispensabilist could say that appropriate operators are simply strings of characters such that when they are used as input, along with the current knowledge state, to **EXEC**, progress in the search for a solution to the puzzle is thereby produced.

How could the notion of progress be defined syntactically? Well, recall that one of the elements of a problem space is u_0, the set of those knowledge states which count as a solution to the puzzle at hand. In SYS, **SOLSET** is responsible for producing a characterization of this set. The output of **SOLSET** is delivered to **TEST**, which then creates an algorithm which sorts members of u_0 from other knowledge states. The dispensabilist can use **TEST** to get a syntactic characterization of the knowledge states in u_0 (solution states): a given string of characters is a solution state

just in case it is a knowledge state and would elicit a positive response if used as input to **TEST**. Some knowledge states (continuing to think of these as strings of characters) will not be solution states, but will be very close to solution states. For example, many knowledge states will have the property of differing from a solution state by just one character. It is not hard to imagine that such properties can be used to develop an overall measure of the extent to which any given knowledge state differs syntactically from a solution state. First we would develop a measure of the extent to which any two knowledge states are syntactically different. It could be as simple as the number of places in which the two strings have the same character divided by the average length of the two strings. But, it is possible to be more sophisticated--we could, for example, give a bonus for having one entire syntactic component in common, or having the same number of syntactic components. In any case, having defined a measure of the syntactic difference between any two knowledge states, the measure of a knowledge state's closeness to a solution state would simply be the smallest number m such that for some solution state the syntactic difference between the knowledge state and that solution state is m. Using this notion of syntactic closeness to a solution state, the dispensabilist can now define the notion of syntactic effectiveness.

An operator is syntactically effective (relative to a given knowledge state, puzzle and problem space) just in case, when the operator is used together with the knowledge state as input to **EXEC**, the output of **EXEC** is a knowledge state that is syntactically closer to a solution state than the current knowledge state.[19]

---

19. There are other ways that one might define the notion of syntactic effectiveness. For example, one might define the distance between

Appropriate operators will often be syntactically effective, and vice-versa. When the procedure designated by an appropriate operator is applied to a knowledge state, the resulting knowledge state typically contains more information about a solution to the puzzle. And, typically, if not always, a knowledge state that contains more information about a solution is a knowledge state that is syntactically closer to a solution state. Hence, typically, when a knowledge state and an appropriate operator are used as input to **EXEC** the result will be a knowledge state that is syntactically closer to a solution state than the original. Therefore, an appropriate operator will typically be a syntactically effective operator.

At first glance, then, it is plausible that PICKOP--the capacity to select an appropriate operator given a knowledge state--can be recharacterized as the capacity to select a syntactically effective operator given a knowledge state. Clearly, however, the capacity to select an appropriate operator is not precisely identical to the capacity to select an effective operator. Some appropriate operators are not effective, and some effective operators are not appropriate. Perhaps the dispensabilist can argue that the discrepancy is so trivial that, for the

---

knowledge states on the basis of the minimum number of steps between the two knowledge states. Two knowledge states are separated by a single step if one can be gotten from the other by applying one of the procedures in the problem space being used. This way of putting it appeals to the procedures that the operators in the problem space designate. But, the dispensabilist could instead say that knowledge state A is a single step from knowledge state B if there is some operator in the problem space such that when it and A are the inputs to **EXEC**, B is the result. With this notion of distance, the dispensabilist could once again say that an operator is syntactically effective if it results in a knowledge state that is closer to a solution state than the original knowledge state. The comments that follow, particularly the remarks about "dead ends" apply as well to this definition of syntactic effectiveness or any other variations on the theme.

purposes at hand, the capacity to select an appropriate operator and the capacity to select a syntactically effective operator may, in fact, be considered the same capacity (under different descriptions).

I don't, in fact, think that it can be argued that the difference between the notion of an appropriate operator and a syntactically effective operator are trivial. Often in solving problems we find ourselves pursuing dead ends--following strategies which, though perfectly reasonable, end up bringing us no closer to a solution to the problem at hand than we were when we embarked on them. Newell and Simon's theory of puzzle solving is designed to allow for the pursuit of dead ends. Most of the problem spaces they consider are designed in such a way that a knowledge state can encode within it a means of indicating where things stood before a certain speculative strategy was begun. In this way, when a strategy does not pan out, an operator can be selected which has the effect of canceling the strategy out.[20] Newell and Simon describe the advantage of having these operators as follows:

> There are several reasons why a system will usually be more efficient that retains some capability for returning to previous knowledge states. A knowledge state may contain false information--as a result of errors in processing or recall. A knowledge state may also contain conditional information, where assumptions are made deliberately in order to work out their consequences (the conditional assignments of cryptarithmetic and the alternative moves of chess are both examples of conditional information). (Newell and Simon, 1972, p. 816.)

The important point about the pursuit of dead ends is that it will generally involve the application of procedures that produce knowledge states syntactically <u>further</u> from a solution state than the original knowledge state. In other words, the pursuit of dead ends involves the

---

20. See Newell and Simon, 1972, p. 273 for an example.

selection of operators that are not syntactically effective, although they are appropriate. Suppose, as Newell and Simon do, that the pursuit of dead ends is sometimes an important aspect of the way we solve puzzles, and therefore an important aspect of the way Solver solves puzzles. Then **PICKOP**'s sometimes selecting an ineffective operator will be an inherent in the way Solver solves puzzles, and it would be a major mistake to attempt to explain Solver's ability to solve puzzles in terms of **PICKOP**'s capacity to select an effective operator, even if it generally does select an effective operator. Given the importance that Newell and Simon place on dead ends, then, the attempt to recharacterize PICKOP as the capacity to select a syntactically effective operator will not work.

5.5

In 5.1 and 5.2 I argued that the capacity PICKOP is meaning dependent. We cannot judge an operator to be appropriate or not without knowing the procedure that is designated by the operator and knowing how to interpret the knowledge state that the procedure will be applied to. In section 5.3 I considered the possibility of syntacticizing PICKOP by means of a complete list of knowledge state-operator pairs, where the operator is appropriate for the knowledge state it is paired with. This possibility was rejected on the grounds of simplicity: it would obscure our explanation with a huge amount of irrelevant information. Finally, in section 5.4, I considered a syntacticization in which the notion of an appropriate operator is replaced with the notion of a syntactically effective operator. I argued that although most appropriate operators are syntactically effective operators and vice-versa, the two notions differ in important ways, and thus that the capacity to select a syntactically

effective operator cannot be regarded as identical to the capacity to select an appropriate operator.

I conclude that there is no apparent way to recharacterize the capacity PICKOP, and that PICKOP, in E, is, therefore, essentially meaning dependent. In section 2, we had an example of a meaning dependent capacity which could be syntacticized, but at the price of narrowing the scope of the accompanying explanation. Given the assumption that we want the explanation to be as general as possible, we were able to say that the explanation was weakened by the syntacticization and that the capacity was, therefore, meaning dependent. Here we do not depend on the assumption that the explanation should, if possible, be of broad scope. Instead, the argument relies on the fact that the only syntacticization available obscures the explanation with an enormous amount of irrelevant information.

Before proceeding to an example of another kind of reason for holding a capacity to be meaning dependent, I want to consider the present example just a bit longer. In particular, I want to consider an argument to the effect that there must be another way to provide a syntactic recharacterization of the capacity PICKOP. The argument is based on the "formality condition."

The formality condition, originally proposed by Fodor, and very widely accepted, places a constraint on theories of mental processes.[21] It says that mental processes must apply to mental representations in virtue of the formal--i.e. syntactic, or non-semantic--properties of the mental representation.

---

21. See Fodor, 1980, p. 64.

Now, consider the processes involved when Solver selects an operator. The formality condition ought to hold since we are assuming that Solver solves puzzles just like people do. Also, Solver is composed of conventional digital computers, and the formality condition holds almost by definition for such machines. So, the process of selecting an operator applies to the relevant representations in virtue of their formal properties. In other words, PICKOP selects an appropriate operator on the basis of the formal properties of its input. This suggests that the relation which holds between the input to PICKOP and the operator it selects can be defined formally. Couldn't, one might ask, this definition be used in turn to provide a syntactic characterization of the notion of an appropriate operator? More generally, if a process satisfies the formality condition, then mustn't there be a way of characterizing the capacity reflected by that process formally or syntactically?

There are a number of problems with deriving a syntactic characterization in this way, but I'll focus on just a couple. Both derive from the fact that capacities characterized in this manner will be much too narrow. Suppose we use the formal algorithm or program used by PICKOP to define the notion of an appropriate operator. Then, there will be one and only one appropriate operator associated with any given set of inputs (a puzzle description, a characterization of a problem space, and a knowledge state). But, in fact, given any particular set of inputs, there may be a number of appropriate operators, all but one of which will be excluded by the proposed definition. Hence, the notion captured by the proposed characterization is not the notion of an appropriate operator, but something much narrower.

An explanation which depended on such a narrower capacity would have two sorts of defects. First of all, since it depends on the manner in which Solver happens to select an appropriate operator, it could not be applied to systems or organisms which select operators in a slightly different way. In other words, the explanation would have the defect of not applying to systems which employ different algorithms than those used by Solver, but still, at a certain level of abstraction (in particular, the level associated with the amount of detail present in the explanation E), solve puzzles in the same way.

Another problem with the proposed explanation has to do with the importance of accounting for the circumstances in which a system fails to exhibit a target ability. As I noted earlier in the discussion of homunctional explanations in section I, there may be some slack between a particular capacity--be it the target capacity or one of the subcapacities in the explanation--and what will actually be achieved by the subcapacities posited to account for that capacity. Suppose, for example, that the capacity C appears in an homunctional explanation, and subcapacities S_1, S_2, ..., S_n, interacting in manner M are posited to account for the system's having capacity C. It may be that there are certain stimuli such that the execution of the subcapacities in the relevant manner will not be sufficient for the system to achieve C. This situation is tolerable so long as such cases can be shown to be exceptional and to correspond to stimuli for which the system actually fails to exhibit the target ability. If the system in question is the human visual system, then among such stimuli will be those that produce optical illusions.

In the case of Solver and the capacity PICKOP, we want to acknowledge that there may be circumstances in which Solver fails to select an

appropriate operator. In fact, these failures may exhibit some sort of systematicity. We would like a detailed homunctional explanation of Solver's abilities to reveal the precise locus of the failures.[22] The failure might result, for example, from the fact that the initial set of subcapacities posited to account for the capacity PICKOP are not always sufficient for the selection of an appropriate operator. Or, the problem might appear later, in the decomposition of one of these subcapacities. Standard homunctional explanations will reveal the locus of difficulty wherever it might be.

Suppose, however, that the syntactic recharacterization of PICKOP is derived from the actual algorithm used by Solver. PICKOP then becomes the capacity to select whatever operator Solver actually selects in a given situation. All of Solver's selections are treated as successes. The causes of Solver's failures to select an appropriate operator will be obscured.

Using the algorithm actually used by Solver to select an appropriate operator to derive a syntactic recharacterization of PICKOP evidently produces a capacity that is too narrow. And, in general, the formality condition does not guarantee that a given nominally meaning dependent capacity of a system can adequately be given a syntactic recharacterization.

---

22. By a detailed homunctional explanation I mean one that contains levels of analysis beyond those included in E.

6. Dispensabilist Explanations and Insight

In this section I will put the homunctional explanation of Solver's ability to solve puzzles aside and focus instead on a particular episode of puzzle solving. The accounts of that particular episode offered by the dispensabilist and indispensabilist respectively will be compared. The point here will largely be tangential to the main discussion, since it is not essentially tied to the structure of homunctional explanations. My claim, that the dispensabilist explanation gives less insight into how the puzzle is solved, could be transferred to the general homunctional explanation which has been at issue since section 3. I won't do this because the argument is too lengthy to make it worth presenting here. Still, it is important to see the kind of effect that dispensabilism can have on the extent to which an explanation of a cognitive process gives us insight into that process.

Suppose once again then that Solver is given the puzzle, DONALD + GERALD = ROBERT; D=5. And, suppose that Solver solves the puzzle in the manner indicated in section 2. Now consider how Solver's activities will be described by the indispensabilist and by the dispensabilist in turn. The broad outlines of the descriptions will be the same. Solver begins by selecting a problem space comprising a solution set, a knowledge set, a set of operators and an initial knowledge state. Solver then produces a sequence of knowledge states, with the step from one knowledge state to the next being guided by the selection of an operator.[23] The accounts will

---

23. Notice that the account will fit neatly with the general homunctional account given earlier. In general, it is easy to transform a general homunctional account of a system's capacity to X into an account of a particular episode of Xing: the more complete the homunctional account, the more detailed the account of the particular episode.

differ in that the dispensabilist regards each of these structures as uninterpreted syntactic objects, whereas the indispensabilist regards each of them as an expression with a meaning or interpretation as well as a syntactic form.

In presenting the sequence of knowledge states and operators, the dispensabilist might produce something like the chart that appears at the end of section 2 and is also presented below:

(D=5)	PC(1)
(D=5,T=0,C1=1)	PC(5)
(D=5,T=0,C1=1,E=9,C4=1,C5=1)	PC(3)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1)	PC(4)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R>6)	PC(2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R>6, R odd)	GN(R)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7)	PC(2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8)	PC(6)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1)	GN(N)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=2v3v6)	TD(N,2)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=3v6)	AV(N)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6)	PC(4)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6,B=3)	GN(O)
(D=5,T=0,C1=1,E=9,C4=1,C5=1,C3=0,A=4,C2=1,R=7,L=8,G=1,N=6,B=3,O=2)	

This chart, combined with the dispensabilist's description of the problem space that Solver selected in solving the puzzle, produces a succinct and cogent account of how Solver solved the puzzle. The description of the knowledge set allows us to interpret each of the knowledge states in the chart above, except for the last one, as a partial solution to the puzzle. And, the description of the set of operators will provide an explanation of what each of the procedures indicated in the column on the right do. Hence, we are able to see each knowledge state as a step toward the final solution of the puzzle, and we are able to see how Solver gets from one step to another. Consider, for example, the step from the tenth line on the chart to the eleventh line. The status of the puzzle represented on the tenth line is:

	11011
	50N485
	197485

N = 2 or 3 or 6	70B970

At this point Solver chooses the operator 'TD(N,2),' which designates the procedure of testing the assignment of the digit 2 to the letter N for plausibility. Of course, if you suppose that N=2, then the third column from the left will sum to 9, and B will have to be assigned the digit 9. But, we already have E=9. So, you can conclude that N does not equal 2. Given that N is equal to 2 or 3 or 6, you can further conclude that N is equal to 3 or 6. And, this is exactly the conclusion that Solver reached by testing the plausibility of N=2. This conclusion is reflected in the knowledge state on the eleventh line above, which differs from the previous knowledge state only in that 'N=3v6' occurs where 'N=2v3v6' occurred before. By looking at the step from the tenth line to the eleventh line, we see that Solver, "knowing" that N is equal to 2 or 3 or 6, tests the plausibility of N being equal to 2; determines that N can't be 2; and, "concludes" that N must be 3 or 6. Thus, we understand to a large degree how Solver got from the ninth to the tenth step in the solution of the puzzle. We can similarly understand each of the other steps in the chart and thereby see how Solver, step by step, solves the puzzle.

So, putting the above chart together with a description of the problem space used, we get a cogent account of how Solver solved DONALD + GERALD = ROBERT; D=5. The explanation is, of course, incomplete--in just the same way that our account, E, of Solver's ability to solve puzzles was incomplete. We still don't know how Solver selected the problem space; and

we don't know in full detail how Solver tested the plausibility of N being equal to 2; etc. Still, it provides an acceptable answer to the question, How did Solver solve the puzzle? It is just the sort of answer we would want if a person had solved the puzzle and we had asked her to explain how she solved it.

Now compare the above account with the one available to the dispensabilist. The dispensabilist can also produce the chart above. But, of course, for the dispensabilist, the symbols on the chart are uninterpreted syntactic objects. Let's emphasize this fact by replacing the set of symbols with ones that do not suggest a particular semantic interpretation. The result is given below:

```
#"%d#                                                              ?#a#
#"%d$&%z$!a%a#                                                    ?#e#
#"%d$&%z$!a%a$)%i$!d%a$!e%a#                                      ?#c#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a#                       ?#d#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+me#                    ?#b#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+me$+~#                 =#+#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g#                    ?#b#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n#                ?#f#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a#            =#;#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a$;%bocoj#    :#;&b#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a$;%coj#      >#;#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a$;%j#        ?#d#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a$;%j$`%c#   =#*#
#"%d$&%z$!a%a$)%i$!d%a$!e%a$!c%z$^%f$!b%a$+%g$}%n${%a$;%j$`%c$*%b#
```

This chart, like the previous one, shows Solver solving the puzzle by going through a sequence of knowledge states. Each step, the dispensabilist may explain, is mediated by the operator listed on the right. The operator triggers a procedure which transforms one knowledge state into the next. And, the dispensabilist use her characterization of u_0, the set of solution states, to prove that the last knowledge state above is, indeed, a solution state. Thus, the dispensabilist can show how Solver, step by step, constructs a solution to the puzzle. But, of course,

the explanation is very unsatisfying. We still don't have any insight into how Solver was able to solve the puzzle.

The lack of insight in the indispensabilist's explanation should be familiar to anyone who has watched a Rubik's cube expert at work without being privy herself to the special techniques used to realign the cube.[24] You can watch each step very carefully. You can see, mechanically, how the person gets from one step to another by rotating a particular face of the cube. And, it is obvious that the state of the cube after the last step is a "solution" to the cube. Still, if one has struggled for hours unsuccessfully attempting to solve the cube, it seems like magic when the expert produces a realigned cube. It is only when the expert reveals her technique, categorizing moves in a variety of ways, that one comes to have some insight into how she solved the cube.

7. Essentially Meaning Dependent Capacities: The Extent of the Problem

The target capacities of homunctional explanations in cognitive science will almost always be meaning dependent. It is impossible to regard a system as doing the sort of things that cognitive scientists want to explain--seeing, hearing, reading, reasoning, problem-solving,

---

24. Rubik's cube, which was popular in the early 1980's, is a plastic cube consisting of 27 smaller cubes. Each of the smaller cubes (except for the center cube) have a different color on each face, the same six colors being used for each cube. Any face of the larger cube, which itself consists of nine smaller faces, can be rotated. These rotations allow one to change the outer appearance of the larger cube. Starting with a cube in which the colors on the faces have been "scrambled," the object is to get each face of the larger cube to be uniform in color. So, for example, the cube is solved if one face is all red, one all green, one all blue, one all yellow, one all white, and the sixth face all orange.

remembering, etc.--without associating an interpretation with some of the system's states. Meaning dependent capacities are, then, quite typical. Meaning dependent capacities will be essentially meaning dependent if they cannot be replaced by syntactic alternatives without substantially weakening the explanation in which they occur. In sections 2 and sections 3 through 5, I discuss two different ways in which syntacticizing meaning dependent capacities can weaken the surrounding explanation: (1) the resulting explanation will sometimes be less general than the original explanation; and, (2) sometimes the only syntactic alternative available will require such an enormously lengthy characterization that it burdens the resulting explanation with too much irrelevant detail. So far we have focused on a single example to illustrate each possibility. In this section I want to consider some other examples, and address the question of how often we can expect to encounter each of these kinds of cases. I will argue that cases of the first sort will be ubiquitous in cognitive science, and cases of the second sort, while less common, will not be unusual.

7.1

The example in section 2 concerned an nomunctional account of our ability to add pairs of numbers. The capacity at issue was the capacity to compute the sum of three digits. The proposed syntactic recharacterization was: the capacity to produce the expression '1' when given two tokens of the expression '0' and one token of the expression '1'; to produce the expression '2' when given two tokens of the expression '1' and one token of the expression '0'; etc. I objected that the resulting explanation would then fail to apply to systems that calculate in bases other than base-10. The example worked because the capacities in the partial homunctional explanation of adding pairs of numbers do not determine, explicitly or

implicitly, precisely the set of syntactic formulae to be used by systems that the explanation applies to.

Homunctional explanations will often have this feature. Theories of cognitive processing can often be given in a fair amount of detail without specifying the exact form of the structures being manipulated. This claim can be thought of as an extension of the tenet which spawned functionalism--that a precise account of how a system processes information can be given without specifying the physical nature of the system. I'll further illustrate the claim with two examples.

Consider a theory of parsing. The theory assumes a particular theory of grammar, the latest version of Government-Binding theory let's suppose. The theory makes assumptions about the kind of structures that are computed during the course of parsing. There will be structures constituting the sequence of words to be parsed, structures constituting words which have been encountered but not assigned syntactic roles, structures constituting syntactic trees, structures constituting partial syntactic trees, etc. The theory will also include a number of specific principles which specify how these structures are built up in the course of parsing. For example, one such principle might have to do with determiners. Perhaps, whenever a determiner is encountered during processing of a sentence, a structure constituting a noun phrase having (at least) a determiner and a head noun[25] is created: the determiner is filled in of course, but the slot for the head noun is left empty. And, perhaps the slot is to be filled by the very next noun that is encountered. This is ensured by allowing empty slots in

---

25. The reader does not need to know what a head noun is in order to follow the discussion.

our syntactic structures to be "starred," and having a general principle which says that when a word of a certain category is encountered and an empty slot for that category is starred, the slot must be filled by that word.

Now, in an homunctional explanation of the ability to parse sentences based on our hypothetical theory, we might include the capacity to respond to a determiner by setting up a noun phrase whose constituents are (at least) the determiner, and a starred empty slot for the head noun. If the theory were presented in the fullest possible detail, and the precise form for all of the structures computed in the course of parsing were given then we may well be able to characterize this capacity syntactically. But, the theory could be given without specifying the precise form of the structures it posits. In particular, at the level of detail used to describe the principle above, it is an open question what the precise form of a noun phrase is. The noun phrase might have the form:

(NOUNPHRASE1 (("The"),(HEADNOUN*))).

But, it could also have the form:

(S1; S1 IS NOUNPHRASE; S1 CONTAINS "the"; S1 CONTAINS HEADNOUN*).

The two forms are clearly very different. Systems using the two forms would have to engage in very different sorts of computations. But, the theory, and hence, the homunctional explanation, can remain silent about which of these forms the structure is to take. And, this is so in spite of the fact that the theory may present a rather detailed account of parsing.

Consider another example. Marr's theory of vision involves the construction of a series of representations of the contents of the visual field. The earliest such representation is called the "raw primal sketch." It is derived from the array of light intensity values encoded at the back of the retina, with this derivation being mediated by the computation of zero-crossings.

A zero-crossing is a point in the visual field where the change in light intensity values is at a local peak. These often correspond to important points in the visual field such as a point on the boundary of an object. The zero-crossings are computed over a number of "channels." Intuitively, each channel blurs the image to a certain degree. The greater the blurring, the more zero-crossings are eliminated, and the greater the chance that the remaining ones are significant. Points that are associated with zero-crossings in all the channels are especially significant and become part of the raw primal sketch.

Now, simplifying somewhat, let's say that the raw primal sketch includes just those points that do appear as zero-crossings in each channel. These points are organized into various kinds of clusters, depending on the rough shape that they form. Among the kinds of clusters are edges, blobs and bars. Representations of the clusters are explicitly constructed and include information about the cluster's position, length, orientation, etc. These representations are the elements of the raw primal sketch.

Figure 5. This sequence, reproduced from Marr, 1981, depicts (left to right) an image, the zero-crossings identified by one channel in processing the image, and, finally, part of the information encoded in the raw primal sketch for the image--the location and orientation of edges.

In an homunctional account of vision following Marr's theories, one of the higher level capacities posited will be the capacity to construct the raw primal sketch. This will be analyzed into a number of capacities such as: the capacity to compute the zero-crossings in a number of channels; the capacity to isolate those points which correspond to zero-crossings at all channels; and, the capacity to include in the representation of the raw primal sketch an explicit representation of those clusters of such points that form an edge, including information about the edge's position, length and orientation; etc.

Now, consider this last capacity. In particular, consider the representation to be constructed. We must include in it information about the position of the edge. But, this information could be encoded in cartesian coordinates or in polar coordinates. Again, machines or organisms using one coordinate system will engage in very different sorts of computations from machines or organisms using the other coordinate system. And, any syntactic recharacterization of the capacity based on the

use of one kind of coordinates will not apply to machines or organisms using the other. But, an homunctional explanation based on Marr's theory, at the level of detail I have described it, need not specify which coordinate system is to be used.[26]

The parsing and vision examples illustrate well that detailed homunctional accounts of cognitive abilities can be given without consideration of the precise form that the structures being processed are to take. In such cases, syntacticizations of the posited capacities will typically result in explanations that are not sufficiently general. And, for this reason, such capacities will be meaning dependent and the explanations will constitute counterexamples to the thesis of the dispensability of meaning in cognitive science.

7.2

If the arguments above are correct, then the following situation will be common in cognitive science. A single partial homunctional explanation applies to a variety of (possible) systems or organisms, each of which has the target capacity. As far as the explanation goes, the systems and organisms all work in the same way, but at a more detailed level the systems and organisms work differently. The dispensabilist will not be able to give the general explanation which applies to all of the systems, but will, it has been assumed so far, be able to provide an homunctional explanation for any particular system. The example discussed in sections 3 through 5 suggests that that assumption may not always be correct. There,

---

26. In fact, Marr did specify the use of cartesian coordinates, and he may have had important reasons for doing so (Cf. Marr, 1982, p. 73.). But, the important point is that, stated at a certain level of abstraction, the theory can be neutral on this issue.

I considered the explanation of Solver's ability to solve puzzles. The capacity at issue was the capacity PICKOP. In that discussion, I did not rely on considerations of generality at all. The claim was not that a syntacticization of PICKOP would result in an explanation that was too narrow, but that, even as applied to Solver alone, the explanation resulting from a syntacticization would not be as effective as the original explanation.

In a moment I want to describe the crucial elements that made this example work, and suggest some other examples where the same elements are present. But, first a discussion of the general problem of providing syntactic recharacterizations of capacities will be helpful. The discussion will be rather abstract, but not completely unfamiliar, since it is just a general version of much of the discussion in section 5.

Suppose that we are considering a capacity C that occurs in an homunctional account of a particular system S's having a certain cognitive ability. Typically, we can think of a component (of a system) that has the capacity as a black box which takes expressions as its input and generates expressions in response. To keep things simple, we can restrict our attention to deterministic black boxes. Therefore, the black box will be a function, f, mapping a certain domain of expressions onto a certain range of expressions. The original characterization of the capacity tells us the relation, R, that the output of the black box--f(e)--must bear to the input--e--if the component actually has the capacity. And, the capacities of interest are ones that characterize this relation by making implicit or explicit reference to the semantic values of e and f(e) respectively. A syntactic recharacterization of the capacity will describe the relation R

that f(e) must bear to e without referring to any semantic properties of the two expressions.

Two general methods for characterizing R syntactically are of particular interest. First, assuming that the domain of inputs and the range of outputs are both finite, we can just list all the acceptable input/output pairs. A second method is available if R is a function--i.e., if there is only one acceptable output for any given input. The method hinges on the fact that the system S must have a component that has capacity C. This component must use some algorithm for taking an expression e and generating an expression that bears R to e. By describing this algorithm in complete detail, we can characterize a function f* such that o will bear R to i (assuming that R itself is a function) just in case o=f*(i).

The problem with the first method, of course, is that a list of all the acceptable input/output pairs may be so long that the explanation becomes literally incomprehensible. The second method has two hazards. First, if, according to the original capacity, there is more than one acceptable output for a given input, then the syntactically characterized capacity will be "narrower" than the original capacity. Since I won't be concerned with this hazard in the present discussion, I won't bother discussing the problems it creates. The second hazard and the one we will be concerned with here has to do with the fact that there may be some slack between the capacity C and what is actually achieved by f*. That is, it may be that for some input i, f*(i) does not really bear R to i. For some such i the original input to the system that generates i is an input on which the system fails, i.e. an input for which the system fails to manifest the ability A, even though, in general, it does have the ability.

Frequently, it is important that such inputs be treated as exceptions to be explained, as in the case of visual illusions. If f* is used to define a new relation R', then f*(i) automatically bears R' to i. Hence, S will be considered successful on all inputs. Failures are then covered up and not explained.

All of the above, rather abstract, discussion should have sounded somewhat familiar. For it parallels the argument given in section 5 concerning the capacity PICKOP and the explanation of Solver's ability to solve puzzles. The inputs in the case of PICKOP, it will be recalled, are knowledge states (I am assuming again that the puzzle being solved and the problem space being used to solve it are fixed) and the outputs are operators. A given output is acceptable if it designates an operator that is appropriate if the current status of the puzzle is represented by the input. The relation R* between two expressions thus defined must be characterized syntactically if the capacity PICKOP is to be syntactically recharacterized. In section 5 I argued that any such recharacterization would substantially weaken the explanation of Solver's ability to solve puzzles. The argument hinged on three points: (1) A list of acceptable input/output pairs would be so long as to render the resulting explanation literally incomprehensible; and, (2) It may be assumed that there are inputs for which Solver--PICKOP in particular--fails to select an appropriate operator. These failures need eventually to be explained, whereas using the complete algorithm used by PICKOP to syntactically characterize a slight variant of R* would treat the failures as successes. Hence, using this method of syntactically characterizing R* weakens our explanation of Solver's ability to solve puzzles. (3) No other plausible methods for syntactically characterizing R* seemed to be available.

In order, then, to find other examples of meaning dependent capacities without appealing to considerations of generality, we need only find examples where the three elements above are present. The most difficult requirement to satisfy is that there be no plausible methods, other than the two methods above, for syntactically characterizing the relevant relation, since the number of methods that could be proposed is limitless. Still, I think there are plenty of other cases where all three elements above are present.

Imagine, for example, an account of the ability to read. Such an account will involve issues of determining co-reference. In particular, an homunctional explanation of the ability to read will probably include the capacity to determine whether two expressions are co-referential. Perhaps the capacity will appear more than once in the explanation, each occurrence dealing with different kinds of sentences, or different kinds of pairs of expressions, or contexts that differ in other ways. And, details of the context might be built into the characterization of the capacity. Still, the capacity to determine whether two expressions are coreferential will appear in some guise.

The input for such a capacity will include representations, probably incomplete, of the sentence or sentences in which the two expressions occur; representations of some of the preceding sentences; some sort of representation synthesizing the most important information conveyed thus far in the passage; etc. The output will take one of two values, one indicating that the expressions are co-referential and the other indicating that they are not. Obviously, a listing of all acceptable input/output pairs would be too long. Even if the input contained nothing more than an unparsed 50-word section of the passage, the number of possible inputs

would be much greater than 10^{50}. Also, it is clear that whatever algorithm is used by the system or organism in question will fail on certain inputs, and that it will be important to explain these failures. It is not at all uncommon to read a piece of text and take two expressions to be co-referential when, in fact, they are not intended to be and a more careful reading of the passage makes this evident. A good account of the human ability to read will not treat such failures as successes.

Finally, consider the question of whether any other plausible method of syntactically characterizing the relation between inputs and acceptable outputs is available. Assume, in considering the question, that the capacity in question encompasses sentences such as (S1), where pragmatic considerations must be brought to bear:

(S1) By the time the father found the boy, he was in despair.

In this example, whether 'the boy' and 'he' are co-referential depends on whether the information in previous sentences suggest that the boy, but not the father, might be in despair. Whether it does is determined by the information in the input, which would include information about previous sentences. But, it determines it in a very subtle way. Outside of constructing a sophisticated algorithm for making such judgments, there clearly won't be any way of sorting cases of co-reference from cases of distinct reference on the basis of the syntactic features of the input.[27]

The key to this example is that the considerations that determine whether a particular input/output pair is acceptable are rather open-ended.

---

27. By syntactic features I mean, of course, the syntactic features of the input itself, not the syntactic features of the sentences that the input is about.

Consider, for example, the myriad of ways of suggesting that "the boy" might be in despair. This open-endedness has the consequence that (a) any particular algorithm for determining co-reference is likely to fail on some cases; and (b) except for using a sophisticated algorithm for determining co-reference, or simply listing all the acceptable input/output pairs, there won't be any plausible way to characterize acceptable input/output pairs on the basis of their syntactic properties.

We can find other examples of capacities that satisfy the three criteria above by duplicating this feature of open-endedness. That is, we want capacities that determine input/output relations the satisfaction of which is determined by an open-ended set of factors. Such examples should not be too difficult to generate. To briefly sketch another example, suppose that the target capacity is the capacity to formulate complex plans in some particular domain. Suppose that one of the posited subcapacities is the capacity to identify cases when two goals or subgoals conflict. Whether two goals or subgoals conflict may depend in a complex way on knowledge about the problem domain that is not explicitly encoded in the representations of the goals. Again, the factors that could determine that two goals conflict will be quite open-ended and any particular algorithm for identifying conflicts is likely to fail on some cases.

These examples and the analysis above make it clear that the explanation of Solver's ability to solve puzzles is not unique in its particular impact on the issue of dispensabilism. The important feature of the explanation was that it contained a meaning dependent capacity which the dispensabilist could not syntacticize without substantially weakening the explanation, even as it applies to Solver alone. We now see that other homunctional explanations will have this feature as well.

8. Conclusion

My aim throughout this essay has been to defend the concern of the Representational Theory of Mind with the meaning of mental representations. Such concern has come under attack by Stich, who has suggested that questions about the meaning of cognitive symbols can and should be discarded. My strategy has been to challenge the assumption that issues involving meaning can be ignored without paying a significant price. I have attempted to demonstrate that within a widely accepted framework for explanations in cognitive science claims about the meaning of mental representations have a central role to play.

The explanatory framework I have assumed is that of homuncular functionalism. Within this framework, the abilities of systems and organisms are explained by analyzing them into hierarchies of interacting subcapacities. When the ability to be explained is cognitive, the subcapacities often involve the manipulation of cognitive structures. And, we have seen that the characterization of the subcapacities frequently involves the association of interpretations or meanings with these structures. Recharacterizing the capacities so that they do not associate meanings with the structures being manipulated is frequently not plausible. In some cases, any such recharacterization would be excessively lengthy and difficult to comprehend. More frequently, the explanations resulting from such recharacterizations will be too narrow, applying to particular systems or small ranges of systems in cases where the original explanation applied very broadly.

To a large extent, the fact that syntactic recharacterizations of meaning dependent capacities are not always plausible, is a consequence of the structure of homunctional explanations. A complete homunctional explanation of a cognitive ability ends with subcapacities the possession of which can be accounted for directly by the physical constitution of the system in question. But, even when such an explanation is cut off part way down the hierarchy of subcapacities, the result is an explanation. And, I as discussed in the previous section, such explanations, even fairly detailed explanations, may not make any commitment, implicit or explicit, to the precise syntactic form of the structures it posits. It is not surprising, then, that characterizing the subcapacities in a way that depends only on syntactic properties of these structures is often not plausible.

If the existence of essentially meaning dependent capacities in homunctional explanations in cognitive science follows from the structure of homunctional explanations in general, then it is fair to say that homunctionalism and the thesis of the dispensability of meaning in cognitive science are incompatible. And, if homunctionalism is the central mode of explanation in cognitive science, then the price to be paid for eliminating appeals to meaning in cognitive science is very stiff indeed. A successful theory of meaning for mental representation may not be just around the corner, but perhaps some patience is in order.

The Syntactic Theory of Mind and the Collateral Information Problem

The Representational Theory of Mind (RTM), championed most notably by Jerry Fodor, has been offered as an account of propositional attitude states--believing that p, desiring that q, etc. According to RTM, for Fred to believe, e.g., that it's raining, is for Fred to stand in a certain relation to a token of a symbol (or sequence of symbols) in his head, the content of which is that it's raining. These symbols in Fred's head are part of a larger structure of symbols that constitute a language of sorts--a language of thought. And, the sequence of symbols can be thought of, on this view, as a sentence in a language of thought, or, as I shall call it, a mental sentence. Advocates of RTM have argued that propositional attitude states play a crucial role in cognitive science and, therefore, consider RTM to be a crucial element in the foundations of cognitive science.

One of the major challenges facing RTM is to provide a theory of content for mental sentences. Such a theory would answer the question, In virtue of what does a given sentence in someone's language of thought have the content that it does? Ideally, equipped with such a theory, and given the relevant non-semantic facts about Fred's psychology (whatever those might be), we could determine the content of any one of the sentences in Fred's language of thought.[1] The possibility of such a theory has come

---

1. Although RTM only requires that sentences be assigned content, it is often assumed that mental sentences derive their content from the contents of component symbols (together with some sort of composition rules). I will make this assumption, and will, therefore, talk about

under attack in recent years. One line of attack concerns what I shall call the "collateral information" problem (the CI problem for short). This name for the issue stems in part from Hilary Putnam's characterization of the problem as one of "developing a criterion which distinguishes changes in the content of mental signs from changes in collateral information." (Putnam 1983, pp. 146-47.)

Very briefly--I'll go into a bit more detail in section 1--the CI problem stems from the following sorts of considerations. Suppose that Fred, but not Barney, believes that Stanley has no hair on his head. This difference in belief surely does not suggest that the mental signs of Fred and Barney associated with the word "hair" have a different content. Surely, Fred's having this belief is merely a difference in collateral information with respect to the concept of hair. But, suppose that Fred believes that cats don't have hair on their bodies. Or, suppose that Fred believes that humans have no hair on their heads or bodies. Such beliefs suggest something more than a difference in collateral information. There is a strong inclination to think that they indicate a difference in the content of the two signs. It appears that a theory of content will have to give us a criterion for distinguishing beliefs which count as collateral information with respect to a particular concept from those beliefs that indicate a difference in content. But, there are reasons to think that such a criterion might not be possible. It seems, for example, that such a criterion would amount to something like an analytic/synthetic distinction, a distinction which is widely held to be untenable.

---

the content of a symbol or mental sign where the symbol is associated not with an English sentence, but an English word. In a similar vein, I will talk about the content of a concept.

Stephen Stich, in From Folk Psychology to Cognitive Science: The Case Against Belief (hereafter, The Case Against Belief) has used these sorts of considerations to argue against the possibility of an acceptable theory of content for mental sentences. His prescription is to abandon semantic notions in cognitive science altogether. On his view, cognitive science would recognize analogs to propositional attitude states. And, these states would be analyzed as relations to mental sentence tokens. But, the mental sentence tokens, and, hence, the propositional attitude analogs, would only be assigned syntactic properties; mental sentences would remain uninterpreted. Stich thus proposes to replace the Representational Theory of Mind with the Syntactic Theory of Mind (STM).

It is tempting to think that the collateral information problem is a problem of meaning and that sticking to syntactic properties of mental sentence tokens would avoid the problem altogether. Consider, for example, the analogous issues for real languages. It is not uncommon to suppose that if I acquire a series of bizarre beliefs about hair, that the meaning of the word in my idiolect might change. But, who has ever proposed that the acquisition of bizarre beliefs might change the syntactic properties of that term in my idiolect?

I shall argue, however, that the CI problem is not just a problem of meaning. I will argue that it is a problem for STM as much as it is a problem for RTM. And, I will argue that the powerful principle which Stich incorporates into STM in order to avoid the collateral information problem is untenable.

The outline of the rest of the essay will be as follows. In section 1, I will quickly review the collateral information problem and ways of

nandling it. In section 2, I will turn to tne version of tne CI problem put fortn by Stich and the difficulties he believes it creates for RTM. Section 3 will consider STM and tne way that it is supposed to avoid tne difficulties discussed in section 2. In particular, I will isolate a principle which is solely responsible for STM's avoiding the collateral information problem. Section 4 presents a detailed argument for tne conclusion that adherence to the principle isolated in section 3 creates insuperable difficulties for STM. STM must, therefore, abandon tnis principle and face the collateral information problem. Finally, I defend this conclusion against a plausible line of attack in section 5.

1. The Collateral Information Problem

The most well-known formulation of tne Collateral Information problem is given by Putnam in "Computational Psychology and Interpretation Theory" (Putnam, 1983). Putnam argues there tnat versions of RTM that advocate some version of conceptual role semantics[2] won't be able to develop a precise notion of sameness of content.[3]

In tne essay Putnam presents a thougnt experiment involving two children living in tne fictional country of Ruritania. At tne beginning of tne story, Oscar and Elmer are, we are told, "as alike as you please."

---

2. Conceptual, or functional, role semantics is the doctrine that the meaning of mental symbols is given by their conceptual, or functional, role. Field, 1977 presents the most worked out version of this view. See also Harman, 1982 and Block, 1986.

3. Putnam actually poses the CI problem as a problem for versions of RTM that advocate "a verificationist semantics." But, on my reading of Putnam, this is just another name for conceptual role semantics. See Putnam, 1983, pp. 143-144 in this regard. Putnam's arguments apply to any version of RTM in whose theory of content inferential role is a crucial determinant of meaning.

Let's assume that they are molecule-for-molecule identical. Furthermore, the only difference in their environments is that in the south of Ruritania, where Oscar lives, pots and pans are normally made of aluminum, whereas in the north of Ruritania, where Elmer lives, pots and pans are normally made of silver. In both dialects, the term 'grug' refers to the stuff pots and pans are made of. That is, in the north, 'grug' means silver and in the south, 'grug' means aluminum.

Oscar and Elmer are old enough when the story begins (age ten, let's say) so that both have acquired a concept that they express with the word 'grug.' Oscar and Elmer appear to have nearly identical beliefs about 'grug' (not surprisingly, given that they are molecule-for-molecule twins). In particular, Oscar and Elmer will say precisely the same things about 'grug.' Of course, the terms do have distinct reference in the two idiolects. And, the indexical terms they use, such as 'here,' 'my mother,' etc. will also have distinct reference. But, if we accept some version of individualism, the doctrine that the nature of a person's purely psychological states supervenes on the states of the person's brain, then we must say that Oscar and Elmer's purely psychological states are at this point identical. For, by assumption, they are in identical brain-states (assuming that one's brain states are supervenient on the state of the molecules in one's brain). In particular, we must say that the mental representations associated for them with the word 'grug' have the same narrow content.[4] Let's introduce some notation and call the mental

---

4. The term 'narrow content' refers to that aspect or determinant of content that supervenes on the states of a person's brain. That aspect or determinant of content that does not supervene on the states of a person's brain and, therefore, takes into account facts about one's linguistic community and the objects in the world with which one interacts, is called wide content. The importance of the distinction

representations of the two ten-year-olds $G_{O,10}$ and $G_{E,10}$ respectively. Individualism guarantees, then, that the narrow contents of $G_{O,10}$ and $G_{E,10}$ are identical.

Oscar and Elmer now proceed to learn the difference between the stuff called 'grug' in their part of the country and the stuff called 'grug' in the other part of the country. Oscar, for example, learns that the stuff he calls 'grug' is much lighter than the stuff they call 'grug' in the north of Ruritania. He learns that the two substances are different chemical elements. And, he learns that in the United States, the stuff they call 'grug' in the north of Ruritania is much more expensive than the stuff he calls 'grug.'

As adults, Oscar's concept of grug and Elmer's concept of grug are as different as my concept of aluminum is from my concept of silver. There appears, then, to be as much reason to believe that the contents of $G_{O,25}$ and $G_{E,25}$ are distinct as there is reason to believe that the contents of my concepts of aluminum and silver are distinct.

But, Putnam asks, if the contents of $G_{O,10}$ and $G_{E,10}$ are identical, and the contents of $G_{O,25}$ and $G_{E,25}$ are distinct, at what point did the contents of G_O and G_E first diverge? By hypothesis, all that happened is that Oscar and Elmer learned lots of facts about grug. Furthermore, the facts they learned were very pedestrian, the sort of facts that we acquire all the time. For this reason, it is difficult to come up with a plausible answer to Putnam's question. Suppose, for example, we say that the

---

between narrow and wide content is very widely agreed upon among advocates of RTM who accept individualism. (But, see Burge, 1986 for an attack on individualism.)

contents of the two concepts first diverge when one of the two children learns the atomic number of (their) grug. Then, it would seem, whenever an American child learns the atomic number of silver, the child's concept of silver comes to have a new content. And, if contents are supposed to be very stable at all, this does not seem plausible.

There is nothing crucially unique about Putnam's particular example (though it is particularly effective). The important point is that one can easily construct cases in which the steady accumulation of what appears to be collateral information apparently results in a mental representation coming to have a new content. It then becomes uncomfortable to have to say just when the change in content occurred.

There are, naturally, some ways out of the CI problem. I'll discuss two in particular. The first possibility is to embrace a graded notion of content according to which the acquisition of collateral information always changes contents, though only slightly. Another way to escape the CI problem is to insist that the acquisition of collateral information is accompanied by other important changes in psychological state and shift the burden of accounting for transformations of content to these other changes.

Ned Block alludes to the first approach in "Advertisement for a Semantics for Psychology." He suggests that the "crude dichotomy of same/different meaning" may have to be replaced by "a multidimensional gradient of similarity of meaning..." (Block, 1986, p. 629.) The chief problem with this approach to the CI problem is that a graded notion of content is very counterintuitive. On this view of content, every time I see a cat, and thus come to believe that there is now a cat in front of me, the content of my concept of cat changes ever so slightly. Most people, I

believe, have the strong intuition that the meaning of the concept does not change at all in such a circumstance, not even a little bit. Of course, if the theoretical reasons for adopting a graded notion of meaning are powerful enough, it would be reasonable to ignore these intuitions.

Fodor rejects the idea of a graded notion of content[5] and pursues the second approach mentioned above. In Psychosemantics: The Problem of Meaning in the Philosophy of Mind (Fodor, 1987), Fodor advocates the view that the narrow content of a symbol is a function that maps contexts into extensions. Put in terms of Putnam's famous twin-earth thought experiments, Fodor's view is that the content of the mental symbol shared by my doppelganger and me (the one associated with the word 'water') is a function. Applied to my context here on earth, the function picks out H_2O. Applied to my doppelganger's context on twin-earth, the function picks out XYZ. The function in question is determined by the counterfactual connections that hold in the relevant context between the mental symbol at issue and objects in the world. Fodor's proposal provides a nice response to Putnam's challenge. The reason G_O and G_E came to diverge in content is that, "whereas at first tokenings of 'grug' would have been elicited from either child by either aluminum or silver, at the end only silver controls 'grug' for Elmer and only aluminum controls 'grug' for Oscar." (Fodor, 1987, p. 94). The relevant question, then, is when aluminum, e.g., ceases to be in the relevant counterfactual relationship with G_E. And, Fodor

---

5. See Fodor, 1987, pp. 57-58.

points out, Putnam has not given any reason for thinking that there will be a particular problem in isolating such a point.[6,7]

The reasons for being wary of Fodor's way out of the CI problem concern the theory of content it rests on. The theory raises a number of difficulties, one of which I'll discuss in a brief way. The problem I have in mind concerns the existence of the relevant causal relations in the case of theoretical concepts. To see the difficulty, suppose that the relation in question holds between a mental symbol and a kind of object in the world just in case the presence of object of the relevant kind in someone's environment is guaranteed to cause a token of the symbol in the person's head. For observable entities one might hope to specify general environmental conditions under which such a relation will hold. But, to take an example that Fodor discusses, it is obvious that this can't be done for theoretical entities such as protons.[8] Fodor struggles mightily to modify the causal relation so that it could hold between protons and the relevant symbol. But, in doing so, he raises as many new questions regarding the proposal as he answers.[9]

---

6. Of course, this is not a criticism of Putnam's argument since acceptance of a conceptual role, or verificationist, semantics for mental representations is a premise of Putnam's argument.

7. Fodor's response to Putnam does seem to leave open the possibility that mental symbols acquire new contents more often than we might think. For it now seems as though it may not be uncommon for the acquisition of apparently collateral information to change the truth of the counterfactuals that determine the content of a particular symbol.

8. For a nice discussion of these problems applied to an earlier version of Fodor's theory of content, see Block, pp. 657-660.

9. See Fodor, 1987, pp. 112-127.

I do not mean to suggest that the responses to the CI problem suggested by Block and Fodor will not prove to be successful. The point is simply that the CI problem does seem to require the advocate of RTM to embrace controversial proposals. If it were true that the CI problem is so linked to issues of meaning that it does not arise for STM, then it might be reasonable for someone who is otherwise sympathetic to RTM but unsatisfied with the responses to the CI problem just discussed, to abandon RTM in favor of STM. It is a question of some interest then whether the collateral information problem arises for STM. That is the question I shall be pursuing in the rest of this essay. My claim will be that the CI problem does arise for STM and, therefore, that one should be wary of opting for STM as a means of escaping it.

2. The Story of Mrs. T and its Alleged Implications for RTM

In this section, I want to discuss Stich's version of the collateral information problem, and the difficulties he believes it creates for RTM. This will set the stage for section 3 in which I will describe how Stich thinks that STM can avoid such difficulties.

Stich's version of the CI problem is presented in a series of examples, one of which is introduced in the following passage:

> Let me ... describe in a somewhat idealized way the history of Mrs. T, a real person who was employed by my family when I was a child. As a young woman, around the turn of the century, Mrs. T had an active interest in politics and was well informed on the topic. She was deeply shocked by the assassination of President William McKinley in 1901. In her sixties, when I first knew her, she would often recount the history of the assassination and spell out her analysis of the effects it had had on the politics of the day. As Mrs. T advanced into her seventies, those around her began to notice that, though her reasoning seemed as sharp as ever, her memory was fading. At first she had trouble remembering recent events: who had been

elected in the Senate race she had been following; where she had left her knitting. As time went on, more and more of her memory was lost.... Some while before her death, something like the following dialogue took place:

S: Mrs. T, tell me, what happened to McKinley?
Mrs. T: Oh, McKinley was assassinated.
S: Where is he now?
Mrs. T: I don't know.
S: I mean, is he alive or dead?
Mrs. T: Who?
S: McKinley.
Mrs. T: You know, I just don't remember.
S: What is an assassination?
Mrs. T: I don't know.
S: Does an assassinated person die?
Mrs. T: I used to know all that, but I just don't remember now.
...
S: But you do remember what happened to McKinley?
Mrs. T: Oh, yes. He was assassinated.

(Stich, 1983, p. 55.)

In discussing this example, Stich first notes that it seems clearly wrong to say that Mrs. T, at the time of the conversation, believed that McKinley was assassinated. To believe that someone was assassinated, one must have, it seems, at least a rudimentary understanding of assassination, and this Mrs. T lacks. But, Stich argues, if she doesn't believe that McKinley was assassinated, it would appear that there isn't anything that she believes. For no sentence p does it seem right to characterize the state that leads Mrs. T to say, "McKinley was assassinated," as the belief that p.[10] Furthermore, Stich argues, even if we were intimately familiar with the progress of Mrs. T's loss of memory, it is doubtful that we would be able

---

10. It is important to note that it is the folk psychological notion of belief that is at issue in Stich's discussion of Mrs. T. If RTM is willing to embrace a theory of content that doesn't always square with the judgments of folk psychology, then it has more room to maneuver here. Although Stich seems to insist that RTM accept the unrefined folk notion of belief, I don't see why it should be bound to do so. I don't see, for example, why in principle RTM couldn't adopt a graded notion of content even if such a notion would conflict with the folk notion of belief.

to say exactly when Mrs. T stopped having the belief that McKinley was assassinated. Stich adds some details to the story, and asks if a certain point is the point at which Mrs. T stopped having the belief that McKinley was assassinated. "The answer," he writes "surely, is that there is no answer. We have entered the penumbra of vagueness where, apart from a special conversational context, there is just no saying whether the content sentence is applicable." (Stich, 1983, p. 142.)

It is the issue of the vagueness of when she ceased believing that McKinley was assassinated that makes the example of Mrs. T a version of the CI problem. The story of Mrs. T is, in a sense, just the story of Elmer and Oscar in reverse. Whereas Elmer and Oscar's concepts of grug underwent a change in content as they acquired information about grug a little bit at a time, Mrs. T's concept of assassination undergoes a change in content (in particular, a loss of content) as she loses information about assassination a little bit at a time. The intuition in both cases is that it would be impossible to justify any particular choice for the point in time when the change in content took place.

There is an important difference in the two cases in that the beliefs acquired by Elmer and Oscar all seem to be collateral to the concept of grug. In Mrs. T's case, on the other hand, some of the beliefs that she loses--the belief that people who are assassinated die, for example--seem to be essential to the concept of assassination. Stich's example, then, relies more heavily on the claim that it will be impossible to determine precisely which beliefs are essential in this way and which are not. In other words, Stich's example must rely on something like the traditional arguments against the analytic/synthetic distinction. Putnam's case is somewhat independent of such arguments: since all of the beliefs at issue

would seem to fall on one side of such a distinction, the ability to make such a distinction wouldn't necessarily help. Again, however, in both examples a change in the content of a concept is caused by a gradual accumulation or loss of beliefs related to the concept. And, it is argued, it is impossible to justify any particular choice for the point in time when the change in content took place.

Stich argues that the example of Mrs. T creates two difficulties for RTM. Both difficulties stem from the fact that the generalizations of an RTM-based theory typically relate mental state types in virtue of their content. Since an RTM-based theory won't be able to assign a content to Mrs. T's cognitive state, no generalizations of such a theory will apply to this state.[11] But, Stich notes, the state in question may still enter into what appear to be law-governed interactions with other of her mental states. For example, Stich suggests,

> ... it may turn out that, even as she is on the brink of death, if we tell Mrs. T, "If McKinley was assassinated, then he is buried in Ohio," she would reply "Well, then, he is buried in Ohio." It would be tempting to conclude from this, and a pattern of similar responses, that while her long-term memory was largely destroyed, her inferential abilities remain intact--the generalizations governing inference which applied before her illness continue to apply. (Stich, p. 142.)

So, Stich concludes, some of the generalizations governing inference ought to continue to Mrs. T's cognitive state. But, the generalizations of an

---

11. Stich unjustifiably ignores the possibility of RTM-based theories containing generalizations which apply to cognitive states in virtue of something other than their content. RTM is free to characterize cognitive states in any number of ways--having a particular syntactic form, being generated by a particular cognitive system, etc.--other than in terms of its content. Each such way of characterizing cognitive states brings with it the possibility of new generalizations. All RTM is committed to is that propositional attitude states, and, hence, characterizations in terms of content, play a central role in the explanation of behavior.

RTM-based theory, which, according to Stich, must apply to cognitive states in virtue of their content, cannot.

The second problem that the example of Mrs. T is alleged to create for RTM has to do with the vagueness and sensitivity to conversational context of the question of when Mrs. T ceased believing that McKinley was assassinated. The generalizations of an RTM-based theory only apply to Mrs. T's cognitive state so long as it can be assigned a content. Since it is a vague and context-sensitive matter when her psychological state ceased having a content, it is a vague and context-sensitive matter precisely when the generalizations of an RTM-based theory applied to the state. The application of the generalizations of a scientifically respectable theory, Stich claims, should not be a vague or context-sensitive matter.

In sum, Stich argues that the example of Mrs. T (and other similar examples) show (a) that the generalizations of RTM-based theories do not apply as widely as they should and (b) that the question of the extent of their application is plagued with vagueness.

3. STM and Avoiding the Collateral Information Problem

In this section, I will present the Syntactic Theory of Mind that Stich advances in The Case Against Belief, paying particular attention to the manner in which it is supposed to avoid the difficulties discussed in section 2. It will be seen that one principle in particular plays a crucial role in allowing STM to escape these difficulties. In section 4 I will argue that although this principle overcomes the collateral information problem, it creates other, insuperable difficulties for STM.

Stich gives the following thumbnail description of STM:

> The basic idea of STM is that the cognitive states whose interaction is (in part) responsible for behavior can be systematically mapped to abstract syntactic objects in such a way that causal interactions among cognitive states, as well as causal links with stimuli and behavioral events, can be described in terms of the syntactic properties and relations of the abstract objects to which the cognitive states are mapped. More briefly, the idea is that causal relations among cognitive states mirror formal relations among syntactic objects. (Stich, p. 149.)

So, before beginning to construct generalizations with which to explain the behavior of her subjects, someone building an STM-based theory must specify a formal language containing the relevant syntactic objects and delineate principles according to which cognitive states can be assigned to these syntactic objects. Stich does not say much about the formal language in which the syntactic objects of STM-style theories ought to be built. But, in all of his examples he assumes that the formal language will be of the sort standardly used to express first-order quantification theory. Adopting the terminology used in talking about formal languages, Stich refers to the syntactic objects of STM-based theories as wffs.

In answer to the question of how cognitive states (specified in some neutral manner--neurologically, for example) are assigned to wffs, Stich writes:

> If we wish to view tokens of the states postulated by an STM-style theory as sentence tokens, then we must say something more about the individuation or typing of these sentence tokens. Since the motivation for viewing hypothetical neurological state tokens as sentence tokens is to describe causal relations by adverting to syntactic ones, we must ask just which syntactic relations must be mirrored for the neurological state tokens to count as sentence tokens. There are, I think, three rather different answers that might be given. One idea is to insist that if a neurological state token is to count as a token of a sentence it must satisfy all the generalizations specified by the theory.... A second idea ... is to specify a set of essential generalizations which a neurological state must satisfy if its tokens are to count as tokens of a given sentence type. Further generalizations may be added and modified as necessary without altering the account of typing. But this approach, too, has its shortcomings. It

> is hard to see what motivation there can be for distinguishing a special set of generalizations as the essential ones, hard to see how the divide between essential and non-essential generalizations could be anything but arbitrary. A third idea is to evade the issue by insisting only that to count as a token of a sentence type, a neurological state must satisfy some substantial number of the cluster of generalizations included in a theory, without specifying any particular generalizations that must be satisfied, nor exactly how many must be satisfied. (Stich, p. 152.)

Stich does not choose among the three methods for determining when a given neurological state is to be mapped to a given syntactic object, though he seems to tilt toward the third possibility. For our purposes, however, the importance of this passage lies in the common assumption underlying each of the three possibilities: neurological states are to be individuated cognitively according to their causal role. More specifically, since it is assumed that an STM-based theory will be able to state generalizations which collectively specify the causal role of every cognitive state, for a neurological state to be of a particular cognitive type is for it to satisfy the generalizations that apply, according to the theory, to cognitive states of that type.[12] This view is clearly at work in the quoted passage; the only question is whether the neurological state has to

---

12. Stich is well aware that the process of determining which generalizations of the theory apply to which states is a holistic process. For example, suppose that an STM-based theory has a generalization that says that if a subject is in a cognitive state mapped to the syntactic object A*B and is in a cognitive state mapped to the syntactic object A, then the subject will not be in a cognitive state mapped to the syntactic object B. Whether a given neurological state token, N let's say, is to be mapped to the syntactic object A*B may depend partially on whether this generalization applies to N. But, of course the decision as to whether it does depends on what syntactic objects other neurological states--the ones to which N is related in the appropriate way--are mapped to. And, this in turn will depend partially on whether the generalization in question applies to them, which will depend in turn on what syntactic object N is mapped to. The STM-style theorist trying to map various neurological states to syntactic objects will evidently have to coordinate these decisions in the holistic manner familiar from the case of radical translation.

satisfy all the relevant generalizations. Regardless of how that question is answered, the commitment to typing by causal role evidently dictates that all that matters in determining whether a given neurological state is of a given cognitive type is what generalizations it satisfies. Hence, we have the following principle:

> (G) The only facts relevant to determining what type a mental sentence (specified as a token of a neurological state type) is a token of, are facts about which generalizations of the theory the neurological state satisfies.

It should be added that typing cognitive states, according to STM, involves not only assigning them to syntactic objects but assigning them to certain broad classifications. The idea is that to capture the distinction between, for example, (using folk terminology) the belief that John is ill and the fear that John is ill, an STM-based theory should classify the former as a B-state and the latter as an F-state, but map them to the same syntactic object. Which of these classifications applies to a given cognitive state will again be determined by causal role.

We can now recast Stich's brief description of STM as follows: generalizations detailing the causal interactions among cognitive states, of which there are various kinds, are to be stated in terms of relations between the syntactic objects they are mapped to; the syntactic objects are wffs in a formal language akin to ones in which first order logic with quantification is usually expressed; and, which wff a cognitive state (neurologically specified) is mapped to, as well as which kind of cognitive state it is, is determined by which of the generalizations it satisfies.

In a moment we will supplement this description with an additional principle. But, it is important to note that as it stands there is reason

to believe tnat STM will be vulnerable to the collateral information problem. Consider, for example, the cognitive state which might lead Mrs. T to say, "The President of the United States has just been assassinated." More specifically, consider C1, the cognitive state that would have led Mrs. T at age 40 to say, "The President of the United States has just been assassinated," and C2, tne cognitive state that might lead the senile Mrs. T to say, "The President of tne United States has just been assassinated." For tne young Mrs. T, coming to be in C1 would have had many consequences. In non-STM terms, she would have come to expect the Vice-President to assume the presidency. She would have expected the next day's newspaper to be dominated by a discussion of the event and its impact. She would have expected many people to be shocked and saddened (and might have been shocked and saddened herself). For the elderly Mrs. T, coming to be in C2 would, presumably, nave none of these consequences.

Each of the predictions above about what the young Mrs. T would have come to believe and expect is generated by applying folk psychology to what we know about Mrs. T. In other words, folk theory contains lots of generalizations about what reasonably knowledgeable, socialized people living in the United States will think if they learn that the President of the country has just been assassinated. And, we might expect an STM-theory to have analogous generalizations (especially if, as Stich advertises, it captures all the generalizations RTM can capture). All such generalizations apply to C1--that's how I was able to predict what Mrs. T would have come to think had she been in C1. But, of course, none of them apply to C2. Hence, C1 satisfies a very different set of generalizations than C2. So, since STM-based theories type cognitive states on the basis of the generalizations of the theory that they satisfy, it seems that C1 is

of a different type than C2. And, of course, this change is a consequence of Mrs. T losing lots of B-states. We can now raise the question, Just when did the (non-actual) state C1 come to be of a different type? Having to answer such questions would, of course, create a collateral information problem for an STM-based theory.

It turns out that STM is able to avoid this difficult question. But, in order to do so, it requires an additional principle. It is a principle Stich first mentions in a general discussion of individuation of belief-like states by causal role. Stich writes:

> The causal patterns of interest to narrow causal accounts of typing are not merely those that have obtained among actual states, but also those that would obtain among nonactual though possible states. The essential point is that, for a narrow causal theorist, the type identity of a mental state is determined by its potential causal interactions with other mental states, with stimuli, and with behavior. Its type identity does not depend on the other mental states the subject happens to be in at the moment in question. (Stich, p. 54.)

In his discussion of STM, Stich makes it clear that STM-based theories must adhere to this principle, that the type identity of a cognitive state does not depend on the other cognitive states the subject happens to be in at the moment in question:

> For a syntactic theory, however, ideological similarity poses no problem, since the characterization of a B-state does not depend on the other B-states that the subject happens to have. A B-state will count as a token of a wff if its potential causal links fit the pattern detailed in the theorist's generalizations, regardless of the further B-states the subject happens to have. (Stich, p. 158.)

Putting this principle--that what wff a B-state is a token of is independent of what other B-states a subject also happens to be in--together with the fact that what wff a B-state is a token of depends on what generalizations apply to the B-state (token), we get a principle about

generalizations of STM-based theories. Specifically, whether a generalization applies to a given B-state token must itself be independent of which other B-states the subject happens to be in. If not, what other B-states a subject happens to be in could affect which generalizations apply to a given B-state token and then affect in turn the type identity of that B-state token, which, of course, the principle in question does not permit. Hence, STM is committed to the following:

> (I) If a generalization of an STM-based theory applies to a B-state token of a given subject, then it applies independently of which other B-states the subject happens to be in (or not to be in).

Principle (I), and the principle from which it was derived, are used by Stich to argue that STM avoids the two problems with RTM cited earlier. Let's see how this works in Mrs. T's case. The idea is this: for an STM-based theory, Mrs. T's B-states will be individuated not by assigning them content sentences but by mapping them to wffs. What wff a given B-state is mapped to will be independent (by the principle from which (I) was derived) of what other B-states she happens to be in. So, no matter how much of her memory she loses--in other words, no matter how many B-staes she ceases to be in--the remaining B-states will still be mapped to the same wffs. If the B-state which leads Mrs. T to say that Mckinley was assassinated was mapped to the wff 'Am' before she began to lose her memory, it will be mapped to the same wff at the time of the conversation quoted at the beginning of section 2; the identity of her B-state does not change as her memory fades.

The vagueness and context sensitivity innerent in an RTM-based theory's judging whether Mrs. T believes that McKinley was assassinated at a given point in time apparently has no analog for an STM-based theory.

Furthermore, by (I), the generalizations which applied to Mrs. T's B-states before apply to the ones that remain. So, the syntactic version of the generalization that if someone believes that p, and comes to believe that if p, then q, then the person will come to believe that q, still applies to Mrs. T. Hence, the STM-based theory can explain why, when we say to the senile Mrs. T, "If McKinley was assassinated, then he is buried in Ohio," she responds, "Well then, he is buried in Ohio." Thus, an STM-based theory avoids the difficulty that confronted RTM-based theories here: RTM theories could not apply the folk version of the generalization because there wasn't anything that Mrs. T could be said to believe.

The idea behind principle (I) is that the type identity of a cognitive state should be determined not by its actual causal connections with other mental states, stimuli and behavior, but rather by its potential causal connections. Clearly, this insistence on the importance of potential as opposed to actual causal connections avoids the collateral information problem. I shall argue in section 4, however, that identifying cognitive states by their potential rather than actual causal connections has disastrous consequences. Roughly put, the idea will be that any two beliefs of the same logical form will have the same potential causal connections even if they have very different actual causal connections. I will then conclude that STM must give up principle (I) and face the collateral information problem.

4. Distinguishing B-States with the Same Logical Form

In this section I will argue that there is no way an STM-based theory can distinguish between the B-state that leads Mrs. T to say, "McKinley was assassinated" and the B-state that might lead her to say, "McKinley was

elected." I'll begin by snowing tnat an STM-based theory ought to be able to distinguish these two states, that simply accepting that these two states are one and the same is not a live option for such a theory. This will set the stage for tne argument that the two states will be indistinguishable for an STM-based theory.

4.1 Two Conversations with Mrs. T

Recall Stich's supposition that "even as she is on the brink of death, if we tell Mrs. T 'If McKinley was assassinated, then he is buried in Onio,' she would reply 'Well tnen, he is buried in Ohio.'" A generalization of the following sort is supposed by Stich to explain wny this is so:

> (1) For all subjects S, and all wffs A and B, if S has a B-state mapped to A --> B and if S comes to have a B-state mapped to A, tnen S will come to have a B-state mapped to B. (Stich, p. 155)

Along witn a few modest assumptions, tnis generalization will, in fact, engender an adequate explanation of Mrs. T's behavior.

Here are the required assumptions. First, we must assume that some particular B-state leads Mrs. T to say, "McKinley was assassinated." Let's call that B-state B1.[13] Let W1 be the wff that B1 is mapped to. Next, we have to assume that there is a particular B-state that leads Mrs. T to say, "Well then, he is buried in Ohio." Let's call that B-state B2 and tne wff

---

13. I'll use expressions such as 'B1' to refer to neurologically characterized states of Mrs. T. In virtue of their causal connections to other of Mrs. T's neurological states, to stimuli, and to her behavior, these neurological states also count as tokens of B-states in our STM-based psychological theory, and, hence, get mapped to wffs in the formal language in which the generalizations of our theory are cast. I'll use expressions such as 'W1' to refer to such wffs.

it is mapped to W2. We must now assume tnat telling Mrs. T, "If McKinley was assassinated, tnen he is buried in Onio," leads to a particular B-state. Let's call it B3. We will have to assume that B3 gets mapped to the wff W1 --> W2.[14]

Verbal stimulus/behavior	B-state stimulus/behavior is connected to	Wff B-state is mapped to
"McKinley was assassinated"	B1	W1
"Well then, he is buried in Onio"	B2	W2
"If McKinley was assassinated, then ne is buried in Onio"	B3	W1 --> W2

Given these assumptions, we can explain that prior to our conversation, Mrs. T was in B-state B1; that as a result of our saying to her "If McKinley was assassinated, tnen he is buried in Onio," she came to be in B-state B3; that she thereby came to satisfy the antecedent of generalization (1); that as predicted by that generalization, she came to be in a B-state mapped to tne wff W2, i.e. B2; and, that she said "Well then, he is buried in Onio," as a result of being in B-state B2.

Imagine now, however, that we had had a different conversation witn Mrs. T. Imagine that instead of saying to Mrs. T, "If McKinley was assassinated, then he is buried in Ohio," we tell her, "If McKinley was elected, then he is buried in Ohio." And, suppose, as is perfectly consistent with what has been said so far about Mrs. T, that this elicits no response. Suppose, in other words, that in whatever sense Mrs. T

---

14. Of course, when I write "the wff W1 --> W2," I mean the wff formed by concatenating the wff W1 with the symbol '-->' and concatenating the result with the wff W2.

remembers that McKinley was assassinated, she does not remember that McKinley was elected.

If we are to continue to maintain that generalization (1) explains Mrs. T's saying, "Well then, he is buried in Ohio," in the case of the first conversation, we will have to claim that generalization (1) does not apply in the case of the second conversation. In particular, we will have to insist that saying, "If McKinley was elected, then he is buried in Ohio," does not lead Mrs. T to have a B-state that gets mapped to the wff W1 --> W2. For if it did, generalization (1) would apply just as it did before, and would predict that she will come to be in B-state B2 and say as before, "Well tnen, he is buried in Ohio." And, of course, she doesn't.

If saying to Mrs. T, "If McKinley was elected, then he is buried in Onio," doesn't lead Mrs. T to nave B-state B3--the one mapped to tne wff W1 --> W2--then what B-state does it lead her to have? It's certainly conceivable that it doesn't lead her to have any B-state at all. But, it would be odd if the utterance, "If McKinley was assassinated, then he is buried in Onio,"--the second half of which is, we may presume, completely unfamiliar to Mrs. T--causes a B-state in her, but the utterance in which "assassinated" is replaced with "elected" does not. In any case, we may certainly suppose that the latter state does lead to some B-state in Mrs. T. Let's call that B-state B4. What wff will B4 be mapped to? The association between uttering, "If McKinley was assassinated, then he is buried in Ohio," and the wff W1 --> W2 leads to an obvious suggestion: B4 should be mapped to the wff W0 --> W2, where W0 is the wff that B-state B0 is mapped to, and B0 is the B-state Mrs. T comes to have if she is simply told, "McKinley was elected." Let's suppose that this is so. We have so far posited the following B-states and wffs:

Verbal stimulus/behavior	B-state stimulus/behavior is connected to	Wff B-state is mapped to
"McKinley was assassinated"	B1	W1
"Well then, he is buried in Ohio"	B2	W2
"If McKinley was assassinated, then he is buried in Ohio"	B3	W1 --> W2
"If McKinley was elected," then he is buried in Ohio"	B4	W0 --> W2
"McKinley was elected"	B0	W0

As I noted before, if we are to explain, using generalization (1), why Mrs. T responds as she does in the two conversations, we will have to insist that when we tell her, "If McKinley was elected, then he is buried in Ohio," she does not come to have a B-state mapped to the wff W1 --> W2. If she did, we would expect her to come to have B-state B2 and to respond, "Well then, he is buried in Ohio." And, she doesn't. Given the suppositions we have made, summarized in the table above, this means that we must insist that the wffs W1 and W0 be distinct. Since, according to STM, assigning a wff to a B-state is just another way of saying what type the B-state falls under, this means that B-states B1 and B0 must be assigned to distinct types.

The upshot of our attempt to explain Mrs. T's responses in the two conversations is that in order to explain why she says, "Well then, he is buried in Ohio," in the one conversation and merely shrugs her shoulders in the other, an STM-based theory must be able to distinguish the B-state which leads Mrs. T to say from time to time, "McKinley was assassinated,"

(B1) from the B-state she comes to have if we tell her, "McKinley was elected," (B0) by assigning them to distinct types (and thereby associating them with distinct wffs). I shall argue in the rest of this section that given principles (G) and (I) discussed in section 3, this will not be possible.

As I indicated in section 3, STM-based theories individuate B-states according to their causal role. And, according to STM, the causal role of a B-state is captured by the generalizations of the theory that it satisfies. This commits STM-based theories to the following principle, also discussed in section 3:

> (G) The only facts relevant to determining what type a mental sentence (specified as a token of a neurological state type) is a token of, are facts about which generalizations of the theory the neurological state satisfies.

From (G) it follows that, if an STM-based theory is to assign B-states B1 and B0 to distinct types, there must be at least one generalization of the theory which is satisfied by one of the B-states but not the other. Stich countenances three sorts of generalizations in STM-based theories: generalizations specifying connections between a cognitive state and other cognitive states; generalizations specifying connections between cognitive states and stimuli; and, generalizations specifying connections between cognitive states and behavior. Among generalizations specifying connections between cognitive states and other cognitive states, I will distinguish between those which hold of all cognitive states of a given form--all B-states mapped to wffs of the form A --> B, where A and B may be any wffs, for example--and those that do not. This leaves four classes of

generalizations. I will treat each class in turn and argue that no generalization of tnat class could apply to B1 or B0 but not the other.

Before I turn to tne four classes of generalizations, I want to make a simplifying assumption. The assumption is that the wffs B1 and B0 are mapped to--W1 and W0--are what I'll call simple wffs. By a simple wff I mean a wff consisting of a predicate later followed by an individual constant (Recall that we are assuming that the formal language in which the wffs of the theory are to be constructed is of the sort standards used to express first-order quantification theory). Eventually I will drop this simplifying assumption. For the moment, however, let W1 be the wff 'Am' and let W0 be tne wff 'Ew,' understanding that 'A' and 'E' may turn out to be the same predicate letter and 'm' and 'w' may turn out to be the same individual constant. I turn now to the four classes of generalizations, which might justify assigning B0 and B1 to distinct types.

4.2 Generalizations Based on Schematic Connections Between B1 (or B0) and other Cognitive States

I use the phrase "schematic connections" to refer to those connections which follow from generalization schemas in which no specific predicate letters or individual constants appear and into which any wff can be substituted to produce a valid generalization. In other words, a schematic connection between cognitive states is one based solely on the way the respective wffs are built up out of connectives and quantifiers. Included in this class of generalizations would be instances of schemas that express: (a) rules of deductive logic, such as, from A and A --> B, infer A, or, from A and B, infer A & B; (b) the practical syllogism--if you believe that A --> B and you desire that B, then you will probably desire that A; (c) the phenomenon of "belief perseverance"--if you come to believe

that A and thereby come to believe that B, and then, later come to believe not-A, you will still be more likely to believe that B than you were before[15]; and, many others.

The following, an instance of generalization schema (1), discussed earlier in this section, is an example of a generalization based on a schematic connection:

(2) For all subjects S, if S is in a B-state mapped to the wff 'Am' and S comes to have a B-state mapped to the wff 'Am --> Bm,' then S will come to have a B-state mapped to the wff 'Bm.'

It might be thought that this generalization will distinguish between B1 and B0, since it seems to apply to B1 but not to B0. But, notice that a very similar generalization does apply to B0, namely:

(2*) For all subjects S, if S is in a B-state mapped to tne wff 'Ew' and S comes to have a B-state mapped to the wff 'Ew --> Bm,' then S will come to have a B-state mapped to the wff 'Bm.'

If the existence of (2) is to be made to distinguish between B0 and B1, then we will have to be willing to claim that (2) and (2*) are not merely the same generalization in different guises. But, without begging the question at hand--whether 'Am' and 'Ew' are really the same wff--that claim cannot be justified.

The same problem will arise for any generalization which expresses, as I have called it, a schematic connection between the cognitive states it applies to. For, given any such generalization which applies to, e.g., B0, the generalization schema from which it was generated can be applied to

---

15. Stich cites this phenomenon, and uses it to construct a generalization schema that might be part of an STM-based theory. See Ross, 1977 on the subject of belief perseverance.

produce a parallel generalization which cannot be distinguished from the first one without begging the question at hand. So, generalizations based on schematic connections between B0 or B1 and other cognitive states will not distinguish the two states.

4.3 Generalizations Based on Non-Schematic Connections Between B1 (or B0) and other Cognitive States

There are certainly plenty of generalizations based on schematic connections between B1 (or B0) and other cognitive states. But, as we have seen, none of them will be able to distinguish between B1 and B0. What about generalizations based on other connections between cognitive states? What we need is a connection between B1 (or B0) and some other cognitive state where the connection is not based solely upon the way the wffs the states are mapped to are built up out of connectives and quantifiers. Here is an example of such a generalization:

> (3) For all subjects S, if, for some individual constant 't,' S has or comes to have a B-state mapped to the wff 'At' (corresponding to the belief that the individual named by 't' was assassinated), then S has or will come to have a B-state mapped to the wff 'Dt' (corresponding to the belief that the individual named by 't' is dead).

The idea behind this generalization is that if someone believes or comes to believe that a particular person was assassinated, then they believe or will come to believe that the person is dead. And, indeed the two beliefs are normally associated in this way. An STM-based theory, however, is not likely to find room for such a generalization. To see why not, we must recall the second principle discussed in section 3:

(I) If a generalization of an STM-based theory applies to a B-state token of a given subject, then it applies independently of what other B-states the subject happens to be in (or not to be in).

Consider now the young Mrs. T (young enough that her memory is intact, but old enough that McKinley has already been assassinated). If (3) is part of some STM-based theory, then it ought to apply to the young Mrs. T. After all, she is, we may presume, normal in every way. And, in fact, the young Mrs. T presumably has a B-state mapped to the wff 'Am' (B1), and also has a B-state mapped to the wff 'Dm'; the generalization apparently applies to B-state B1 of the young Mrs. T. However, it does not apply to this B-state independently of what other B-states she happens to be in. We know this because, by hypothesis, she goes on to lose many of her other B-states, and solely in virtue of the loss of these B-states, the generalization comes no longer to apply to B-state B1 of Mrs. T. The elderly Mrs. T has a B-state, B1, corresponding to the belief that McKinley was assassinated, but none corresponding to the belief that McKinley is dead. By principle (I), then, generalization (3) doesn't really hold of the young Mrs. T. And, since the young Mrs. T is perfectly normal, there is no reason to think that it holds of anybody.

Principle (I) places a substantial constraint on any generalization based on connections between two cognitive states. The connection must be strong enough that it will continue to hold regardless of what other cognitive states a subject happens to be in or happens not to be in. Since the young Mrs. T is perfectly normal, and since the elderly Mrs. T differs from her younger self only in having undergone a loss of memory, the connection must be strong enough to survive in the elderly Mrs. T. And, of course, the whole point of Mrs. T is that the connections between cognitive

states that normally nold, don't hold in Mrs. T. In particular, given any alleged connection between B1 (or B0) and some other B-state, if the connection is not of the form of a logical inference, and, hence, a schematic connection, it will be easy to suppose that the connection does not survive Mrs. T's senility. Hence, any such connection between B1 (or B0) and another cognitive state will not be strong enough to form the basis of a generalization of an STM-based theory. Evidently then, the category of generalizations based on non-schematic connections between cognitive states will be of no use in distinguishing B1 from B0.

4.4 Generalizations Based on Connections Between B1 (or B0) and Stimuli

It is quite natural to expect that connections with stimuli will allow us, finally, to distinguish between B1 and B0. After all, we identified B-state B0, as well as B-states B3 and B4, according to the verbal stimuli which cause Mrs. T to be in these states. B1 is, by definition, connected with Mrs. T's hearing the utterance, "McKinley was assassinated." Presumably, there is no such connection between B0 and Mrs. T's hearing the utterance, "McKinley was assassinated." In a moment, I want to consider the question of whether we can formulate a generalization of our STM-style theory based on the connection between B1 and the utterance, "McKinley was assassinated," that will apply to B1 and not to B0, and thereby distinguish the two states. But, first I want to argue that if <u>any</u> connection between B1 (or B0) and stimuli will yield such a generalization, it will be a connection involving a linguistic stimulus.

The connections of interest between cognitive states and stimuli obviously run in one direction only; stimuli produce cognitive states, and not, in general, the other way around. So, what we need is some kind of

stimulus which will generally produce in people cognitive states of the same type as B0 (or B1). Put in folk psychological terms, we need to think of a stimulus such that, if a person is exposed to it, and the person does not already believe that McKinley was elected, the stimulus will produce that belief in them. We must keep in mind, however, that the connection has to be strong enough that the resulting generalization will satisfy principle (I). In particular, the connection will have to be strong enough that it survives even in Mrs. T. Now, what sort of non-linguistic stimulus can we expose someone to in order to get them to believe that McKinley was elected. Perhaps we could show them a picture of *a* campaign celebration in which McKinley is giving an acceptance speech. If this does allow the person to infer tnat McKinley was elected, it is only in virtue of that person's knowing a substantial number of facts about election campaigns, as well as having the ability to recognize McKinley (which is, of course, unlikely). But, of course, Mrs. T doesn't have any of the relevant facts about elections and about McKinley at her disposal. There isn't the remotest possibility that we can produce B-state B0 in Mrs. T just by showing her a picture of McKinley giving a campaign speech (unless, of course, it triggers a recovery of some of her lost memory). And, no other non-linguistic stimulus is any more likely to do the trick. Exposure to non-linguistic stimuli will generally produce in people the belief that McKinley was elected by way of complicated inferences, some of the premises of which Mrs. T is not in possession of.

Evidently, the only kind of stimuli whose connections with the B-state B0 might be direct enough to survive in Mrs. T are linguistic stimuli. The connections involving linguistic stimuli that seem most likely to allow us to distinguish between B0 and B1 are the ones between B1 and tokens of the

English sentence, "McKinley was assassinated" and between B0 and tokens of the sentence, "McKinley was elected." I want now to consider the question of whether such connections will finally provide a generalization that distinguishes between the two B-states in question. Because these connections represent the most plausible source for such a generalization, I will proceed very carefully here, ultimately considering three different ways that a distinguishing generalization might be framed. Once, however, we have set each of these possibilities aside the argument that an STM-based theory has no basis for assigning B1 and B0 to different types will be nearly complete.

Let's begin by attempting to construct a generalization based on the connection between tokens of the sentence, "McKinley was assassinated," and B-state B1. Here is such a generalization:

> (4) For all subjects S, and all names 'N,' if S is a speaker of English with normal hearing abilities and a token of the sentence, "N was assassinated," is uttered in S's vicinity, then S will come to have a B-state mapped to the wff 'An,' where 'n' is some individual constant.

As it stands, this generalization is false. Generally, exposure to utterances of p do not lead people to believe that p unless the utterer is thought to be a reliable source of the relevant sort of information, and the utterance of p is taken as an attempt to express the speaker's belief that p. And, as Paul Grice has shown, it is quite a complicated matter to say when utterances of p will (or ought) to be taken as attempts to express the speaker's belief that p.

Grice argued more particularly that we, as communicators, share an adherence to certain conversational maxims, and that the appearance of violating these maxims can be used by a speaker so that a particular

utterance conveys to her conversational partner any of a number of the speaker's beliefs. In STM jargon the Gricean claim is this: suppose, as is suggested by generalizations such as (4), that there is a correspondence between English sentences and B-states, so that for each sentence of English, there is a B-state corresponding to it that utterances of the sentence "standards" lead to. Then, the claim is that by seeming to break a conversational maxim, a speaker could use an utterance of p to lead her conversational partner to have any of a number of B-states other than the one that corresponds to p. In the case at hand, S may have a B-state corresponding to the belief that the person uttering, "McKinley was assassinated," would be breaking a conversational maxim unless she meant something other than that McKinley was assassinated. S would then come to have a B-state other than the one mapped to 'Am.' In sum, for the utterance actually to lead to a B-state mapped to the wff 'Am,' it must be the case (now not speaking in STM jargon) (a) that S believes the speaker to be a generally reliable communicator; and (b) that S does not believe that the speaker would be breaking a conversational maxim, if she were attempting to express her belief that McKinley was assassinated. This is to say, of course, that (4) violates (I); the generalization only holds when S has further B-states and does not have others.

It might be thought that this problem can be avoided by building the necessary conditions into the antecedent of the generalization. Thus, we might get something like this:

(5) For all subjects S, and all names 'N,' if S is a speaker of English with normal hearing abilities, and a token of the English sentence, "N was assassinated," is uttered in S's vicinity and S has a B-state mapped to the wff R (corresponding to the belief that the speaker is a reliable communicator), and S does not have a B-state mapped to the wff B (corresponding to the belief that the speaker would be breaking a conversational maxim if she were to attempt to express the belief that N was assassinated with the utterance, "N was assassinated"), then S will come to have a B-state mapped to the wff 'An,' where n is some individual constant.

But, this generalization too is inadequate. Even if S does not now have a B-state mapped to the wff B, that is, does not now believe that the speaker would be breaking a conversational maxim if she were to attempt to express the belief that N was assassinated, she may come to have this B-state after having heard the utterance. What is needed to make the generalization adequate is a set of B-states such that if the subject has these B-states, the subject will not come to have a B-state mapped to the wff B. But, such a set is surely not in the offing. There are just too many ways that S might come to believe that the speaker's meaning what she uttered would break a conversational maxim. This being the case, there doesn't appear to be any way of constructing a generalization in our STM-based theory based on the connection between utterances of, "McKinley was assassinated," and the B-state B1. And, since this connection offered the best hope for a generalization based on a connection between stimuli and B1 (or B0), we may conclude that such a generalization will not be available.

It is important to note that the arguments used above to reject (4) and (5) as plausible STM generalizations need not rely at all on the details of Grice's theory of communication. I used the Gricean theory in two ways. First, I used it to make the point that hearing utterances of p only lead a person to believe that p under certain very particular

conditions. This point, which Grice stressed, should be uncontroversial. Given this point, it is clear that (4) is in conflict with principle (I). We must modify (4) to build more stringent conditions into the antecedent of the conditional. I then used the Gricean story about conversational maxims as a way of considering what would have to be built into this antecedent. If some other story about the conditions under which utterances of p lead people to form the belief that p is correct, then some other attempt to modify the antecedent to the conditional in (4) would be appropriate. But, it still seems doubtful that we could build enough into the antecedent to the conditional to guarantee that an utterance of p will lead the subject to form the belief that p. Regardless of what the correct story about the path from hearing utterances to forming beliefs (or belief-like states) is, there are, on any plausible view of communication, too many ways that the path can be blocked.

Though it is true that the arguments rejecting (4) and (5) do not rely on details of Grice's theory, there is one technical point that it does rely on. Let's call those cases where a speaker's utterance of p does not ultimately result in the listener's forming and maintaining the belief that p, non-standard cases. The assumption I have made is that in the non-standard cases the listener never forms, even on a tentative basis, the belief that p. One could accept the Gricean point, but maintain that in non-standard cases the listener first forms the tentative belief that p, then discovers that the belief is not warranted and therefore rejects it. Let us call this proposal the direct translation proposal, since it supposes that the listener directly translates the utterance that p into a tentative belief that p and only then asks questions about whether the

speaker is really reliable, whether she could have really meant that p, etc.

An advocate of STM who wanted to distinguish between B-states B1 and B0 could use the direct translation proposal to resuscitate generalization (4). For if the proposal is correct, then it is true that whenever a speaker of English hears an utterance of the sentence p, she will come to be in the associated B-state, even if she ceases to be in that B-state only moments later. Generalization (4) would apply to B-state B1 but not B0 and thus form the basis for assigning the B-states to distinct types.

The only problem with the direct translation proposal is that it constitutes a rather bold empirical supposition. If one accepts Fodor's distinction between input modules and central processes (Fodor, 1983), then the empirical question being addressed is the question of the form of the output of the language module. On Fodor's picture central processes bring to bear everything the subject knows in attempt to determine what the subject ought to believe on the basis of what the input module has reported. Hence, unless the output of the language module <u>is</u> the tentative belief that p, there is no reason to believe that such a belief will ever be formed.

Fodor himself argues that the output of the language module is a representation of an utterance's syntactic and logical form (and no more).[16] Such a representation would not be adequate to characterize what the speaker <u>said</u>, and hence, not adequate as a representation underlying the belief-like state ultimately formed if the speaker is held to be

---

16. See Fodor, 1983, p. 90.

reliable, etc. A representation of the syntactic and logical form of a sentence does not indicate, for example, what all of the referring expressions in the sentence refer to, as a representation of what the speaker said would.

Of course, Fodor may be wrong about what the output of the language module is. But, the important point is that the claim that when one hears an utterance of p, one forms the tentative belief that p before considerations about the speaker's reliability and intentions have a chance to block it, is a very speculative empirical claim. It is doubtful tnat the advocate of STM would want to rely on such a proposal to save STM from having to face the collateral information problem.

The attempt, then, to produce a generalization of an STM-based theory based on the connection between B-state B1 and utterances of the sentence, "McKinley was assassinated," has failed. In general, there is no reason to believe that an STM-based theory will contain generalizations based on connections between B1 or B0 and stimuli which will allow us to distinguish the two B-states.

4.5 Generalizations Based on Connections Between B1 (or B0) and Behavior

As in case 3, my aim here is to argue that our STM-based theory will not contain any of the relevant sort of generalizations. That is, I will argue that the theory contains no generalizations based on connections between B1 (or B0) and behavior. Because the argument will be quite similar to the one given for case 3, I'll be much sketchier with this one.

Once again, the idea is that the connection between B1 (or B0) and any particular behavior is not strong enough to satisfy the constraints of principle (I). As before, the most plausible example of a strong connection between B1 and some piece of behavior involves the utterance, "McKinley was assassinated." In particular, the most plausible example of a strong connection is the tendency of B-state B1 to cause a subject, under certain circumstances, to utter, "McKinley was assassinated." To see that this is the most plausible case, try to imagine what sort of behavior the B-state is more likely to produce.

If we are to construct a generalization based on this connection between having the B-state B1 and saying, "McKinley was assassinated," we must begin by determining the circumstances under which having the B-state will produce this behavior. Roughly, it will do so whenever (a) the subject wants someone to know that they believe that McKinley was assassinated; and (b) the subject believes that saying, "McKinley was assassinated," will lead that person to think that the subject believes that McKinley was assassinated. So far things look pretty good for the generalization we are trying to build. It looks like we will only need to build two B-states into the antecedent of the conditional, and the generalization will hold.

The problem is that a very similar generalization will hold for B0 as well. That is, having the B-state B0 will lead the subject to say, "McKinley was assassinated," if (a) the subject wants someone to know that they believe that McKinley was elected; and (b) the subject believes that saying, "McKinley was assassinated," will lead that person to think that the subject believes that McKinley was elected. The question then becomes whether these two generalizations are in fact distinct. In particular, is

the D-state corresponding to the subject's desire that some person know that they believe that McKinley was elected, distinct from the D-state corresponding to the subject's desire that some person know that they believe that McKinley was assassinated; and, is the B-state corresponding to the subject's belief that saying, "McKinley was assassinated," will lead the person in question to think that the subject believes that McKinley was assassinated, distinct from the B-state corresponding to the subject's belief that saying, "McKinley was assassinated," will lead the person in question to think that the subject believes that McKinley was elected? If the two D-states are identical, and the two B-states are identical, then the two generalizations will be identical. In that case, the same generalization will apply equally well to B1 and B0, and cannot be used to distinguish them. And, of course, we have no reason to assume that either the pair of D-states or the pair of B-states are distinct. The only apparent difference between the two D-states and the two B-states involves, speaking folk-psychologically, the difference between the concepts of election and assassination. And, the question we have been struggling with is whether, for an STM-based theory, that difference is enough to make two cognitive states distinct. We still have no reason to think that it is. Hence, the contemplated generalization connecting B1 with utterances of, "McKinley was assassinated," will not distinguish B1 from B0.

4.6 Conclusion and Review

I've now discussed the four kinds of generalizations which might distinguish B1 and B0. And, we have not found a single generalization which can do the job. We may, therefore, conclude that no generalizations of an STM-based theory will distinguish B1 and B0, and that for an STM-based theory these belief-like states of Mrs. T must be regarded as one and

the same cognitive state. But, now consider the consequences of this outcome.

At the beginning of this section, I described two conversations we might have with Mrs. T. In the first, we say to Mrs. T, "If McKinley was assassinated, then he is buried in Ohio," and she replies, "Well then, he is buried in Ohio." This is Stich's example. In the second conversation, we say to Mrs. T, "If McKinley was elected, then he is buried in Ohio," and she simply shrugs her shoulders. In attempting to explain why Mrs. T responds differently in the two conversations, we posited two distinct B-states, B1 and B0, the former being associated with the sentence, "McKinley was assassinated," the latter being associated with the sentence, "McKinley was elected." We then explained the difference in Mrs. T's responses in the two conversations by supposing that she is in B1 but not in B0. Because she is <u>not</u> in B0, the generalization which accounts for her response in the first conversation does not apply in the second conversation, and there is no reason to expect her to say, "Well then, he is buried in Ohio." But, this attempt to explain the difference in Mrs. T's responses depends crucially on the fact that she is in B-state B0 and not B1. If these two B-states turn out to be the same B-state, this attempt collapses. So, one of the consequences of having to regard B1 and B0 as the same cognitive state is that an STM-based theory will not be able to explain the outcomes of our two conversations with Mrs. T (although in section 5 I'll discuss a way that STM can be modified to avoid this problem).

But, the consequences of the inability of an STM-based theory to distinguish between B-states B1 and B0 go far beyond Mrs. T. Since principle (I) guarantees that the cognitive states that survive in Mrs. T

have not changed in their type identity, the predecessors of B1 and B0 in the younger Mrs. T will also be of the same cognitive type. Since the younger Mrs. T was, as far as we know, completely normal, we may assume that the B-states corresponding to your belief that McKinley was assassinated and your belief that McKinley was elected are of the same type as well. And, since very little in the argument of this section depended on the particular concepts associated with B1 and B0, the same could be said of countless other pairs of beliefs. The argument will go through unless the beliefs in question involve concepts that are so strongly tied to behavior and stimuli that the intervention of other, even bizarre, beliefs cannot break the link. Such concepts might include color or other sensory concepts, but not much more. So, your belief that lambs are gentle and my belief that lions are ferocious will be indistinguishable. Your belief that volcanoes are full of lava and my belief that coffeepots are full of coffee will be indistinguishable; etc. In the next section, I will argue that this situation is intolerable for a theory of cognition. But, first I want to review where things now stand in my attempt to saddle STM with the collateral information problem and also describe the somewhat technical argument of this section in less formal terms. Then, in the next section, I will modify the conclusion of this argument slightly but maintain that STM is still left in an untenable position unless it gives up principle (I) thereby opening itself up to the CI problem.

In section 1 I discussed the collateral information problem as it was raised by Putnam. The main point of the section was to demonstrate that Putnam's arguments creates a difficult problem for RTM that does not have an uncontroversial solution. Section 2 described Stich's version of the CI problem, the example of Mrs. T, and how he sees it creating problems for

RTM. In section 3 I described STM in some detail. I showed that STM would be vulnerable to the CI problem if it were not for principle (I), which holds that the generalizations of an STM-based theory must apply to a particular cognitive state independently of the other cognitive states a subject happens to be in (or not to be in). I then argued in section 4 that principle (I) has the consequence that STM-based theories will not be able to distinguish belief-like states as finely as it must. Roughly put, it will not be able to distinguish beliefs with the same logical form.

Principle (I) captures the idea that in typing cognitive states it is not the actual causal links between that cognitive states and other cognitive states, stimuli, and behavior that matter. Rather, what counts is the state's potential links. The driving force behind the argument in section 4 is the idea that potential links come too cheaply. The belief that McKinley was elected has the same potential causal links as the belief that McKinley was assassinated. The belief that McKinley was elected has the potential to produce the belief that McKinley is dead. And, it is likely to do so if the subject also happens to believe that people who have been elected are dead. The belief that McKinley was elected has the potential to be caused by the hearing of an utterance of, "McKinley was assassinated." And it is likely to be so caused if the subject happens to believe that people normally express the thought that McKinley was elected by uttering, "McKinley was assassinated"; that the utterer is a reliable source of information, etc. The belief that McKinley was elected has the potential to cause the subject to utter, "McKinley was assassinated." And it is likely to do so if the subject happens to desire that people know of her belief, and happens to believe that uttering, "McKinley was

assassinated," will lead people to think that she believes that McKinley was elected.

The causal links between the belief that McKinley was assassinated (or any other belief) and most of the cognitive states and stimuli and behavior to which it is linked depend on the presence of certain other cognitive states. Once these dependencies are spelled out, it becomes clear that if other, similar cognitive states were present instead--for example, the belief that being elected entails dying, instead of the belief that being assassinated entails dying--beliefs other than the original one would have the same causal links that the original belief actually has. In other words, it becomes clear that other beliefs have the same potential causal links as the original belief. And, since potential causal links are all that count toward type identity of a belief-like state, STM is unable to distinguish the two original belief from the others. And, since these new beliefs may be rather dissimilar to the original, STM ends up with a far too coarse-grained individuation of cognitive states.

5. B-states with the Same Logical Form are of the Same Type

In this section I want to make a distinction that was not made in the previous section. The distinction will be between the claim that an STM-based theory can't distinguish two cognitive states and the claim that an STM-based theory must regard two cognitive states as being of the same type.[17] I will ultimately conclude that an STM-based theory can

---

17. It is important to remember that questions like, What makes you think you've got two states if you can't distinguish them? are not appropriate here. For we've assumed following Stich that the two states of interest are originally specified neurophysiologically. The question is whether the two (neurophysiologically specified) states can

distinguish between B-states such as B0 and B1, but must still regard them as being of the same type. This result, I will argue, is still unacceptable, and, therefore, principle (I) must still be rejected. To motivate these refinements, I now want to discuss an argument that claims that B-states B1 and B0 can be distinguished after all.

Suppose, the argument begins, that there is a universal human response to being "informed" of a proposition of the form, 'If p, then p."[18] Call this response the t-response (Among English speakers, the t-response might consist of uttering the words, "Of course; do you take me for an idiot!") Then, our theory will contain the following generalization[19]:

(T) Given any subject S and any wff A, if the subject comes to be in a B-state mapped to the wff A --> A, then S will produce the t-response.

Now, let's introduce some new B-states of Mrs. T. B5 is the B-state Mrs. T comes to be in if we tell her, "If McKinley was assassinated, then McKinley was elected." B6 is the B-state that Mrs. T comes to be in if we tell her, "If McKinley was assassinated, then McKinley was assassinated." Now, the generalization just mentioned will apply to B-state B6 but not to B-state B5. That is, if we tell Mrs. T, "If McKinley was assassinated, then McKinley was assassinated," she produces the t-response. But, if we

---

be distinguished at the cognitive level. This question makes perfect sense.

18. The argument was suggested to me by James Higginbotham.

19. For the sake of simplicity, I am assuming the picture of communication that I rejected in the previous section. That is, I am assuming that telling someone, "If p, then p" leads them to explicitly formulate the normally implicit belief that if p, then p. The argument being presented here could be formulated without this assumption, but not without adding needless complexity.

tell her, "If McKinley was assassinated, then McKinley was elected," she does not. Hence, B6 is to be assigned to a wff of tne form A --> A, but B5 is not.

Now, generalization (1) discussed in section 4.1 should apply to B-states B6, B1 and B0 respectively. This means that B-state B6 will have to be mapped to the wff W1 --> W0. A listing of all the B-states we have now posited and their assignments is given below. I assume there that B-state B6 will be mapped to W0 --> W0, though this will not matter.

Verbal stimulus/behavior	B-state stimulus/behavior is connected to	Wff B-state is mapped to
"McKinley was assassinated"	B1	W1
"Well then, he is buried in Ohio"	B2	W2
"If McKinley was assassinated, then he is buried in Ohio"	B3	W1 --> W2
"If McKinley was elected," then he is buried in Ohio"	B4	W0 --> W2
"McKinley was elected"	B0	W0
"If McKinley was assassinated, then McKinley was elected"	B5	W1 --> W0 **(NOT OF THE FORM A --> A)**
"If McKinley was assassinated, then McKinley was assassinated"	B6	W0 --> W0

Putting the fact that B5 is assigned to the wff W1 --> W0 together with the fact that, in virtue of not satisfying generalization (T), B5 is not to be mapped to a wff of the form A --> A, we can conclude that wffs W1 and W0 must be distinct. Since distinct wffs correspond to distinct cognitive state types, we may conclude that B1 and B0 must be assigned different

types by our STM-based theory. But, isn't this precisely the conclusion which I argued could not be reached in section 4?

An analogy will help us untangle the situation. In "The Meaning of 'Meaning'," Putnam confesses that he cannot tell the difference between beech trees and elm trees. Consequently, he argues, his concept of a beech tree is "exactly the same as" his concept of an elm tree.[20] Putnam evidently believes that the connection between the concepts and the associated phonological and morphological forms is not sufficient to create a difference between the two concepts. If Putnam is right about this, then the way things stand with his concepts of elm trees and beech trees (assuming that his knowledge of beeches and elms has not expanded) is precisely the same as the way things stand with B-states B1 and B0. Putnam surely does not mean to suggest that he has just one concept of the relevant sort. If he did, then he would surely believe that beech trees are elm trees, which, presumably, he does not. The point is not that the two concepts are identical; that he does not believe that beeches are elms proves that. Rather, the point is that although the concepts are distinct, they have precisely the same content (on the assumption that the connection with distinct phonological and morphological forms is not significant in this regard).

Similarly, generalization (T) reveals that B0 and B1 are distinct B-states. If they weren't, then telling Mrs. T, "If McKinley was assassinated, then McKinley was elected" would provoke the t-response. However, there is still no basis for assigning the two B-states distinct

---

20. Putnam, 1975, p. 143.

wffs or for regarding them as being of distinct types for they still satisfy exactly the same pattern of generalizations.

Why, one might ask, can't an STM-based theory resolve this situation by arbitrarily assigning the two B-states to distinct wffs if the B-states are known to be distinct? One can even continue to honor the idea that different wffs correspond to distinct cognitive state types by simply attaching indices to wffs. So, for example, we could acknowledge that B1 and B0 are distinct B-states but of the same type by associating both of them with the wff 'Am,' but attaching different indices. Thus, B1 might be associated with the indexed wff 'Am^1' and B0 with the indexed wff 'Am^0.' In general, we could associate with every distinct pattern of generalizations a single cognitive state type and a single wff, but an indefinite number of indexed wffs. It would be understood that when generalizations of the theory quantify over wffs, they quantify over indexed wffs. Hence, a generalization might refer to all (indexed) wffs of the form A^i.

There is, as far as I can see, nothing wrong with this proposal. It maintains STM's commitment to individuating cognitive states according to the generalizations they satisfy. At the same time, it allows the explanation of why Mrs. T reacts differently in the two conversations described in section 3 to go through. For all that was required there was that B-states B1 and B0 be assigned distinct wffs. For these purposes a distinction captured by arbitrarily assigned indices will be sufficient.

The proposal does not, however, change the fact that B-states B1 and B0 must still be regarded as cognitive states of the same type. The point becomes clear when we consider intersubjective comparisons. Since indices

are attached arbitrarily to the wffs that distinct B-states are mapped to, the pattern of attachments will be different in different individuals. Suppose, for example, that Mrs. T has a twin. The B-state that leads Mrs. T to say, "McKinley was assassinated" might be assigned to the very same wff with the very same index as the B-state that leads Mrs. T's twin to say, "McKinley was elected." If these are the only B-states the two women have, then our STM-based theory would not be able to regard the two women as having different B-states.

And, what goes for Mrs. T and her twin goes for me and you as well. An STM-based theory still won't be able to distinguish between my belief that McKinley was elected and your belief that McKinley was assassinated, or between your belief that lambs are gentle and my belief that lions are ferocius; etc. Evidently, as far as intersubjective comparisons are concerned, the individuation of cognitive states of an STM-based theory will still be too coarse-grained.

That Stich needs to be able to quantify over more finely-grained cognitive states is evident in his discussion of an argument against STM proposed by Patricia Churchland. In setting up the argument, Stich writes:

> One of the ways in which a content-based theory may specify connections between stimuli and behavior is by linking specific stimulus types to specific beliefs; thus:
>
> (11) For all subjects S, when an elephant comes into view, S will typically come to believe that an elephant is in front of him.
>
> This sort of generalization can be mimicked easily enough by a syntactic theory:

(12) For all subjects S, when an elephant comes into view, S will typically come to have a sequence of symbols, E, in his B-store.

(Stich, 1983, p. 178.)

Stich goes on to discuss the possibility that RTM-based theories could produce more general forms of such a generalization. The important point for us, however, is that generalization (12) relies on the ability of an STM-based theory to regard the state I am in when an elephant is in front of me and the state someone else is in when there is an elephant in front of them as being cognitive states of the same type. There is nothing wrong with this, but the state in question will be the same as the one I would be in if a tiger rather than an elephant were in front of me. Hence, an STM-based theory will get into trouble when it attempts to mimic the following generalization: for all subjects S, when an elephant comes into view, but not a tiger, S will typically come to believe that an elephant is in front of him, but not come to believe that a tiger is in front of him.

To summarize: the argument in section 4 turns out to have been slightly overstated. Generalizations such as (T) above may be capable of identifying B-states that satisfy the same pattern of generalizations but must nevertheless be assigned distinct wffs. STM can be modified to allow for such distinctions by using some sort of indexes: B-states that satisfy the same pattern of generalizations must still be considered to be of the same type, but can be distinguished by attaching distinct indexes to the wffs they are mapped to. This move allows us to explain why Mrs. T responded differently in the two conversations discussed in section 3. However, since B-states of the same logical form are still regarded as being of the same type, the individuation of B-states will still be too

coarse-grained. This problem manifests itself particularly in contexts where B-states are being compared across individuals.

We see, then, that principle (I) leaves STM with a notion of cognitive state that is unacceptably coarse-grained. But, we saw earlier, in section 2, that without principle (I), STM will have to face the collateral information problem. Perhaps STM can find some other solution to the collateral information problem. But, until this has been shown, it would be ill-advised for anyone to abandon RTM in hopes of an easy way out of it.

Eliminative Connectionism and Cognition as Pattern Association

1. Connectionism

Cognitive science has seen an explosion of interest in connectionist, or parallel distributed processing (PDP)[1], models in recent years. A connectionist model consists of a very large set (a network) of individual processing units (nodes). Each node is connected to many, if not all, of the other nodes. Each node is associated with a level of activation, which changes over time. Often the activation level is allowed to take on any value between an upper and a lower limit. In other models the activation levels must assume discrete values. In some cases the nodes are restricted to an activation level of 1 or 0; i.e., they are either "on" or "off." Each node receives input from the nodes it is connected to. The connections are weighted, and the input one node receives from another is typically the product of the activation level of the second node and the weight of the connection. The activation level of each node is

---

1. The word "distributed" in the phrase "parallel distributed processing" refers to a feature which many connectionist models do not have. In particular, "distributed representation" refers to the fact that in some models concepts (or whatever one wants to call the items being encoded in the network) do not correspond to individual nodes in a network or even to a small set of nodes. Instead, they correspond to patterns of activity over a very large number of nodes. The same large set of nodes may be used for a number of concepts. Which concept is active at any given time depends on the pattern of activity in the set of nodes at that time. Since only some connectionist models use distributed representation, it would seem that "parallel distributed processing" is not synonymous with "connectionism," and shouldn't be used in that way. However, Rumelhart, McClelland and the PDP Research Group have written a two-volume work entitled Parallel Distributed Processing, which is something of a connectionist bible and contains many models which don't have distributed representation. I follow them in using the two phrases interchangeably.

periodically (continuously in some models) updated as a function of the input it is receiving from the nodes it is connected to. This function, usually called the activation rule, may be probabilistic. Updating of activation levels usually occurs in all nodes simultaneously.

A "run" of a PDP network typically begins with the activation levels of some nodes being set manually. This setting of activation levels constitutes the stimulus for which the network must generate a response. The clock then begins to run, and activation levels for all nodes are then determined on the basis of activation levels at previous times, connection weights and the activation rule. Typically, activation levels of the various nodes stabilize after some number of cycles. This stable set of activation levels constitutes the network's response to the stimulus. Methods for interpreting activation levels, either individually or collectively, allow the network to be seen as a model of some cognitive process. Sometimes means are provided for adjusting the connection weights in a given network on the basis of "experience." Learning, or cognitive development, is modeled in this way.

In sum, connectionist networks are large sets of simple processing units, which are densely interconnected, and operate in parallel. Connectionists have often advertised such networks as alternatives to models in cognitive science that are based on standard, serial computer architectures.[2] Three differences between connectionist architectures and

---

2. Feldman and Ballard, for example, write,

> The fundamental premise of connectionism is that individual neurons do not transmit large amounts of symbolic information. Instead they compute by being appropriately connected to large numbers of similar units. This is in sharp contrast to the conventional computer model of

conventional computer architectures (Von Neumann architectures) are usually cited. First, there is the issue of parallelism vs. seriality. In conventional architectures, processing occurs one step at a time. In PDP networks, processing occurs in all nodes at the same time. Hence, the networks are said to be "massively parallel." The second architectural difference is that there is no centralized control in connectionist architectures. A Von Neumann machine has a central processing unit, or executive, which is the locus of control in the machine.[3] The third difference is that the notion of a symbol being operated on is not fundamentally involved in connectionist architectures. The operations of reading and writing symbols are primitive operations in Von Neumann architectures. Because they take notions of storing, retrieving, and transforming symbols for granted, models based on conventional computer architectures are sometimes characterized as symbol-processing models. In sum, then, connectionist models have often been advertised as alternatives to symbol-processing models.

The function of the node in a PDP network is obviously supposed to be reminiscent of the role of a neuron in the brain. It is not hard to imagine how connectionist notions might apply to collections of neurons. Nodes could be individual neurons; activation levels could be time-averaged

---

> intelligence prevalent in computer science and cognitive psychology. (Feldman and Ballard, 1982, p. 208.)

3. Connectionists tend to depict the contrast between connectionist and non-connectionist computer architectures in stark terms. In fact, however, there are many architectures which could hardly be called connectionist, but which incorporate either elements of parallelism or decentralized control. Parallel (but non-connectionist) architectures in particular, such as that presented in Hillis, 1985, have been much publicized of late. For an example of a non-connectionist architecture with decentralized control, see Agha, 1986.

firing rates; connection strength could correspond to the strength of synaptic connections; etc.[4] Although there is some disagreement among connectionists about the extent to which their networks should be understood as directly modeling what occurs in the brain, almost all connectionists cite something like "biological plausibility" as an advantage of connectionist models.[5] The idea that PDP networks are to be thought of as networks of neurons invites an advocate of old-fashioned symbol-processing models to claim that she and the connectionist are simply looking at cognitive processing at different levels of abstraction. Of course, she says, if you are looking at what the individual neurons are doing, then conventional computer architectures might not be appropriate. But, that doesn't mean that they aren't appropriate at a higher level of abstraction. Similarly, if you look at conventional computers at a low enough level, processing does not consist of the execution of complex rules or the manipulation of complex symbolic representations: at a low enough level it consists solely of transistors switching currents in response to the switches of other transistors.[6]

---

4. On the other hand, there are many different ways of making this sort of talk precise, and its not at all clear what the most appropriate way would be.

5. See, for example, J. A. Anderson, 1983, pp. 799-803; Feldman and Ballard, 1982, p. 206; and Rumelhart, 1984, pp. 60-61. McClelland, Rumelhart and Hinton are more restrained, but still cite biological plausibility as part of the appeal of PDP models (McClelland, Rumelhart and Hinton, 1986, pp. 10-11). Smolensky, it should be noted, appears to be somewhat wary of resting much weight on the appeal to biological plausibility (Smolensky, 1987, pp. 7-8).

6. In regard to the issue of levels of abstraction, see the exchange between Broadbent (Broadbent, 1985) and Rumelhart and McClelland (Rumelhart and McClelland, 1985).

This obvious response to the connectionist challenge has led to much discussion in the literature on the question of the extent to which connectionist and symbol-processing models are pitched at different levels, and if they are, the extent to which connectionism should still be viewed as an alternative to work in the symbol-processing tradition. These questions about the proper relation between connectionist and symbol-processing models constitute the topic of this paper. In particular, I will be concerned with a view, which, borrowing the terminology of Steven Pinker and Alan Prince (Pinker and Prince, 1988), I will call eliminative connectionism.

Eliminative connectionism says that connectionist models and symbol-processing models are pitched at different levels of analysis, but that the level of analysis associated with the former is more fundamental. Symbol-processing models are, on this view, only "approximate" or metaphorical descriptions of an underlying connectionist reality. Symbol-processing models do not accurately characterize actual cognitive processes at any level of abstraction.

I will be focusing especially on a paper by Paul Smolensky, entitled "On the Proper Treatment of Connectionism" (Smolensky, 1987). I will claim that the advocacy of eliminative connectionism on the part of Smolensky and others is a consequence of taking a particular kind of connectionist model, pattern association, as a paradigm for cognition in general. In other words, eliminative connectionism rests on a view of cognition as pattern association. I will argue, based on considerations drawn from within the connectionist framework, that the view of cognition as pattern association is untenable. And, I will argue that once the view of cognition as pattern association is abandoned, there is no reason to accept the eliminativist

connectionist's view of the proper relation between connectionist and symbol-processing models. I will conclude that it would be wise for the connectionist to take a more open-minded view about the proper relation between connectionist and symbol-processing models than that suggested by eliminative connectionists.

2. Eliminative Connectionism.

2.1 Views on the Relation Between Connectionist and Symbol-Processing Models

In a recent article, Steven Pinker and Alan Prince discuss the different stances that can be taken on the question of the relation between connectionist and symbol-processing models. They describe three possible views on the matter. The views are not supposed to cover all the possibilities. Rather, they represent three points on a continuum of views. A useful way of thinking about the continuum is to consider the relative importance within each view of symbol-processing accounts of cognition. In the first view, which we may think of as defining the rightmost point in the continuum, symbol-processing accounts loom large. According to this view, which Pinker and Prince label implementational connectionism, connectionist models describe the elementary information processes out of which the algorithms described in symbol-processing accounts are built. Research into connectionist models might produce new suggestions about the primitive information processes of the brain, but this would function mainly to select from among the already available symbol-processing accounts of cognition.

The view in the center of the spectrum is called revisionist-symbol-processing connectionism. On this view groups of connectionist networks will again implement the primitive procedures out of which symbol-

processing algorithms are constructed. Symbol-processing accounts will thus retain much of their importance. But, the view allots greater importance than implementational connectionism does to connectionist models. For on this view, the sort of primitive information processes used by the correct symbol-processing algorithms will largely be discovered by research into connectionist models.

For example, connectionists have had success in developing networks that perform pattern completion (much more on this later). If such a network contains information about baseball teams, one might be able to retrieve all the information stored about a particular team by feeding the network the names of a few players. The revisionist-symbol-processing connectionist might conclude that a memory fetch involving the retrieval of an item from memory by producing an incomplete description of the item, should be regarded as a primitive process in symbol-processing models of cognition.

According to revisionist-symbol-processing connectionism, the primitive processes available to symbol-processing algorithms are largely unknown, and, therefore, the shape of the correct symbol-processing accounts will undergo substantial change based on the results of research into connectionist models. Hence, although symbol-processing accounts play a crucial, if not the crucial, role in the ultimate account of cognition, the direction of cognitive science will largely be driven by connectionist research.

The view on the far left of Pinker and Prince's continuum is eliminative connectionism, the position at issue in this paper. According to this view, many of the symbolic structures that get processed in symbol-

processing models will typically not correspond to any discrete components of connectionist models. Similarly, the individual processing steps of symbol-processing algorithms will not correspond to discrete events in the operation of connectionist models. Groups of connectionist networks do not implement symbol-processing algorithms at all on this view. Symbol-processing models may approximately capture the relation between the input to these models and their behavior. In so far as they do, symbol-processing descriptions of connectionist models will have heuristic value in generating behavioral predictions, and provide an "approximate" description of the behavior of the model. But, the symbol-processing description of cognitive processes will not be literally true at any level of analysis. Symbol-processing models of cognition are clearly of secondary importance on this view; connectionist models are where all the "action" is.

The position I have just described encompasses many ideas. But, the heart of the view concerns: (1) the claim that connectionist networks do not implement symbol-processing algorithms (in the sense that, for example, some assembly-level code implements an algorithm specified in a programming language); and, (2) the claim that symbol-processing descriptions of cognition are not, therefore, true characterizations of actual cognitive mechanisms at any level of abstraction. The upshot is, of course that connectionist models are more fundamental and more important than their symbol-processing counterparts.

The claim of non-implementation can be explicated in terms of the impossibility of mapping elements of a symbol-processing algorithm onto the events in the underlying connectionist network. For example, the eliminative connectionist might claim that individual steps of symbol-

processing algorithm intended to model a cognitive process will not map in any neat or simple way onto events or sets of events in the operation (either during the processing of the mature networks, or in their training history) of the connectionist networks that correctly describe the cognitive process. However, to repeat, the heart of the eliminative connectionist view is simply that ultimately connectionist networks will not turn out to implement symbol-processing algorithms, and, therefore, no symbol-processing algorithms will accurately describe actual cognitive processing.

2.2 Eliminative Connectionists

A number of authors appear to espouse eliminative connectionism. Consider, for example, the view of connectionism presented by Daniel Dennett in "The Logical Geography of Computational Approaches (A View From the East Pole)." Dennett describes PDP systems in the article, emphasizing that the values computed by the nodes in a PDP network do not by themselves have any "external-world semantic role" and asks, "How then do we ever get anything happening in this system that is properly about Chicago?" Dennett answers that on the connectionist view in question, there is a higher level of description at which one can attribute external-world significance to the activity of the networks. But, the relationship between the elements posited at this level is not computational, but "statistical, emergent, holistic." Hence, he concludes:

> in this vision the low, computational level is importantly unlike a machine language in that there is no supposition of a direct translation or implementation relation between the high-level phenomena that do have an external-world semantics and the phenomena at the low level. If there were, then the usual methodological precept of computer science would be in order: ignore the hardware since the idiosyncrasies of its particular style of implementation add nothing to the phenomenon ... (Dennett, 1986, pp. 69-70.)

Of course, in symbol-processing models of cognition, at least some of the symbols that are manipulated are supposed to receive external-world interpretations. Hence, symbol-processing models are supposed to be what Dennett calls higher level descriptions. But, at that level, the interactions and relationships between elements is not computational according to Dennett's understanding of connectionism. Since symbol-processing models ascribe computational relations to elements at this level, these models do not accurately characterize cognitive processes on Dennett's picture.

Douglas Hofstadter paints a similar picture in "AI: Subcognition as Computation." He uses a metaphor to illustrate doubts about conventional approaches to AI. The brain is compared to a colony of ants, with ants playing the role of neurons, and teams of ants corresponding to concepts or thoughts. Hofstadter then questions whether there exist formal, computational rules at the level of the teams that describe the behavior of the colony. The idea, of course, is that the only level at which the brain is truly computational is at the level of the neurons, and that at any higher, more abstract level, the brain is no longer computational.

Hofstadter concludes:

> The premise of AI is that thoughts themselves are computational entities at their own level. At least, this is the premise of the information-processing school of AI, and I have very serious doubts about it. (Hofstadter, 1983, p. 278.)

Hofstadter is not talking here specifically about an underlying connectionist structure. But, elsewhere in the paper, he makes favorable indirect references to connectionist research. And, it is no secret that

Hofstadter is sympathetic to the connectionist program.[7] To the extent that Hofstadter can be thought of as a connectionist, he is clearly an eliminative connectionist.

The widely noticed "bible" of connectionism is, Parallel Distributed Processing, a two-volume work by David Rumelhart, James McClelland and the PDP Research Group. The main locus of the metatheoretical, or ideological, elements of the book is Chapter 4 of Volume I, an essay by Rumelhart and McClelland entitled, "PDP Models and General Issues in Cognitive Science." Here Rumelhart and McClelland opt also for the eliminative connectionist view. Setting up part of the discussion, they write:

> It might be argued that a model such as, say, schema theory or the ACT* model of John Anderson (1983) is a statement in a "higher level" language analogous, let us say, to the Pascal or LISP programming languages and that our distributed model is a statement in a "lower level" theory that is, let us say, analogous to the assembly code into which higher level programs can be compiled.

But, they respond

> Pascal code will, in general, compile into only a small fraction of the possible assembly code programs that could be written. Since there is every reason to suppose that most programming that might be taking place in the brain is taking place at a "lower level" rather than a "higher level," it seems unlikely that some particular higher level description will be identical to some particular lower level description. We may be able to capture the actual code approximately in a higher level language--and it may often be useful to do so--but this does not mean that the higher level language is an adequate characterization. (Rumelhart and McClelland, 1986a, pp. 124-125.)

---

7. When Dennett presented a version of his paper (Dennett, 1986) at the Sloan Conference at MIT in 1984, he had Hofstadter give a brief presentation on connectionist research as an adjunct to the paper. Furthermore, Dennett credits Hofstadter with being a major source of the ideas in his paper, ideas which are rather sympathetic with the connectionist program.

Later, Rumelhart and McClelland introduce the idea that higher level symbol-processing models are "approximately" correct descriptions of an underlying connectionist structure.

> We view macrotheories as approximations to the underlying microstructure which the distributed model presented in our paper attempts to capture. As approximations they are often useful, but in some situations it will turn out that an examination of the microstructure may bring much deeper insight. (Rumelhart and McClelland, 1986a, p. 125.)

Terming the macrotheories approximations actually suggests that the theories might provide an important, true description of the structure of cognition, just as Newtonian mecnanics provides an important, true (provided that the laws are not claimed to hold precisely) description of physical reality. But, Rumelhart and McClelland go on to write:

> Tnus, although we imagine that rule-based models of language acquisition--the logogen model, schema theory. prototype theory, and other macrolevel theories--may all be more or less valid approximate macrostructural descriptions, we believe that the actual algorithms involved cannot be represented precisely in any of those macrotheories. (Rumelhart and McClelland, 1986, p. 126.)

Tne pnrase "the actual algorithms involved" betrays, I believe, the fact that when the macrotheories are characterized by Rumelhart and McClelland as "valid approximate macrostructural descriptions," the validity lies in their heuristic value and not in their corresponding to actual mental processes.

My main target in this paper will not be Dennett, Hofstadter or Rumelhart and McClelland. My arguments will apply directly to Rumelhart and McClelland, but their main target is Paul Smolensky's essay, "On the Proper Treatment of Connectionism." To understand Smolensky's statement of eliminative connectionism, we must begin by reviewing some of his terminology. The first piece of terminology we need to look at, "the intuitive processor," is introduced in the following passage:

> What kinds of programs are responsible for behavior that is not conscious rule application? I will refer to the virtual machine that runs these programs as the intuitive processor. It is (presumably) responsible for all of animal behavior, and a huge portion of human behavior: perception, practiced motor behavior, fluent linguistic behavior, intuition in problem solving and game playing ... (Smolensky, 1987, p. 3)

Clearly Smolensky believes that the operation of the intuitive processor constitutes the bulk of cognition. On anybody's view, only a very minor part of cognitive processing is the conscious application of rules.

Smolensky discusses the nature of the activity of the intuitive processor. He considers and rejects the hypothesis (labeled (3b)) tnat "the programs running on the intuitive processor are composed of elements--symbols--referring to essentially the same concepts as are used to consciously conceptualize the task domain," and comments on it as follows:

> This hypothesis has provided the foundation for the symbolic paradigm for cognitive modeling. Cognitive models of both conscious rule application and intuitive processing have been programs constructed of entities which are <u>symbols</u> both in the syntactic sense of being operated on by "symbol manipulation" and in the semantic sense of (3b). Because these symbols have the conceptual semantics of (3b), I will call the level of analysis at which these programs provide cognitive models the <u>conceptual level</u> (Smolensky, 1987, p. 4:)

For Smolensky, one of the characteristic features of traditional symbol-processing models is that its level of analysis is the conceptual level. If connectionist models could be regarded as implementations of symbol-processing models, then such models could be characterized at the conceptual level. The intuitive processor, it turns out, is properly described as a connectionist network. Smolensky directly confronts the question of the extent to which the intuitive processor can be characterized at the conceptual level in the following passage:

> The entities in the intuitive processor with the semantics of conscious concepts of the task domain are <u>complex patterns of activity</u> over many units.... The interactions between

> individual units are simple, but these units do not have conceptual semantics: they are subconceptual. The interactions between the entities with conceptual semantics--interactions between complex patterns of activity--are not at all simple. Interactions at the level of activity patterns are not directly described by the formal definition of a subsymbolic model; they must be computed by the analyst. Typically, these interactions can be computed only approximately. There will generally be no precisely valid, computable formal principles at the conceptual level; such principles exist only at the level of the units--the subconceptual level. (Smolensky, 1987, p. 6.)

Symbol-processing accounts of cognition are generally presented at the conceptual level. But, "no precisely valid, computable formal principles" apply to the intuitive processor at the conceptual level. Hence, no symbol-processing account will provide a completely accurate description of the activities of the intuitive processor. As with Rumelhart and McClelland, it is possible to interpret Smolensky here as leaving an important role for conceptual level accounts, since they might be approximately, if not precisely, valid. But, I think that the tone of the passages above makes it clear that when Smolensky says that there are no precisely valid, computational principles that apply to the intuitive processor at the conceptual level, he means that there are no such principles that have anything beyond heuristic value.[8]

---

8. This interpretation is reinforced later in the paper when Smolensky illustrates the relation between conceptual level theories and subconceptual accounts with examples. He writes of a certain set of conceptual level representations, for example, "They are informal, approximate descriptions--one might even say they are merely metaphorical descriptions--of an inference process too subtle to admit such high-level descriptions with great precision." (Smolensky, 1987, p. 23.) Smolensky seems to me to believe that conceptual level descriptions will, in general, be "merely metaphorical."

3. Pattern Association

My aim in this paper is to show that the eliminative connectionist views of both Rumelhart and McClelland (Rumelhart and McClelland, 1986a) and Smolensky are based on a very specific and overly simple view of the form of connectionist models of cognition. In particular, their view seems to be based on the assumption that connectionist models of cognition will be networks called pattern associators. More specifically, they seem to assume that connectionist models will consist either of simple pattern associators or of the kind of pattern associators Smolensky has studied under the rubric "harmony theory." In the present section I will describe these kinds of networks and explain why undue attention to them might lead one to espouse eliminative connectionism. In section 4 I will demonstrate the extent to which McClelland and Rumelhart and Smolensky rely on these sorts of networks in developing their ideological positions, charging them with subscribing to a view of cognition as pattern association. A careful look at the connectionist research program itself, section 5 will argue, suggests that the view of cognition as pattern association is overly simple. And, finally, I will show that a more realistic picture of connectionist models of cognition is not nearly so conducive to eliminative connectionism.

3.1 Pattern Associators and Auto-Associators

Connectionist networks are often designed to be pattern associators. When a pattern associator is proposed as a model for a cognitive task, the task is conceived as that of generating the right response to a stimulus. The stimulus, or input, might be a verb and the response, or output, the past tense form of the verb. Or, the input might be a person and the

output their occupation. Etc. One subset of the nodes in a pattern association network are designated as input nodes. Another subset is designated as output nodes. There may be further, "hidden" nodes, but there needn't be. Each element in the set of possible stimuli is associated with a distinct pattern of activity among the input nodes. Similarly, each possible response is associated with a distinct pattern of activity among the output nodes. If the connection weights are set properly, then, when a stimulus is encoded "manually," the network will generate the appropriate response. In other words, if the pattern of activity associated with a particular stimulus is imposed on the network externally, and the network is then allowed to run and stabilize, the pattern of activity that emerges among the output nodes will be the one associated with the correct response to the stimulus.[9]

Usually, the point is not simply to show that a network can be designed that associates the appropriate response to a given stimulus. Rather, the point is to show that the network can learn this association. This is done by providing the network with a "learning rule," and supplying it with "experience." Each unit of experience consists in the following sequence of events. A particular stimulus is encoded manually in the normal fashion. The network is allowed to run, thereby generating a particular pattern of activity among its output nodes. A special teaching input is then supplied to each node, indicating what the correct response to the given stimulus was for that node. Finally, connection weights are

---

9. When the activation rule for the network is stochastic, sometimes "running" the network involves changing the activation function over time in accordance with a measure called "temperature." The basic idea is that the activation values of the nodes should be highly random at first (when the temperature is high), and become less so as time goes on (during which time the temperature drops and the network "freezes.")

updated in accordance with a learning rule based on differences between the network's response and the correct response as encoded in the teaching input. A great variety of learning rules have been proposed, but in effect they all say: if input node A was on, and output node B was off but was supposed to be on (i.e. the corresponding teaching node was on), then increase the weight of the connection between node A and node B; if input node A was on, and output node B was on but was supposed to be off, then decrease the weight of the connection between node A and node B. Connectionists have been quite successful in designing networks that learn to associate particular stimuli with particular responses in this manner.

A special case of pattern association that is of particular interest here is the case where the set of input nodes and the set of output nodes are identical. In this case, the pattern associator is called an auto-associator. If the individual nodes correspond to meaningful features, then auto-association can be interpreted as providing the network with a partial description of a situation or object and asking the network to complete the description. In other cases auto-association is best understood as feeding the network a noisy input and asking the network to eliminate the noise.

3.2 The Room-Schemata Example

One particular example of auto-association, or pattern completion, plays a central role in Smolensky's paper and in a paper cited at an important juncture in Rumelhart and McClelland, 1986a. In the rest of this section, I'll discuss the example and its lessons in some detail.

The network in question consists of forty nodes. Each node corresponds to a "feature" of a room. The features are things such as,

coffee-pot, window, large, [has] books, etc. The activation level of each node ranges between 0 and 1. As always, the activation level of node x_j at time t+1 is a function of the activation level of node x_j at time t, the activation levels of other nodes at time t, and the weights of the connections between those nodes and x_j. The activation levels of nodes are updated one at a time.

The weights connecting the nodes were derived in the following way.[10] First, eighty sample feature-sets were constructed. This was done by asking someone to imagine a living room, e.g., and then running through each of the forty features, having the person say for each one whether or not it applied to the room they were imagining. (The person was asked to imagine a living room in one-fiftn of the cases. Similarly with kitchens, bathrooms, offices, and bedrooms.) Given these eighty feature-sets, the weight between any two nodes (features) was set as a function of the following four joint probabilities: the probability tnat both the first and second feature are included in a given feature-set, the probability that the first feature, but not the second, is included in a given feature-set, etc.

A "run" of the network consists of "telling the network that a given room has certain features", i.e. manually setting the activation level of a handful (usually two or three) of nodes at 1, and then letting the activation levels of all the nodes change in accordance with the specified

---

10. Rumelhart et al. could have set the weights by equipping the network with a learning rule and providing it with experience. Presumably, they refrained from doing so out of expediency, perhaps because of the computing expenses that would have been involved (The computational effort required to simulate the training of a connectionist network on a serial computer is apparently quite substantial).

rules for updating activation levels. The run ends when a stable set of activation levels has been reached. Because of certain aspects of the design of the network, the network will always end up with a certain set of nodes having an activation level of 1, and the rest having an activation level of 0. Hence, we can always interpret the network as telling us what other features are likely to be in the room in question. We can thus view the network as performing a pattern completion task. It is given a partial list of the features of a given room, and must then complete the list. As it turns out, in most cases, the output of the network will be nearly identical to one of just five lists of features. These lists of features correspond to a prototypical kitchen, a prototypical bedroom, a prototypical living room, a prototypical bathroom, and a prototypical office.

Among connectionists, there is a popular way of conceiving of what the network does which helps to explain why it is so likely to end up describing a prototypical kitchen, or a prototypical bedroom, etc. Consider first a very simple network containing just two nodes whose activation levels range between 0 and 1. We can create a function, G (goodness-of-fit), which associates with every pair of activation levels for the two nodes, a measure of how consistent the activation levels are with the connection weights. So, for example, if the two nodes are mutually inhibitory, G will be greatest for the pairs (1,0) and (0,1). We can modify G to take into account the nodes' initial activation levels. So, if we begin with the first node having an activation level of .5 and the second node having an activation level of .2, and modify G accordingly, G ought to be greatest for (1,0) since this is most consistent with both the connection weights and input activation levels. We can graph G (either

the original version or the modified version) as a surface lying over the square in the xy-plane whose corners are (0,0), (0,1), (1,0) and (1,1). Connectionists call these surfaces landscapes. Given our function G, together with a method for modifying it for a particular set of input activation levels, we can think of the network as having a goodness-of-fit landscape (the one determined by the original function) which then gets distorted depending on the input activation levels given to the network.

It turns out that there is a function G of the sort described such that the behavior of the room-schemata network and other networks of its kind is accurately characterized as climbing a path of steepest ascent toward a peak (or local maximum) in some distortion of its goodness-of-fit landscape. The particular distortion of the landscape it climbs is determined by the input to the network.

The room-schemata network contains 40 nodes. So, for this network, G would have to be plotted in 41-space. The graph would "lie over" a 40-cube. Because the activation rule for the network tends to drive nodes into activation levels of 1 or 0, the peaks in the goodness-of-fit landscape all occur at corners of the 40-cube. It turns out that the room-schemata network has five dominant peaks in its goodness-of-fit landscape. The vectors of activation levels that these peaks lie over correspond, of course, to the five lists of features that the network tends to produce. In other words, the dominant peaks in the goodness-of-fit landscape correspond to the prototypical kitchen, bedroom, living-room, bathroom, and office discussed above. Because these peaks are so dominant, whenever the input to the network consists in turning on just a handful of features, one or more of the five peaks will dominate the relevant distortion of the goodness-of-fit landscape, and the path of steepest ascent will lead to one

of them. This explains why the network is so likely to settle into a state associated with one of the five prototypical rooms.

Obviously, it is impossible to visualize the goodness-of-fit landscape in 41-space, but Rumelhart et al. present several three-dimensional cross-sections of the landscape and some of its distortions. Some of these are reproduced below in Figure 1. Figure 1A is a three-dimensional cross-section of the original landscape. Figure 1B is the landscape associated with an input in which the "oven" node alone is turned on. And, figure 1C is the landscape associated with an input in which the "bed" node alone is turned on.

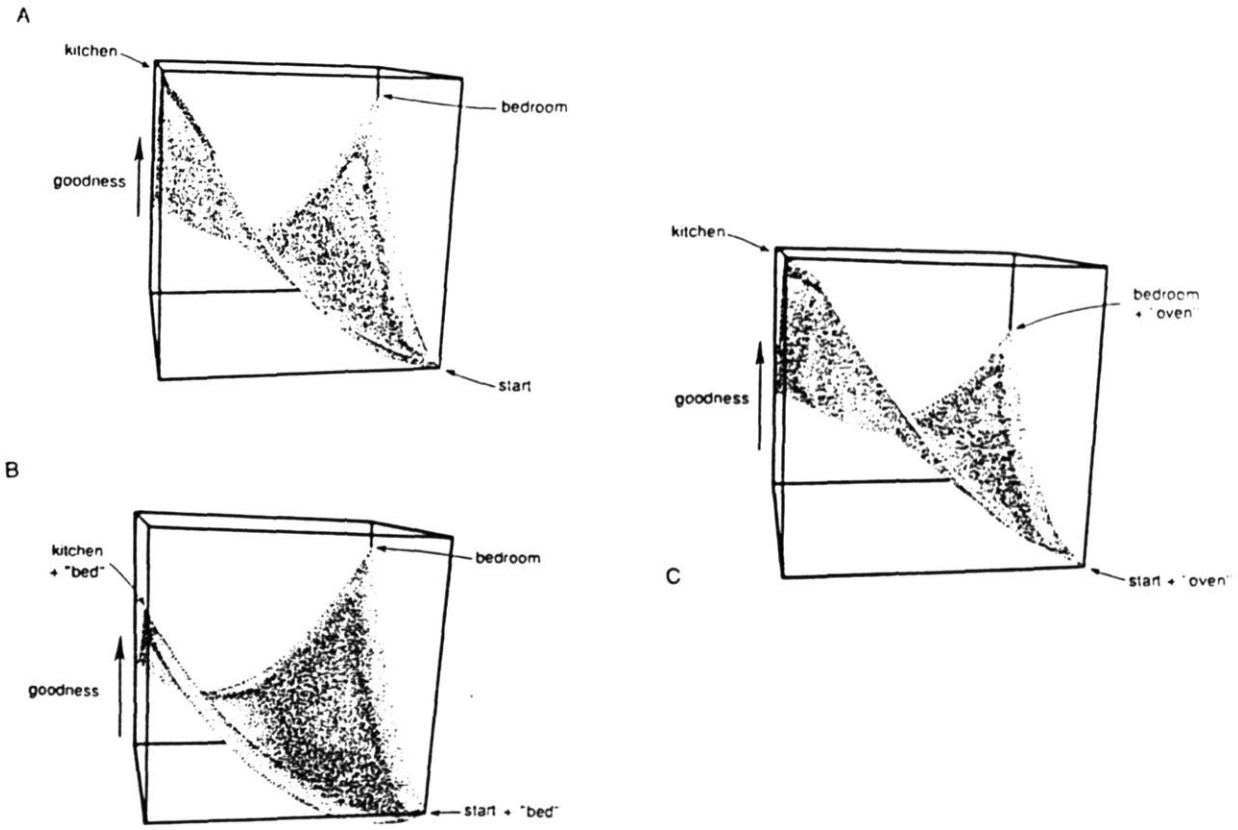

Figure 1. Reproduced from Rumelhart et al., 1986

3.3 Lessons of the Room-Schemata Example.

Rumelhart et al. (Rumelhart et al., 1986), and Smolensky (Smolensky, 1987) both make much of the example just described. Their point is basically this. The network behaves as if it had encoded in it somehow five schemata describing a typical kitchen, a typical bathroom, etc. There is an inclination to think of it as having schemata, complete with subschemata, variables with default values, etc.[11] Thinking of the network in this way allows us to generate good predictions of the networks behavior. But, in fact, the network has no schemata encoded in it all. The information associated with the schemata is all encoded in the network in the weights of the connections. The only thing in the network corresponding to these schemata are patterns of activity that turn out to be highly stable.

Smolensky writes,

> Looking closely at the harmony landscape we can see that the terrain around the "bathroom" peak has many of the properties of a bathroom schema: variables and constants, default values, schemata imbedded inside of schemata, ... The system behaves as though it had schemata for bathrooms, offices, etc., even though they are not "really there" at the fundamental level: these schemata are strictly properties of a higher-level description. They are informal, approximate descriptions--one might even say they are merely metaphorical descriptions--of an inference process too subtle to admit such high-level descriptions with great precision. Even though these schemata may not be the sort of object on which to base a formal model, nonetheless they <u>are</u> useful descriptions that help us understand a very complex inference system. (Smolensky, 1987, p. 23)

---

11. The notion of a schema referred to here is the one introduced in cognitive psychology by Rumelhart (Rumelhart, 1975). For present purposes, Rumelhart's schemata are equivalent to Minsky's frames (Minsky, 1975). Schank and Abelson's scripts (Schank and Abelson, 1977) are also similar, but not applicable to the case at hand.

In discussing the differences between conventional approaches to schemata and the present network, Rumelhart et al. stress the flexibility of the network. In a conventional schema-processing algorithm, for example, they note that a decision has to be made about "which aspects of a given schema are constant and which are variable" and "which aspects of the situation are part of the schema and which are not." (Rumelhart et al., 1986, p. 37) In the network, such decisions needn't be made. A feature's being part of a schema consists only in its being strongly linked to many other of the features in some group. The question of how strong these links must be for the feature to be part of the schema is irrelevant to the functioning of the network.

Rumelhart et al. consider the flexibility of the PDP approach to schemata an advantage of the approach. They also use it to illustrate their view of the relation between symbol-processing and PDP models. In a way, the network operates as if it had a schema in the conventional sense. However, the differences show up if the network is given unusual inputs. For example, given the input, bed, sofa, and ceiling, the network seems to "blend" the bedroom schema with the living-room schema. (Rumelhart et al., 1986, p. 34) This illustrates the general point that "we can often run up against phenomena in which our high-level descriptions will not do, we must describe the system in terms of the underlying processes to understand its behavior." (Rumelhart et al., 1986, p. 56)

The flexibility that Rumelhart et al. stress follows largely from the use of a simple pattern associator. Not only does the network not make a decision about when a feature will be part of a schema and when it will not, it isn't even possible for such a decision to be reflected in the network. Because the network only contains nodes for individual features,

only connections between a feature and individual features can be encoded. Similarly, it would not be possible for the invoking of a schema itself to be a distinct event within the network. Other aspects of a conventional symbol-processing model of schema-application are also unlikely to be reflected in a simple pattern associator. For example, because a pattern associator encodes almost all of the statistical associations available (due to the learning rule), the sources of input to any given feature will be numerous. When the activation level of a particular feature will begin to climb will, for this reason, vary a great deal from one pattern completion to another. A simple pattern associator is, therefore, unlikely to reflect any particular algorithm for filling the slots in a schema or frame sequentially.

As a general rule, then, it would seem unlikely that a simple pattern associator could be regarded as implementing a symbol-processing algorithm. Since all the nodes are dedicated either to input features or to output features, no nodes are available to represent higher-level structures. And, since the learning rule for adjusting weights encourages all observed statistical associations to be reflected in the connection weights, the manner in which an output pattern emerges in a pattern associator will not likely reflect a particular algorithm. If we assume that the connectionist models that are to account for human cognitive abilities are simple pattern associators, eliminative connectionism is certainly an attractive view. I now want to show that the same can be said of a slightly more complicated pattern associator, the kind studied by Smolensky in his work on "harmony theory."

3.4 Harmonium networks

The networks Smolensky studied in harmony theory[12] are auto-associators with one layer of hidden units and a differential activation equation. As with many auto-associators, the task confronting a harmonium network[13] is pattern completion. A stimulus pattern of activity is encoded in the input/output layer of nodes. The pattern can be interpreted as a list of features, or, as Smolensky prefers, "microfeatures", which provides a partial description of an object or situation. The goal of the network is to complete the description by filling out the pattern of activity in the input/output layer of nodes.

Smolensky calls the nodes in the hidden layer "knowledge atoms." A given knowledge atom has a positive connection to some of the microfeatures, a negative connection to others, and no connection to the rest. Harmonium networks are unusual in that there are no intra-layer connections, only connections between nodes in the hidden layer--knowledge atoms--and nodes in the input/output layer--microfeatures. The connections between a given knowledge atom and the microfeatures define three subsets of the input/output layer--the microfeatures that are positively connected to the knowledge atom, those that are negatively connected, and those that are unconnected. Potentially, there is a different knowledge atom for every possible three-way partition of the set of microfeatures.

The potential number of knowledge atoms is enormous for input layers of even modest size--3^N where N is the number of microfeatures. Not all of

---

12. Harmony theory is presented in Smolensky, 1986.

13. The term is Smolensky's.

the potential knowledge atoms will be used. But, for theoretical reasons, it is likely that a substantial portion of them will be. Smolensky is particularly interested in using probability theory to study the capabilities of harmonium networks. The theorems Smolensky proves in Smolensky, 1986 concern how the network should and will complete a particular pattern given its exposure to a variety of patterns in "the environment." It is assumed that a pattern in the environment consists of the presence of certain microfeatures and the explicit absence of others. It is further assumed that the network contains a distinct knowledge atom for every such pattern "observed" by the network. (Smolensky, 1986, pp. 226-229)

It is possible for a number of patterns to be observed on a single occasion. Suppose, for example, the environment presents a pattern that includes six features, with four others explicitly excluded and two features neither included nor excluded. The pattern would contain as subpatterns six different patterns that include five features, exclude four, and neither include nor exclude three others. Any or all of these subpatterns could be observed depending on how observation is supposed to work for the network in question.[14]

All of these technical details are suggestive of a very large number of knowledge atoms, and this is clearly what Smolensky has in mind. Consider, for example, the following passage discussing how a harmonium network would handle letter-perception:

---

14. The possibility of observing subpatterns is clearly suggested when Smolensky considers having some of the knowledge atoms in a word-recognition network be digraph units, e.g. W in position 1 together with A in position 2. (Smolensky, 1986, p. 204) The possibility also seems to be in the general spirit of the discussion on pp. 201-208.

> In harmony theory, the idea is that there would be a set of representation nodes [nodes in the input/output layer] dedicated to the representation of the presence of letters independent of their shapes, sizes, orientations, and so forth. There would also be a set of representations for graphical features, and for each letter there would be a multitude of knowledge atoms, each relating a particular configuration of graphical features with the representation of that letter [Smolensky's emphasis]. (Smolensky, 1986, pp. 217-218)

The fact that harmonium networks contain very large numbers of knowledge atoms is important. Higher-level structures in a harmonium network do not correspond either to individual nodes in the input/output layer or to individual knowledge atoms. Rather, they correspond to large groups of knowledge atoms. It is possible that a given higher-level structure will be best represented by one particular knowledge atom. Suppose, for example, that we were to construct a harmonium network for the room-schemata example. There is a potential knowledge atom corresponding to the prototypical bedroom. The knowledge atom would have positive connections to all those features that are contained in the prototypical bathroom. There might be some features to which it had no connections, and it would have positive connections to the rest. Perhaps this prototypical knowledge atom would be included in our harmonium network. But, it needn't be. As long as many knowledge atoms which have considerable overlap with the prototypical knowledge atom are present in the network and are assigned sufficient strength, the prototypical knowledge atom itself need not be present. And, even if it is present, its role is not really distinguishable from that of many similar knowledge atoms.

Learning in the harmonium network will consist of particular knowledge atoms acquiring greater strength as they are observed in the environment more and more frequently. Since a pairing of even two features constitutes a knowledge atom, once again all statistical associations between features

may be learned and then exert an influence on the emergence of an output pattern during pattern completion. As with simple pattern associators, then, we would not expect a harmonium network to implement a sequential symbol-processing algorithm.

3.5 Summary

Harmonium networks differ from simple pattern associators in that a layer of knowledge atoms is added. Each knowledge atom is associated with a set of input features that it includes and a set that it excludes. Harmonium networks are like simple pattern associators in that the activity in the network is very diffuse. There are a very large number of knowledge atoms, many of which differ from others by a very small degree. The learning procedure for harmonium networks, like the learning procedure for simple pattern associators, results in all statistical associations being learned reflected in connection weights; in this case, all statistical associations between a given feature and a set of features are reflected in the weights of connections between an input feature node and a knowledge atom. When an output pattern emerges in a harmonium network during pattern completion, just as in a simple pattern associator, a great many nodes will be involved. Activity is spread over a great number of similar knowledge atoms.

In sum, representation, learning and subsequent processing are diffuse in harmonium networks, just as in simple pattern associators. I now want to coin a term that will cover both simple pattern associators and Smolensky's harmonium network. To emphasize the diffuseness of representation and processing, I will say that both are unstructured

pattern associators.[15] I won't try to delineate such a class of pattern associators with any precision. What is important is: (a) to have some term that covers both simple pattern associators and the pattern associators studied in harmony theory; and (b) to have a rough idea of what they have in common, so that when the time comes we'll have no trouble recognizing that certain models are importantly different from these.

What we have seen in this section is that unstructured pattern associators are not likely candidates to be implementations of symbol-processing algorithms. Therefore, if it is assumed that connectionist models of cognitive processing typically consist of a single unstructured pattern associator, or even multiple unstructured pattern associators operating independently of one another, eliminative connectionism is extremely plausible. In other words, someone who thinks of cognition as unstructured pattern association ought to be an eliminative connectionist. In the rest of the paper I will argue that, in fact, Smolensky, and to a lesser extent McClelland and Rumelhart, do hold to a view of cognition as unstructured pattern association (I will sometimes refer to this view simply as the view of "cognition as pattern association"). I will argue further that this view is overly simple, and that a more realistic view of how connectionist models might account for cognitive abilities, is not nearly so conducive to eliminative connectionism.

---

15. Recall that harmonium networks are pattern associators with a single hidden layer (the layer of knowledge atoms).

4. Pattern Association and Eliminative Connectionism

4.1 Pattern Association and Smolensky's Eliminative Connectionism

We saw in section 2 that Smolensky is an eliminative connectionist. Symbol-processing accounts cannot, he insists, precisely describe the activity of the intuitive processor. But, why does he believe this? I want now to suggest the following answer. Smolensky conceives of the intuitive processor as a single pattern associator of the sort studied in harmony theory (Smolensky, 1986). And, such networks, as we saw in section 3, are, like simple pattern associators, particularly conducive to eliminative connectionism. In other words, Smolensky advocates eliminative connectionism because he is implicitly wedded to the view of cognition as unstructured pattern association.

To see how Smolensky's eliminative connectionism is rooted in a view of cognition as pattern association, let's begin by looking at the details of the intuitive processor. Its important to realize that just because the intuitive processor is thought to have a connectionist architecture does not mean that the intuitive processor is a single connectionist network. It could very well be a number of connectionist networks linked together. This might raise the question of why it should then be thought of as a single processor. But, there might be any number of reasons for this. Perhaps, for example, there is a single network in which the ultimate output of the processing of all the other networks is represented. This would be a reason for thinking of the networks as forming a larger whole. Nevertheless, although the intuitive processor could be thought of as a set of interlinked connectionist networks, it is pretty clear that Smolensky

thinks of it as a single network.[16] Immediately after stating that the intuitive processor has a connectionist architecture, he writes,

> The kind of connectionist model I will consider can be described as a network of very simple processors, units, each possessing a numerical activation value that is dynamically determined by the values of the other processors in the network. The activation equation governing this interaction has numerical parameters which determine the direction and magnitude of the influence of one activation value on another; these parameters are called connection strengths or weights. The activation equation is a differential equation .... A network is sometimes programmed by the modeler, but often a network programs itself to perform a task by changing its weights in response to examples of input/output pairs for the task. The learning rule is the differential equation governing the weignt changes. (Smolensky, 1987, p. 5.)

All of these assumptions generate "the connectionist dynamical system hypothesis":

> The state of the intuitive processor at any moment is precisely defined by a vector of numerical values (one for each unit). The dynamics of the intuitive processor is governed by a differential equation. The numerical parameters in this equation constitute the processor's program or knowledge. These parameters may change according to a learning equation. (Smolensky, 1987, p. 5)

If the intuitive processor were a set of networks, it would make more sense to say that the state of the processor at any moment is defined by a set of vectors, one for each network. It would also be more natural to regard each network as having its own activation equation.[17] Similarly with the learning equation. That Smolensky thinks of the state of the processor as

---

16. It should be emphasized that the criticisms I will go on to level against the view of cognition as pattern association apply both to (a) views that envision single pattern associators as models of cognitive processes; and (b) views that envision multiple pattern associators operating independently as models of cognitive processes. Smolensky's view, however, seems to be of the former variety.

17. I only say more natural because, if we assume that the connections between the networks are of the standard connectionist sort, and if the general principles governing activation were the same in each network, then it would be possible to regard the set of networks as a single network with a single activation equation.

being defined by a single vector, and his assumption without argument or comment that the rules governing update of activation and modification of connection strengths should be expressed in a single activation equation and a single learning equation, make it evident that Smolensky thinks of the intuitive processor as a single connectionist network.

The passages above also serve to make another point. Smolensky assumes that the activation equation will be a differential equation.[18] This assumption is borrowed from his work on harmony theory, in which he studied connectionist networks governed by a differential activation equation. Those networks also happened to be auto-associators with a single layer of hidden units. Smolensky is evidently thinking of the intuitive processor as a single network of the kind he studied in harmony theory.

There is yet another manner in which the intuitive processor could easily have been complicated. The ultimate output of the intuitive processor could be the outcome of several processing cycles. By a processing cycle, I mean the period of time which begins when a network receives an input and ends when the pattern of activity of the network has stabilized. In many systems the activation equation changes over the course of such a processing cycle in accordance with a measurement called "temperature." In the beginning of the cycle the temperature of the system

---

18. The significance of a differential activation equation is that a differential equation is continuous. So, in the Smolensky's networks, activation levels are updated continuously, not just at discrete units of time. This feature sets up a parallel between the activity of such networks and the activity of thermodynamic systems, a parallel that Smolensky makes much of in Smolensky, 1987. For our purposes, the important parallel is the one between the networks that appear in harmony theory and the intuitive processor.

is high. This means that there is a large random component in the assignment of activation levels. The temperature cools over the course of the processing cycle. The intuitive processor could work as follows. It begins with an initial input. The temperature of the system is raised and then lowered, leading it to settle into a particular pattern of activity. This pattern then becomes the input for the next processing cycle, during which the temperature is again raised and then lowered. Eventually, after some number of processing cycles, the network settles into a final pattern of activity, constituting its ultimate output. Smolensky specifically proposes something like the above scenario as an account of conscious rule application. Among the intermediate patterns of activity are the explicit rules being followed. (More on this later.) However, Smolensky also specifically assumes that this is not how the intuitive processor functions. The contrast is described in the following passage:

> Using the stored rules the network can perform the task. the standard learning procedures of connectionist models turn this experience performing the task into a set of weights for going from inputs to outputs. Eventually, after enough experience, the task can be performed directly by these weights. The input activity generates the output activity so quickly that before the relatively slow interpretation process has a chance to reinstantiate the first rule and carry it out, the task is done. (Smolensky, 1987, p. 12)

In sum, Smolensky assumes that the intuitive processor is a single connectionist network, with a differential activation equation that produces a response to an input in a single processing cycle. He all but explicitly assumes that it is a pattern associator of the kind he studied in harmony theory.

That Smolensky regards the intuitive processor as an unstructured pattern associator can further be seen in the concluding section of the paper, where he illustrates the relation between connectionist accounts of

the intuitive processor and symbolic accounts at the conceptual level (Smolensky, 1987, pp. 21-23). Smolensky discusses two specific examples in this section. One example is taken from his work on harmony theory, and is, therefore, a harmonium network. The other example is the room-schemata example, which is, as we have seen, a simple auto-associator.

4.2 Pattern Association and Rumelhart and McClelland's Connectionism

Both the detailed assumptions that he makes about the intuitive processor--which, one must keep in mind, is supposed to be responsible for the vast majority of cognition--and the examples he uses to illustrate the relation between connectionist and symbol-processing models, show that Smolensky, in "On tne Proper Treatment of Connectionism," relies on a view of cognition as unstructured pattern association. Rumelhart and McClelland, in "PDP Models and General Issues in Cognitive Science," do not give as elaborate a picture of cognition as Smolensky does. It would be unfair to charge them with the view of cognition as pattern association. Still, their view of the relation between connectionist and symbol-processing models is heavily influenced by the pattern association paradigm. They write:

> The basic perspective of this book [Parallel Distributed Processing] is that many of the constructs of macrolevel descriptions such as schemata, prototypes, rules, productions, etc. can be viewed as emerging out of interactions of the microstructure of distributed models. These point are most explicitly considered in Chapters 6, 14, 17 and 18. (Rumelhart And McClelland, 1986a, p. 125.)

Chapter 6 is Smolensky's essay on harmony theory. Chapter 14 is Rumelhart et al., 1986, the central example of which is the room-schemata model described above--a simple auto-associator. In chapter 17, McClelland and Rumelhart present a PDP model of human learning and memory. Each of the

specific models described in the chapter are simple auto-associators.[19] Finally, chapter 18 presents Rumelhart and McClelland's model of the acquisition of the past tense. The model is essentially a simple pattern associator. In some simulations the pattern associator is augmented by a secondary output layer, but this additional layer is not fundamental to the model.[20] It turns out that all of the models that illustrate Rumelhart and McClelland's (Rumelhart and McClelland, 1986a) basic perspective on the relation between connectionist and symbol-processing models are unstructured pattern associators. If, as I shall argue, unstructured pattern associators are too simple to serve as a general paradigm for cognition, then the foundations of both Smolensky's and Rumelhart and McClelland's eliminative connectionism will be shaken.

5. Problems with the View of Cognition as Pattern Association

We have seen that in espousing eliminative connectionism, Rumelhart and McClelland (Rumelhart and McClelland, 1986a) to some extent, and Smolensky to a greater extent, have relied on a view of cognition as unstructured pattern association. My aim in this section is to cast doubt

---

19. McClelland and Rumelhart actually contemplate a model of human memory consisting of a large number of interconnected modules, each one of which would be a simple auto-associator (McClelland and Rumelhart, 1986, p. 174). They also contemplate the possibility of augmenting the networks with hidden units (McClelland and Rumelhart, 1986, pp. 209-14). However, the PDP models that are presented in the chapter, and whose performance is compared with human performance on a variety of learning tasks, are all individual modules, and, hence, consist of a single simple auto-associator.

20. Rumelhart and McClelland write, "All learning occurs in the pattern associator; the decoding network is simply a mechanism for converting a featural representation which may be a near miss to any phonological pattern into a legitimate phonological representation. Our primary focus here is on the pattern associator." (Rumelhart and McClelland, 1986b, p. 223

on this view and thereby cast doubt on the eliminative connectionism of these authors. More particularly, I will claim that a careful look at the connectionist research program itself, even that branch of it represented in Parallel Distributed Processing, suggests that unstructured pattern association may not be adequate as a paradigm for the study of cognition.

I will begin in section 5.1 by describing three ways in which structure is sometimes added to PDP models. As I present each of these features I will discuss how the structure in these models make it more likely that they will be implementations of symbol-processing algorithms. In section 5.2 I discuss the importance of these features and reasons for thinking that they or other similar features ought to be incorporated into PDP models that are adequate to account for human cognitive abilities.[21]

It should be emphasized that the discussions in sections 5.1 and 5.2 are supposed to have a cumulative effect. I will discuss three very different ways of adding structure to PDP models, arguing in each case that there is reason to consider complicating PDP models in this way and that doing so makes them much more likely to end up being implementations of sequential symbol-processing algorithms. I do not claim for any of the features that PDP models will absolutely have to have the feature or that models with that feature will necessarily implement conventional models. Nor do I even attempt to show that models with any of the particular

---

21. The ensuing discussion may leave the unknowing reader with a picture of connectionist research according to which simple pattern association is the most basic, fundamental model and other models can be derived from simple pattern association with the addition of the relevant features. This is not an accurate picture. Many connectionist models, particular non-distributed models, are neither simple pattern associators nor variations on that theme. McClelland and Rumelhart's interactive model of word recognition is a prominent example (McClelland and Rumelhart, 1981).

features in question will implement full-blown symbol-processing algorithms. Rather, I show in each case that models with these features are likely to exhibit substantial elements of sequential symbol-processing algorithms. Although my claims for each feature are, therefore, not particularly strong, their cumulative effect, particularly when one considers that the features could be combined, should be to demonstrate that there are a great many ways in which PDP models could be developed, and, perhaps need to be developed, so that they will come to implement symbol-processing models. Hence, it will be seen, one should be extremely wary of assuming, in our current state of ignorance about what fully adequate PDP models of cognitive processes might look like, that they will not in any way implement symbol-processing algorithms.

5.1 Ways of Adding Structure to PDP Models

Smaller Numbers of Special Purpose Hidden Units. The possibility of using hidden units in pattern associators is important because simple pattern associators are subject to the limits on certain perceptron-like devices discovered by Minsky and Papert(Minsky and Papert, 1969). Perceptron-like devices (pattern associators being such devices) without hidden units cannot compute certain functions, such as the logical function of exclusive disjunction. Smolensky, we saw, uses very large numbers of hidden units arranged in a single layer. Also, in his harmonium networks, the connections between a given hidden unit, or knowledge atom, and features in the input/output layer, and, therefore, the significance of the unit, is fixed in advance. Other models have not insisted on a large number of hidden units. They have also formulated methods by which connections involving hidden units can be learned. In such models, the

hidden units often come to play very particular roles in processing. This is in contrast to harmonium networks where the role played by any given knowledge atom is very similar to the role played by many others. Rumelhart and Zipser, 1986; Rumelhart, Hinton and Williams, 1986; and, Ackley, Hinton and Sejnowski, 1985 provide examples of such models.[22]

Although I believe that similar sorts of remarks would apply to other proposals for small numbers of special-purpose hidden units, I want now to focus on one in particular. Specifically, I want to argue that the networks described in Rumelhart and Zipser, 1986 are likely to contain elements of sequential processing. The networks Rumelhart and Zipser studied were not pattern associators. Rather, they were networks of the kind called "regularity detectors." A regularity detector contains an input layer and some hidden units. As with pattern associators, a single input is a pattern of activity over the whole input layer. The network receives a range of inputs during the course of the training period, each with a particular frequency. The system is supposed to discover and come to represent regularities in the input population.[23] In Rumelhart and Zipser's networks, the hidden units are organized into a hierarchy of layers. Within each layer units are organized into clusters. Units within

---

22. Minsky and Papert, incidentally, were dubious about perceptrons because, although their computational limitations could be overcome by adding hidden units, it wasn't clear that adequate learning rules for the hidden units would be forthcoming. Hence, because of the existence of learning rules such as the ones in the references cited above, their reasons for being dubious about perceptrons have largely been overcome. See the Epilogue to the expanded edition of Perceptrons (Minsky and Papert, 1988) for a discussion of this issue.

23. It isn't hard to imagine a regularity detector being merged with a pattern associator so that regularities in the inputs were used to assist in producing the appropriate response to a given input.

a cluster "compete" with each other in the learning process so that different units end up responding to distinct features of the inputs.

In the particular example I want to describe in some detail, there are two layers of hidden units. The first layer contains two clusters, each containing four units. The second layer contains one cluster of two units. The input layer contains 72 units. The nodes in the input layer can be thought of as forming two 6-by-6 square grids. The grid on the left serves as a "teaching" grid and is eventually dropped, so that the input eventually occurs only in the grid on the right. The network will receive a large number of inputs (the input population) in the grid on the right, and the goal of the network is to come to respond to regularities in the input population. Since the final cluster of hidden nodes contains only two nodes, the network will tend to classify the input population into two sorts, with one node lighting up when inputs of the first sort are produced, and the other node lighting up when inputs of the second sort are produced. In the input population for which the network was tested, all the inputs were either vertical or horizontal lines in the 6-by-6 grid. Hence, if the network is successful, one of the top two nodes should come to respond to vertical lines and the other to horizontal lines.

When the network was tested on this input population, each of the units in the first layer learned to respond either to half of the vertical lines or to half of the horizontal lines (in the grid on the right). So, four of the units in the first layer only respond to vertical lines, and the other four units only respond to horizontal lines (though no one node at this layer will respond if and only if a vertical/horizontal line is present). One of the units in the top layer then learned to respond when and only when <u>two</u> of the former units became active. The other unit in the

top layer learned to respond when and only when two of the latter units in the first hidden layer became active. Hence, one of the units in the top layer came to respond when and only when the input was a horizontal line, and the other unit came to respond when and only when the input was a vertical line.

Clearly, the network, which is depicted Figure 2, can be thought of as implementing a particular three-step algorithm for determining whether the activation of some nodes in a 6-by-6 grid of nodes is either a vertical or a horizontal line. Each of the first two steps, which are taken in parallel, consist of determining whether the input falls into any one of four categories. The third step uses the results of the first two steps to classify the input as a horizontal or vertical line (or neither).

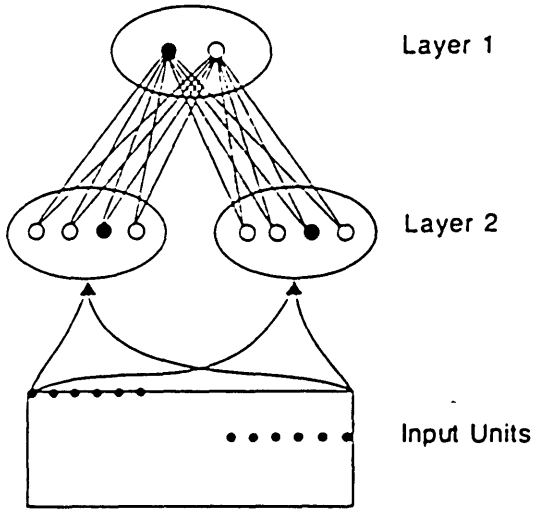

Figure 2. Reproduced from Rumelhart and Zipser, 1986.

The fact that the network can be seen to implement an algorithm is more important than the sophistication of the achievement. Still it is worth noting that the achievement is more notable than it might first appear. The architecture of the network is indifferent to how the input nodes are arranged. It would have been just as accurate to picture them in a single row or in a differently arranged 6-by-6 grid, so that the inputs could not have been thought of as vertical or horizontal lines. The network would still have worked in the very same way, and thereby learned an algorithm for classifying the inputs in an important way.

The network in the example above comes to implement a simple sequential algorithm for recognizing vertical/horizontal lines because of general features of Rumelhart and Zipser's proposal. Hidden units are organized on this proposal into a hierarchy of clusters. The arrangement of the layers together with the associated learning rules have the consequence that any given cluster of hidden nodes comes to detect regularities in the patterns in the layer below it. It is quite natural that such networks would come to implement sequential algorithms for classifying the stimulus encoded in the input layers, with the hidden nodes in the higher layers detecting regularities of greater scope (i.e., applying to more of the nodes in the input layer) and greater abstraction. The greater the number of layers, or the greater the number of clusters within each layer, the greater is the potential for such a network to implement a complex algorithm.

Modular Structure. Connectionist models may be organized into a number of distinct networks or modules.[24] One can imagine, for example, a model consisting of two pattern associators which have been connected so that some or all of the output nodes of one pattern associator are connected to the input nodes of the other. In this manner, the output of the first pattern associator could become the input to the second. McClelland and Rumelhart (McClelland and Rumelhart, 1986) propose a model of human learning and memory which consists of a large number of interconnected networks or modules. "We assume," they write, "that the units are organized into modules. Each module receives inputs from other modules; the units within the module are richly interconnected with each other; and they send outputs to other modules." (McClelland and Rumelhart, 1986, p. 174.) They suppose that a complete system of memory would contain "many hundreds or perhaps many thousands" of modules.

The significance of a modular structure is clear. In general, if a connectionist attempts to model human performance of a given task and proposes a modular structure, each network will be thought of as being responsible for a distinct component of the task. Suppose that the connections are arranged so that processing in any given network does not begin until the network which supplies its input has stabilized. And, suppose that each of the components of the task corresponded to a step in a symbol-processing algorithm. Then, it is easy to imagine the model implementing a sequential (or only moderately parallel) symbol-processing

---

24. The modules under discussion here are not assumed to be modules in Fodor's (Fodor, 1983) sense. For example, there is no assumption of informational encapsulation. When I talk about modular structure, I simply mean a PDP model consisting of a number of distinct, interconnected networks.

algorithm. The restriction that the modules supplying the input to a given module stabilize before the given module begins processing could be relaxed. We would only have to require that the end result in the given module is the same as it would be if it received its inputs all at once, rather than over a period of time. We could relax the assumption still further if any deviations were incidental enough to regard as effects of the implementation. Whether the interactions among modules in such PDP models will generally turn out to satisfy any version of these assumptions is a matter of speculation. But, it is certainly plausible that they will.

Multiple annealings. The final complicating feature of PDP models that I want to bring up is one discussed at length by Smolensky himself. He proposes it to account for conscious rule application, that aspect of cognition that does not involve the intuitive processor. Smolensky supposes that a particular pattern associator contains units for a large number of features pertaining to some problem. Patterns of activity over these features are able to represent both descriptions of the current state of the problem (state-descriptions) and linguistic descriptions of rules pertaining to the problem. In conscious rule application the problem is solved as follows. The statement of the problem provides the initial input to the pattern associator. The pattern associator goes through a single annealing process. This annealing is not sufficient to generate a pattern of activity representing a solution to the problem. Instead the pattern associator generates a rule. In particular, it generates a rule whose antecedent is satisfied by the stimulus state-description. Suppose, for example, that the problem being solved is of the sort found in books of "brain teasers," where a group of people must be seated around a table subject to various constraints. Then, the stimulus state-description might

be that Harry is sitting next to Shirley. And, the rule generated might say that if Harry is sitting next to Shirley, then neither Bill nor Beth is sitting next to Harry.

Once a rule has been generated, it becomes the input to the pattern associator which then goes through another annealing process. The output this time is a representation of the situation that execution of the rule would produce. In the example above, the second annealing process generates a state-description that includes the information that neither Bill nor Beth is sitting next to Harry. This new situation becomes the input for another stage of processing, which generates a new rule. The network thus produces an alternating sequence of state-descriptions and rules such that each rule applies to the state-description that precedes it and each state-description is the consequence of applying the preceding rule to the preceding state-description.

Eventually, after enough rules have been generated and executed, a solution to the problem will be generated. Smolensky proposes that conscious rule application is realized in connectionist networks through multiple annealings of the same network in this way. After enough experience, he supposes, connection weights will be established which will make it possible to generate the pattern of activity representing the solution to the puzzle in a single annealing.

A network of the sort just described (at the point when it still requires multiple annealings to generate solutions) obviously implements a sequential algorithm in order to solve the problems it is given. In fact,

it can be seen to implement a production system.[25] During the stages in which the ultimate output of the network is a rule, the network can be regarded as identifying a production whose antecedent matches the current situation. During the stages in which the ultimate output is a description of the situation resulting from execution of the rule, the network can be regarded as executing the production. In general, when a PDP model involves multiple annealings, the model can be seen as implementing a serial process.

5.2

I have described three different features of PDP models, each of which makes them more structured than unstructured pattern associators. In each case we saw that the additional structure may result in the models implementing sequential algorithms. If these features are combined, of course, the possibilities of implementing standard symbol-processing algorithms are only increased. The obvious question at this point is, will adequate PDP models of human cognitive abilities (if such are forthcoming) have some or all of these features? If there is reason for thinking that they will, then the versions of eliminative connectionism that we have been discussing will be in trouble. I want next to argue that indeed there are reasons to believe that some or all of the above features will be present in adequate PDP models of human cognitive abilities. But, first I want to emphasize that the burden of proof here is really on Smolensky and on Rumelhart and McClelland. Once it is clear that their ideological positions depend on a very particular sort of PDP model, the burden of

---

25. See J. R. Anderson, 1983, chapter 1 for a general discussion of production systems.

proof, it seems to me, is on them to give reasons for thinking that adequate PDP models of human cognitive abilities will be models of that sort. Still, I shall now give some reasons for thinking that they won't.

Small numbers of special-purpose hidden units. The importance of having some hidden units is clear. As I noted earlier, without any hidden units, pattern associators are subject to the limits on perceptrons discovered by Minsky and Papert. Smolensky's harmonium networks have hidden units, but present a different problem. Recall that in a harmonium network, the network comes with a full array of knowledge atoms in place. And, the connections between knowledge atoms and the input/output layer are fixed in advance. All that is learned is the strength of the knowledge atom. The obvious question, then, is how do the knowledge atoms, complete with connection weights, get put in the network. An obvious suggestion would be to have a mechanism for automatically "growing" all the possible knowledge atoms for a given input/output layer, and perhaps eliminating ones that are never used. However, as I noted earlier, the number of possible knowledge atoms is 3^N for an input/output layer of size N. The number of possible knowledge atoms is too large for this suggestion to be biologically plausible even for an input/output layer with 40 microfeatures: the number of nodes required would exceed the number of neurons in the brain by several orders of magnitude. Models with a small number of hidden units in which the strengths of a given units' connections are learned do not have this problem. Installing a few hidden units which obey the appropriate learning rule is easy enough. The learning rule takes care of everything else. In raising this question of how harmonium networks acquire their knowledge atoms, I do not mean to suggest that it is

a fatal problem. Still it does provide one reason for thinking that the other models may be preferable.

Modular structure. The fundamental point I want to raise here is a simple one: PDP models are typically incomplete in many ways (not unlike most models in cognitive psychology). Many aspects of the task in question are assumed to be handled elsewhere, i.e. somewhere other than the network being described. Within a PDP framework, "elsewhere" could mean one of two things, either in an expanded version of the network being described, or in other networks connected in appropriate ways with the one being described. The latter option is, I believe, often more natural and leads to the supposition that the network being described is part of a larger modular structure. I'll now illustrate these points with an example, Rumelhart and McClelland's model of the acquisition of the past tense (Rumelhart and McClelland, 1986b). Much of the discussion relies on Pinker and Prince, 1987, in which the model and some of its shortcomings are discussed at length.

For the sake of this discussion, all we need to know about Rumelhart and McClelland's model is: (1) that it is a simple pattern associator; (2) that the input layer encodes present tense forms of verbs and the output layer encodes past tense forms; (3) that the microfeatures in the layers are purely phonological; and (4) that the model learns by being repeatedly exposed (by some external entity) to present tense forms followed by the correct associated past tense form. Once the model has gone through a period of learning it is supposed to be able to produce the correct past tense form in its output layer when a given present tense form is encoded in its input layer.

The first thing to notice is that in actual practice the network will be conjoined with a mechanism or mechanisms which need to access the past tense form of a given verb in completing some task and use the network for this purpose. One might think that these mechanisms are peripheral to the actual acquisition of the past tense, that all the learning takes place inside the network. But, this isn't so. In some cases, for example, the output of the network will have to be interpreted in a fairly sophisticated way. For some people, myself included, the word "leap" has two more or less equally acceptable past tense forms, "leapt" and "leaped." My answer to the question, what is the past tense of "leap"?--"Well, it really has two past tense forms ..."--can to some extent be accounted for by this PDP model. Perhaps I have a network of the sort proposed and it produces two stable patterns in its output layer given "leap" as input. Still, the fact that I say that "leap" has <u>two</u> past tenses, rather than saying that I'm not sure what its past tense is or that I don't remember must be accounted for. Evidently, part of my knowledge about the past tense of "leap" is contained outside the network itself, and resides in the mechanisms that interpret it.

There also must be fairly sophisticated knowledge of the past tense in the mechanisms that access the network. Pinker and Prince note, for example, that when a verb is derived from a noun, the regular rule for past tense formation is used, citing examples such as[26]:

(a) He flied out to the centerfielder. (not flew)

---

26. Pinker and Prince, 1988, p. 111.

(b) The invading army ringed the city with troops. (not rang)

It appears that in the production of such sentences the network is bypassed, and some mechanism that generates the regular form used. The knowledge that this is to be done must be built into the mechanisms that would normally access the network.

The mechanisms that are responsible for "teaching" the network will be another locus of important activity in the acquisition of the past tense. Presumably teaching occurs as the child comprehends sentences and her language comprehension faculties figure out, using syntactic and semantic clues, that some of the phonological forms being processed must be past tense forms of familiar verbs. Thus, there is a subtle interaction between the mechanisms involved in parsing sentences and constructing their meaning, and the past tense network. The interaction will presumably go both ways: an association between a present tense and a past tense form, even when it is weak, will probably sometimes be one of the clues in determining that a given form is the past tense of a certain verb; the association will then, of course, be reinforced.

Other likely cases of important interactions between the past tense network and mechanisms external to it--other networks if the model is to fit into a PDP framework--could be cited. There is likely, for example, to be some interaction between the past tense network and networks responsible for acquiring related forms such as passives, past participles and verbal adjectives, all of which are governed by the same rules for regular verbs.[27] In sum, a complete account of the acquisition of the past tense

---

27. See Pinker and Prince, 1988, p. 102.

will involve much more than a single network. It must involve a large number of networks interacting in complex ways. And, the locus of the explanation of the past tense will involve the other networks crucially. This claim is not intended as a criticism of Rumelhart and McClelland's model of the acquisition of the past tense. Rather, it is a reminder that their model is a deliberate simplification. Careful analysis of other PDP models would, I believe, lead to similar conclusions. Although a single network is typically posited to account for cognitive abilities, ultimately these networks will generally have to be embedded in larger structures in order to become complete accounts of the ability in questions. Hence, there is reason to believe that ultimately many PDP models of cognitive abilities may have a modular structure.

Multiple annealings. I want, finally, to discuss multiple annealings. The thrust of my remarks here will be that if something like pattern association is fundamental to cognition, then, something like multiple annealings will likely be used in many PDP models of human cognitive abilities. Before arguing this point, however, I want to recall Smolensky's proposal for implementing conscious rule application in a PDP architecture through multiple annealings. The sort of rules Smolensky has in mind are, of course, rules that are expressed in production systems as condition-action pairs, or productions. The proposal, very roughly, is that conscious rule application can be modeled by an auto-associator if (a) states of the network in which a condition obtains, cause the network to settle into a pattern of activity in which a rule with that condition is represented, and (b) states of the network representing a given rule cause the network to then settle into a pattern of activity representing a new situation in which the action sequence of the rule has been executed.

After presenting this proposal, Smolensky suggests that such networks will eventually be able to perform tasks directly without going through multiple annealings.

> Using the stored rules the network can perform the task. The standard learning procedures of connectionist models turn this experience performing the task into a set of weights for going from inputs to outputs. Eventually, after enough experience, the task can be performed directly by these weights. The input activity generates the output activity so quickly that before the relatively slow interpretation process has a chance to reinstantiate the first rule and carry it out, the task is done. (Smolensky, 1987, p. 12.)

The suggestion in this passage, which offers an account of the novice-expert shift, might be used to argue for the relative importance of PDP models.[28] Early in Smolensky's paper, a dichotomy is set up between conscious rule application and the operations of the intuitive processor. Symbol processing models may provide satisfying accounts of conscious rule application. But, he wants to argue, they don't provide satisfying accounts of other kinds of cognition, namely the operation of the intuitive processor, which is best explained in PDP terms. The proposal that conscious rule application rapidly becomes intuitive suggests that conscious rule application will account for a relatively small percentage of cognitive activity: it can't be prevalent in any domain for long without becoming an intuitive process. Connectionist models, then, will account for far more of cognition than will symbol processing models.

There is a major flaw in this argument. Let us accept Smolensky's account of the novice-expert shift. And, let us assume that conscious rule application, with experience, is rapidly transformed into the expert's

---

28. I am not claiming that Smolensky makes the following argument, though it seems to me that he may.

ability to perceive the answer to a problem directly. Still, much of cognition will not be accounted for by pattern associators completing a task in a single annealing. Even for an expert in a given domain, a great number of cognitive tasks can only be completed over an extended period of time. Rumelhart et al. explicitly assume that their networks take on the order of a few hundred milliseconds to settle into a stable pattern of activity.[29] Smolensky does not make this explicit assumption, but he does assume, following Rumelhart et al., that the contents of consciousness are the stable patterns of activity of the intuitive processor.[30] Presumably, when a cognitive task is performed over a number of minutes, the contents of consciousness change a number of times. We can infer, then, that in such a task the intuitive processor will achieve a sequence of stable states prior to completion of the task. And, this, of course, requires multiple annealings.

Such will be the case then with any task that requires extended deliberation. In the novice-expert shift, of course, what originally required deliberation no longer does. But, for even the most expert of experts many tasks will require substantial deliberation. The chess master may have learned to "perceive" checkmate without consciously verifying that each of her possible moves leaves the king vulnerable. But, she still takes several minutes to choose a single move in a serious game of chess. Such examples could, of course, be produced for any other expert we might choose. The physicist recognizes the solution to Physics 101 problems at a

---

29. See Rumelhart et al., 1986, p. 39.

30. See Smolensky, 1987, p. 12.

glance, but presumably will have occasion to deliberate at length when designing experiments to test the very latest theory in particle physics.

Extended deliberation will also be required in many tasks not associated with any sort of expert. The formulation of complex intentions and plans often requires extended deliberation. During such time consciousness is full of mental activity: subgoals are formulated, consequences of possible actions are evaluated, etc. The point here is not that tasks which require extensive deliberation of experts must be achieved by way of conscious rule application. I don't intend to be advocating any particular psychological account of these cognitive abilities. The point is simply that if Smolensky's intuitive processor or some other PDP network is to model extended deliberation, the model will have to settle into many different stable patterns of activities before the task is completed. In other words, extended deliberation can only be modeled by PDP networks which undergo multiple annealings.

6. Conclusion

Before reviewing and summarizing the arguments presented in this paper, I want discuss the polemical situation created by papers such as Rumelhart and McClelland, 1986a and more particularly by Smolensky, 1987. No one doubts that connectionist models are very different, both in spirit and detail, from the symbol-processing models, designed to be implemented in conventional computer architectures, that have dominated cognitive science. One can recognize these differences, however, and still maintain that there is an important place for both connectionist and symbol-processing models. Eliminative connectionists choose not to; instead they believe that connectionism should be taken as and advocated as an

alternative approach to the study of cognition whose ascendance will render some symbol-processing models obsolete, some superfluous, and others merely heuristic.

Neither McClelland and Rumelhart or Smolensky really offers an argument for their version of eliminative connectionism. There is a good reason for this. It is clearly possible to construct connectionist models that do implement symbol-processing models. It would be futile, therefore, to attempt to argue that a connectionist must be an eliminative connectionist. It is quite natural, then, that in advocating eliminative connectionism McClelland and Rumelhart and Smolensky focus on describing what they believe connectionist models should look like and the relation they envision between these models and traditional accounts in the relevant cognitive domains. The views, as a whole, are never really argued for, although parts of the views are.

Someone reading these essays, be they a connectionist or not, wondering whether to take connectionism as, or perhaps whether to pursue it as, a radical alternative to traditional models in cognitive science, cannot critically evaluate arguments given for the view that she should so take it; no such broad arguments have been provided. Of course, the reader can critically evaluate the arguments that are provided, but all of these are in the service of narrower claims. In this situation, it is perfectly reasonable for the critical reader to inquire into the sources of the authors' views about what connectionist models should look like and what their relation to other models ought to be. In particular, it is reasonable to wonder whether the views reflect a very narrow conception of a connectionist model and whether the views would make less sense if the conception were broadened. This is, of course, the tack that I have taken.

Revealing that these views appear to follow from a very narrow conception of a connectionist model is important. Many readers, connectionists and non-connectionists alike, would, I believe, take much more of a skeptical stance toward Smolensky's and Rumelhart and McClelland's versions of eliminative connectionism if they believed them to follow largely from an overly narrow conception of a connectionist model. My aim has been to encourage the development of such skepticism.

Toward that end, I have highlighted the fundamental importance of what I call unstructured pattern association in the views of Smolensky and of McClelland and Rumelhart. This was the aim of sections 3 and particularly section 4. Having established the role of unstructured pattern association, I turned to the question of whether eliminative connectionism retains its plausibility given the possibility of a wider range of connectionist models. In section 5, then, I examined three different ways of modifying unstructured pattern association. I argued that in each case, the possibility that connectionist models implement symbol processing models is substantially increased. Finally, in section 5.2 I argued that these modifications of unstructured pattern association are not mere possibilities. In each case, there are reasons for thinking that connectionist models adequate to account for human cognitive abilities may include such features. We can conclude not only that unstructured pattern association plays a crucial role in Smolensky's and McClelland and Rumelhart's eliminative connectionism, but that the views lose their appeal if an appropriately broad range of connectionist models are considered. A skeptical attitude toward these eliminative connectionist views is, therefore, in order.

REFERENCES

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. 1985. "A Learning Algoritnm for Boltzmann Machines." Cognitive Science 9, 147-169.

Agha, G. 1986. Actors: A Model of Concurrent Computation in Distributed Systems. Cambridge, Mass.: MIT Press.

Anderson, J. A. 1983. "Cognitive and Psychological Computation with Neural Models." IEEE Transactions on Systems, Man and Cybernetics 13, 799-815.

Anderson, J. R. 1983. The Architecture of Cognition. Cambridge, Mass.: Harvard University Press.

Block, N. 1986. "Advertisement for a Semantics for Psychology." In P. French, T. Eunling, and H. Wettstein (Eds.), Studies in the Philosophy of Mind. Volume 10: Midwest Studies in Philosophy. Minneapolis: University of Minnesota Press.

Broadbent, D. 1985. "A Question of Levels: Comment on McClelland and Rumelnart." Journal of Experimental Psychology, General 114, 189-197.

Burge, T. 1986. "Individualism and Psychology." Philosophical Review 95, 3-45.

Cummins, R. 1975. "Functional Analysis." Journal of Philosophy 72, 741-760.

_____. 1983. The Nature of Psychological Explanation. Cambridge, Mass.: MIT Press.

Dennett, D. C. 1986. "The Logical Geography of Computational Approaches: A View From the East Pole." In M. Brand and R. Harnish (Eds.), The Representation of Knowledge and Belief. Tuscon: University of Arizona Press.

Feldman, J. A. and Ballard, D. H. 1982. "Connectionist Models and their Properties." Cognitive Science 6, 205-254.

Field, H. 1977. "Logic, Meaning, and Conceptual Role." Journal of Philosophy 74, 347-375.

Fodor, J. 1968. "The Appeal to Tacit Knowledge in Psychological Explanation." Journal of Philosophy 65, 627-640.

_____ . 1978. "Propositional Attitudes." The Monist 64, 501-521.

_____. 1980. "Methodological Solipsism Considered as a Research Strategy in Cognitive Science." The Behavioral and Brain Sciences 3, 63-73.

_____. 1983. The Modularity of Mind. Cambridge, Mass.: MIT Press.

_____. 1986. "Banish DisContent." In J. Butterfield (Ed.), Language, Mind, and Logic. Cambridge: Cambridge University Press.

_____. 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. Cambridge, Mass.: MIT Press.

Grice, H. P. 1977. "Logic and Conversation." In G. Harman and D. Davidson (Eds.), Semantics of Natural Language. Dordrecht: Reidel.

Harman, G. 1982. "Conceptual Role Semantics." Notre Dame Journal of Formal Logic 23, 242-256.

Haugeland, J. 1978. "The Nature and Plausibility of Cognitivism." The Behavioral and Brain Sciences 1, 215-226.

Hillis, D. 1985. The Connection Machine. Cambridge, Mass.: MIT Press.

Hofstadter, D. R. 1983. "Artificial Intelligence: Subcognition as Computation." In F. Machlup and U. Mansfield (Eds.), The Study of Information: Interdisciplinary Messages. New York: Wiley.

Lycan, W. G. 1981. "Toward a Homuncular Theory of Believing." Cognition and Brain Theory 4, 139-159.

Marr, D. 1982. Vision. San Francisco: Freeman.

McClelland, J. L. and Rumelhart, D. E. 1986. "A Distributed Model of Human Learning and Memory." In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Parallel Distributed Processing: Explorations of the Microstructure of Cognition. Volume 2: Psychological and Biological Models. Cambridge, Mass.: MIT Press.

McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. 1986. "The Appeal of Parallel Distributed Processing." In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations of the Microstructure of Cognition. Volume I: Foundations. Cambridge, Mass.: MIT Press.

Minsky, M. 1975. "A Framework for Representing Knowledge." In P. H. Winston (Ed.), The Psychology of Computer Vision. New York: McGraw-Hill.

Minsky, M. and Papert, S. 1969. Perceptrons: An Introduction to Computational Geometry. Cambridge, Mass.: MIT Press.

Minsky, M. and Papert, S. 1988. Perceptrons: An Introduction to Computational Geometry. Expanded Edition. Cambridge, Mass.: MIT Press.

Newell, A. and Simon, H. Human Problem Solving. Englewood Cliffs, N. J.: Prentice-Hall.

Pinker, S. and Prince, A. 1988. "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition." Cognition 28, 73-193.

Putnam, H. 1973. "Reductionism and the Nature of Psychology." Cognition 2, 131-146.

_____. 1975. "The Meaning of 'Meaning'." In K. Gunderson (Ed.), Language, Mind, and Knowledge, Vol. 7, Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.

_____. 1983. "Computational Psychology and Interpretation Theory." In H. Putnam, Realism and Reason. Vol. 3, Philosophical Papers. Cambridge: Cambridge University Press.

Pylyshyn, Z. 1984. Computation and Cognition. Cambridge, Mass.: MIT Press.

Ross, L. 1977. "The Intuitive Psychologist and His Shortcomings." In L. Berkowitz (Ed.), Advances in Experimental Social Psychology, Vol. 10. New York: Academic Press.

Rumelnart, D. E. 1975. "Notes on a Schema for Stories." In D. G. Bobrow and A. Collins (Eds.), Representation and Understanding. New York: Academic Press.

_____. 1984. "The Emergence of Cognitive Phenomena from Sub-Symbolic Processes." Proceedings of the Sixth Annual Conference of the Cognitive Science Society.

Rumelhart, D. E., Hinton, G. E., and Williams, Ronald J. 1986. "Learning Internal Representations by Error Propagation." In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, Mass.: MIT Press.

Rumelhart, D. E. and McClelland, J. L. 1986a. "PDP Models and General Issues in Cognitive Science." In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, Mass.: MIT Press.

Rumelhart, D. E. and McClelland, J. L. 1986b. "On Learning the Past Tenses of English Verbs." In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models. Cambridge, Mass.: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. E. (Rumelhart et al.) 1986. "Schemata and Sequential Thought Processes in PDP Models." In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models. Cambridge, Mass.: MIT Press.

Rumelhart, D. E., and Zipser, D. 1986. "Feature Discovery by Competitive Learning." In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, Mass.: MIT Press.

Schank, R. C. and Abelson, R. P. 1977. Scripts, Plans, Goals, and Understanding. Hillsdale, N.J.: Erlbaum.

Smolensky, P. 1986. "Information Processing in Dynamical Systems: Foundations of Harmony Theory." In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, Mass.: MIT Press.

Smolensky, P. 1987. "On the Proper Treatment of Connectionism." Technical Report No. CU-CS-359-87. Department of Computer Science, University of Colorado at Boulder. To appear in The Behavioral and Brain Sciences.

Stich, S. P. 1983. From Folk Psychology to Cognitive Science: The Case Against Belief. Cambridge, Mass.: MIT Press.