CONNECTIONISM, NATURALIZED EPISTEMOLOGY, AND ELIMINATIVE MATERIALISM

A Thesis submitted to the faculty of Graduate Studies and Research in partial fulfilment of the requirements for the degree of Master of Arts.

Gordon SF Krieger
Department of Philosophy,
McGill University, Montreal
December 1993

Abstract

The aim of this essay is to explore the potential for an epistemology consistent with eliminative materialism based on work in connectionist modeling

I present a review of the connectionist approach to psychological models that contrasts it with the classical symbolic approach, focusing on the nature of their respective representations While defending the legitimacy of the connectionist approach, I find that its most useful application is as a basis for neuroscientific investigation

Discussing connectionist psychology, I find it inconsistent with folk psychology and therefore consistent with eliminative materialism I argue also for the naturalization of epistemology and thus for the relevance of psychology for epistemology The conclusion of the essay is an outline of connectionist epistemology, which centres around two mathematical analyses of the global activity of connectionist networks I argue that connectionist psychology leads to a version of epistemic pragmatism.

Resume

Le but de cette etude est d'explorer le potential pour une epistemologie compatible avec le materialisme eliminationniste base sur les models psychologique de connectionisme.

Je presente un sommaire de l'approche connectioniste qui etude les differences entre connectionisme et l'approche classique Je défends la legitimite de l'approche connectioniste, mais je trouve que l'utilisation preférable de connectionisme est comme un outil pour l'investigation neuroscientifique

Je trouve que la psychologie connectioniste est incompatible avec la psychologie des gens («folk psychology»), donc compatible avec la materialisme éliminationniste. Je défends la naturalization de l'epistémologie et ainsi la pertinence de la psychologie pour l'epistemologie. La conclusion de l'essai est une sommaire de l'épistemologie connectioniste, qui implique deux analyses mathematiques de l'activite globale des appareil connectionistes: je trouve que la psychologie connectioniste suggere une version de pragmatisme epistemique

Acknowledgments

A number of people were helpful in getting this thesis from head to paper. Firstly, my supervisor, Paul Pietroski, was exemplary in helping me weave a thread out of a tangled knot of ideas, and in helping me realize the sorts of questions I was asking. Paul was also instrumental in allowing me to meet rather pressing deadlines that I had set for myself.

Patricia Sheridan, in returning a previous favour of computer time a hundred times over, allowed me to get a substantial amount of work done during otherwise difficult times. A working computer, good company, and an apartment a great deal warmer than my own were all conducive to writing philosophy. I thank Patricia both for the help and the hospitality.

As well, my parents, Janet and Fred Krieger, were of much needed assistance in filling the few months gap between two incomes, which allowed me time to worry more about philosophy and less about money.

Partial funding for this thesis was provided by Employment and Immigration Canada and by the Ministère de la Main-d'œuvre, de la Securite du revenu et de la Formation professionnelle (Quebec).

Contents

Introduction

The goal of this thesis is to explore the potential for an epistemology consistent with eliminative materialism based on work in connectionist modeling. Eliminative materialism is the thesis that (1) our commonsense understanding of psychology constitutes a theory, ("folk psychology"), (2) folk psychology is a false theory, which (3) will be replaced by, rather than smoothy reduce to, a mature cognitive theory.[1]

There are various arguments for eliminative materialism: folk psychology is a "stagnant research programme", it is not likely to cohere with well established theories in adjacent and overlapping domains, it tells us nothing about learning, language acquisition, sensorimotor coodination, sleep or mental illness.[2] My project here is not to undertake a defense of eliminative materialism, but rather to develop a compatible epistemology. Eliminativism conflicts with traditional epistemologies, because the latter are typically based on the entities and processes posited by folk psychology

Beliefs, for example, are central to both folk psychology and traditional epistemology. Most accounts of epistemology see knowledge as justified true belief, there are many variations on this theme but very little real deviation from it. Since the elimination (as opposed to mere revision) of folk psychology will inevitably include the elimination of belief, eliminativism will require a significant shift in epistemology. A new understanding of cognition requires a new understanding of epistemology. But we do not eliminate epistemology because, like psychology, it is a field of inquiry, a series of questions; those questions, or most of them, remain.

In the thesis I explore the epistemological implications of "connectionism", a relatively recent approach to cognitive modeling. In investigating epistemology in this manner, I am practicing what is referred to as "naturalized" epistemology this is an approach that sees empirical inquiry as relevant to epistemology. There are a variety of views of epistemology that count as naturalistic in this sense; I discuss this approach in the third chapter.

On the psychological side, I argue in the second chapter that connectionism and folk psychology are incompatible. So whatever we learn about epistemology from connectionism,

---

1 This statement of eliminative materialism largely follows that of the Churchlands, in eg: Paul Churchland (1981), or Patricia Smith Churchland (1986). Some may know eliminative materialism under the different name "San Diego Imperialism"

2 All from Churchland (1981).

it should prove to be compatible with eliminative materialism. To this end, I should note at the outset my assumptions regarding the fate of folk psychology. Throughout the thesis, I will assume that, in order ultimately to escape elimination, folk psychology must prove compatible with our best account of the mechanisms of cognition By "compatible with" I mean something like "smoothly reducible to" or "equivalent". If the structures and processes posited by our best account of cognition are significantly unlike those posited by folk psychology, then we must reject folk psychology and adopt the superior theory. I have left the reading of compatibilty open, because my aim here is not to give an account of inter-theoretic relations, but rather to view folk psychology as a rival of other accounts of cognition

This view of folk psychology seems to me quite reasonable, but it is not universal There are a number of people who share this view of the vindication of folk psychology, including those who think that folk psychology will in fact be vindicated (eg, Jerry Fodor, William Lycan), as well as those who do not (eg, Stephen Stich, Paul and Patricia Churchland). As well, there are a variety of views of folk psychology amongst those who disagree with these requirements for the vindication of folk psychology, although this view is typically held by people on friendly terms with folk psychology (eg, Daniel Dennett, Donald Davidson). So, properly understood, my claims concerning folk psychology are conditional, based on the adequacy of this view of vindication.

Similarily, I will throughout the thesis understand cognitive modeling and cognitive science in general to be concerned with describing the internal mechanisms of cognition, rather than (merely) the explanation of behaviour. My interest in connectionism is in the implications it may have for an understanding of cognition. Such a view of connectionism is important, but perhaps not crucial, for the connectionist epistemology developed in the last chapter, as well as for the discussion of folk psychology.

That leaves us with the first chapter, which is a review of connectionism; it contains all of the connectionist background necessary to understanding the rest of the thesis.[3] I concentrate on contrasting connectionism with the classical sentential approach to cognitive modeling, and on justifying its legitimacy as a model of cognition.

---

[3] While all of the necessary information can be found in the first chapter, connectionist neophytes may find my overview somewhat opaque. Those seeking a more involved introduction should look to Bechtel and Abrahamsen (1991), presently the only book length introduction to connectionism. The first four chapters of Rumelhart, McClelland and the PDP Research Group (1986) are also helpful.

Chapter one
An analysis of connectionism

The goal of this chapter is to present a review of connectionism, in order to both provide the background necessary to understanding the rest of the thesis, and to analyze the various features of connectionist models with the aim of contrasting them with more standard approaches to cognitive modeling. Recall that my concern with cognitive modeling and cognitive science is the description of the internal mechanisms of cognition. This will be more important in later chapters, but it is best repeated as we venture into connectionism.

Connectionist modeling aims to model cognition by using elementary processors organized in networks, where each unit has connections with several other units. The units in a connectionist model employ only the barest imitations of some of the gross functions of neurons. The elementary units in a connectionist network are simple processors, each of which computes a (positive real number) activation value from the activation value of neighboring units together with the (real number) weight of the connection between the units. The input to a connectionist network is provided by the activation of input units (units that covary with some feature of the environment). The networks are structured so that activation propagates through a network to the output units, whose activation values constitute the output of the system. Different sorts of connectionist models will differ in the properties of their units, and in the means of propagation of activation (differ in the structure of connectivity between units).

Connectionism in its present form has its genesis in the late 1950s and early 1960s, particularly in the work of Frank Rosenblatt (eg. 1962), although its emphasis on parallel processing and self-organizing networks owes many debts to earlier work, including the Associationist school and to the work of Donald Hebb. Rosenblatt's "perceptrons" were networks of simple, neuron-like elements given simple classification tasks. Perceptrons were simulated on digital computers and subjected to formal mathematical analyses, two techniques basic to modern connectionism. Rosenblatt's aim in his work was not to model any specific part of the nervous system, but rather to

> study ... lawful relationships between the organization of a nerve net, the organization of its environment, and the "psychological" performances of which it is capable... The model is not the terminal result, but a starting point for exploratory analysis of its behavior.[1] (1962: 28)

---

[1] Rosenblatt (1962: 28).

Rosenblatt's analysis led him to the "perceptron convergence theorem": given any "world" (a string of data) and a classification task for which a solution exists, a perceptron will yield a solution to the task in a finite time

Perceptrons fell out of favor in part due to the work of Minsky and Papert (1969) The problem with perceptrons was not that the convergence theorem was false, but rather that the abilities of perceptrons were such that there were many ordinary classification tasks for which no perceptron solution existed By organizing perceptrons into multiple layers, solutions for these tasks could be achieved, but only with astronomical complication (Later, with connectionism, such restrictive complication vanished with a slight modification of the perceptron model.) The climate of the Artificial Intelligence community at the time was such that the fall of the perceptron was generally taken to spell defeat for any sort of quasi-neural, parallel processing approach to Artificial Intelligence, in favor of the serial, program writing approach favored by Minsky and others.

The "new connectionism" began to rear its head in the early 1980s. Some of the earlier work is exemplified by Feldman and Ballard (eg 1982). The units in their networks were each used to represent a particular concept: "redness" or "largeness" for example. This use of units is known as local representation. In contrast, most of the more recent work in connectionism employs a group of processing units to represent a single concept or entity Connectionist work of this sort is known as "Distributed Connectionism" or sometimes as "Parallel Distributed Processing" ("PDP"); in distributed connectionist networks, the tasks of representation and of processing are distributed across the units in a network. A particular representation is fully distributed if every unit of the network is involved in the representation. I will reserve discussion on how states of a connectionist network may qualify as representations until later in this chapter

The representations in a distributed connectionist network involve sets of units; they are usually referred to as active representations because representations involve the activation levels of a particular set of units; when the network is dormant it does not represent. Which units are involved in a particular representation depends on the particular representation and the structure of the network. In some networks, the active representation will be the pattern of activation over all of the units at a particular time. In feedforward networks (networks structured in distinct layers) different levels are active at different times, so the active representation will usually be the pattern of activation over the units in one layer. The active representation of a network is a product only of the given input and the various weights of the connections between the units

Connectionist representations of this sort are said to be distributed because it is the pattern of activation itself that constitutes a representation, rather than the activation of individual units, or of the activation of groups of units at different times As patterns of activity in a network, distributed representations of this sort may be analyzed as an ordered set or vector. Further, since the activity of an individual unit need not be non-zero to be part of a set of activation values, particular units need not be active in order to play a representational role

In connectionist networks, representation and information storage are two distinct tasks While representation is left to the activity of units, the task of information storage is left to the weights of the various connections between processing units When a network is dormant, receiving no input and doing no processing, it represents nothing. We have seen that the active representation is a function solely of input and the weight of the connections between the units. So while connection weights are not part of any particular representation, they mediate representation by the role they play in producing the proper activation patterns for a given input Connection weights are the modifiable aspect of a connectionist network, it is by altering the connection weights that an experimenter, or the network itself, may improve the network's accuracy in terms of providing the appropriate output for the given input

The moral of all of these opaque comments about connection weights is that distributed connectionist models blur the line between representation and processing Representation in connectionist networks is more of a process than a tool of processing (That is, "representation" is more of a verb than it is a noun) McClelland, Rumelhart and Hinton put it this way:

> The representation of the knowledge is set up in such a way that the knowledge necessarily influences the course of processing Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear; it is part and parcel of the processing itself [2] (1986: 32)

One cannot read the literature explaining the advantages of connectionism (what Fodor and Pylyshyn call the "polemical literature") without being well prepared to distribute a few grains of salt where necessary, and the preceding quote is a good example It is, despite the confidence of its authors, rather unclear what gets to count as knowledge at all, a fortiori what counts as knowledge in a connectionist network And further, most of what McClelland, Rumelhart and Hinton say above can be said of classical, program writing

---

2 McClelland, Rumelhart, and Hinton (1986: 32)

models "knowledge" necessarily influences processing, and knowledge access (information retrieval) is "part and parcel" of processing

For example, an "expert system" is a sentential AI program, the purpose of which is to embody a base of knowledge in some domain, and usually consists of a large base of if-then "production rules". There are expert systems for medical diagnosis, for example. If it knows anything, an expert system knows by virtue of its production rules. (We could imagine a rather simplistic medical diagnosis production rule like "IF COUGH THEN COLD") Since an expert system embodies its knowledge in rules of inference, one could make the same claims about expert systems that McClelland, Rumelhart and Hinton make about PDP models in the quote above. the representation of the knowledge is set up in such a way that knowledge necessarily influences processing, and the information relevant to the situation is part and parcel of the processing So the difference between connectionist and sentential AI doesn't seem to be the relationship between representation and processing.

If there is some genuine difference between classical and connectionist models that McClelland, Rumelhart and Hinton have captured in the quote above, it concerns the mode of representation in the two types of models (I will take this claim for granted here, for the sake of introducing the unique method of representation used in distributed connectionist networks Later sections will address the debate as to whether connectionist representations are genuinely different from those of symbol systems)

In connectionist models, representation is said to be distributed across the units of the network As we saw above, features of the world are represented by patterns of activation over many units, and an individual unit can be involved in different representations. Most of the advantages that connectionism claims over symbol systems can be seen as a product of this method of representation.

There is a sense, hinted at above, in which the connection weights play a representational role, insofar as they store information by having one weight instead of another But again, it is a queer sort of representation. Not only is the representation distributed in the same sense as the active representation was (by involving more than one unit), but the weight of a single connection is also involved in (the mediation of) representing more than one entity. This sort of information storage is called superpositional storage.

For example, in a network that can learn to associate the terms "dog", "cat" and "bagel" with three different types of input,[3] the connection weights are set with the weights appropriate to such a task. One might ask, reasonably enough, as to where the information necessary for the proper recognition of dogs is stored The answer, as we should all know

---

[3] McClelland and Rumelhart (1986).

now, is that it is stored in (or by) the connection weights - all of them. But where then is the additional information necessary for recognizing cats and bagels? In the very same connection weights - all of them. Connectionist networks can use all of the same units for various functions. If it makes any sense at all to speak of the connection weights as storing information, then they must be said to store in a superpositional fashion. Information is not stored in specific locations or at addresses: it is smeared together and spread over the network

This style of information storage is importantly different from that of the standard, program-writing approach to Artificial Intelligence (AI) The problem of access to relevant knowledge, which is a particular instance of the frame problem, has proved a serious stumbling block for the "Good Old Fashioned AI" approach to cognitive modeling The problem, basically, is that there is far too much information to go around, and behavioral efficiency (and survival!) requires determining the information relevant to a given situation quickly, while ignoring the rest, so long as it remains irrelevant When one is being chased by a tiger, for example, it's best not to concentrate on the colour of the leaves Of the classical AI systems that try to deal with this problem, most are painfully slow, and equally slow organisms would make a nice meal for the slowest of predators.

It is an unfair caricature of both the frame problem and sentential AI to say that the frame problem cripples classical AI, but is not a problem for connectionism. Connectionist networks, at least as they have been used so far, usually deal with much smaller domains than those tackled by sentential AI; with less information to deal with, the frame problem is less likely to be a real problem Yet there is some difference between the two approaches in terms of the frame problem, and the difference seems to lie in the different means of information storage employed in the two models. But an articulation of the difference is want of an explanation of exactly how "knowledge" is more implicated in processing in connectionist networks than it is in sentential models

§

The debate over the nature and sufficiency of connectionist representations is at the heart of much of the debate over the adequacy of connectionist models of cognition. There are some, particularly Fodor and Pylyshyn (1988), who think that connectionist models of cognition are untenable because connectionist representations do not meet certain

requirements, they keep connectionist models from meeting certain architectural constraints required of any account of cognition.

The classical, "symbolic" view of cognition holds that mental representations are structured. There can be atomic or unstructured representations, but most useful representations, on this view, will be molecular. Molecular representations are made up of other representations, which are themselves either atomic or molecular, and the different parts can be put together in different ways: representations have a combinatorial syntax. The semantic content of these sorts of representations is a function of the syntactic structure of the representation and the semantic content of its constituents. Included in this view of cognition is the thesis of the correlation of syntax and semantics of mental states: differences in the content of mental states are mirrored in differences in their syntax

Computation, on this view, is the manipulation of mental representations. Cognitive processes are structure sensitive. they apply to representations by virtue of their syntactic structure, content does not play a role in computation. But once again, the correlation of syntax and semantics is at play here: syntactic transformations of representations will make sense from a semantic point of view Just as "Q" can be derived from "P&Q" (where "derived from" is understood as a syntactic relation), so too does the latter truth-functionally entail the former.

Generally, this "classical" sententialist view of cognition is of a materialist stripe the symbols that are the constituents of structural representations are seen as physical states of whatever system is being considered. They may be fairly complex, scattered sorts of states, but they must be syntactically and semantically atomic in order to count as the sorts of symbols with which this view is concerned. The view of computation the classical conception of cognition presents us with is not so much a view of the physical process of computation, but rather a demand for a certain mapping function: a physical system can be seen as computational if it is possible to map states of the system onto formulae in a computing language so that semantic relations among the formulae are preserved by computation.[4] Computation on the classical view is basically a change of state from one string of symbols to another· it is the manipulation and transformation of structured representations There is no shortage of reasons for adopting the classical, symbolic view of cognition One can appeal to the constituent structure of mental states in order to explain what seem to be features of cognition; it allows us to explain the productivity of thought; to explain how unbounded expressive power arises from finite means. It explains the coherence of our inferences and the systematicity of cognition

---

[4] Fodor (1975· 73).

Connectionist representations and processes differ considerably from those involved in the classical symbolic view They lack, at least on first analysis, the combinatorial structure and semantics of symbolic representations; and certainly if connectionist representations (or whatever we might call them) lack structure, then the processes by which they are manipulated must be defined over something other than structure, unlike symbolic processes.

Fodor and Pylyshyn (1988) think that all connectionist representations are atomic, atomic in the sense that they have no structure, and no constituents. Connectionist processes therefore must be sensitive to something other than structure, again going against classicism. This leads Fodor and Pylyshyn to the conclusion that a connectionist account of cognition is inadequate, because, unlike classicism, such an account would be unable to explain obvious features of cognition such as the productivity and systematicity of thought, inferential coherence, etc.

Consider, as Fodor and Pylyshyn would have us, two machines that draw the inference from "A&B" to "A" and "B": one classical and one connectionist. The classical machine has a tape upon which different expressions are written; the machine is constructed so that whenever a token of the form "A&B" appears that will cause the machine to write tokens of both "A" and "B" onto the tape Simple enough. Fodor and Pylyshyn's idea of a connectionist network that performs this inference involves three nodes: one ascribed the content "A&B", another "A" and a third "B". The network is arranged so that when the "A&B" node is activated, so too are the "A" and "B" nodes

In the classical machine, tokens of "A&B" literally have as their constituents tokens of both "A" and "B", and it is because of this that the semantics of the expression "A&B" is determined "in a uniform way" by the semantics of its constituents. Neither of these things is true of the imagined connectionist device That device is constructed so that there is a causal connection between tokenings of "A&B" and of "A" and "B", but there is no structural (eg part-whole) relation between them. The connectionist tokenings of "A&B" are, despite the ampersand, atomic and not molecular; they have no constituent parts and no structure; they are syntactically and semantically atomic.

Because connectionist representations are all atomic, connectionist devices, it is argued, cannot account for the obvious features of cognition that are so smoothly explained by the symbolic view.

> .. [S]ince the Connectionist architecture recognizes no combinatorial structure in mental representations, gaps in cognitive competence should proliferate

---

5 Fodor and Pylyshyn (1988: 15-16).

> arbitrarily ... [Connectionist architecture treats mental representations] not as generated sets but as lists. But lists, qua lists, have no structure; any collection of items is a possible list. And, correspondingly, on Connectionist principles, any collection of (causally connected) representational states is a possible mind So, as far as Connectionist architecture is concerned, there is nothing to prevent minds that are arbitrarily unsystematic. But that result is preposterous.[6]

Surprisingly little is said about the nature of representations in the connectionist literature Most of the literature concerning connectionist representations concerns their causal role in the operation of particular connectionist models. The extent to which states of a connectionist device merit the name "representation" is usually ignored Paul Smolensky is responsible for much of the work in which the nature of connectionist representations is given due consideration. To properly understand Smolensky's account of connectionist representations and how that contrasts with the understanding of Fodor and Pylyshyn, we should first consider the levels of analysis that Smolensky takes to be relevant when studying a connectionist device

Smolensky thinks that there is level of analysis (of connectionist devices) that lies between the symbolic level (the level at which cognition is described in the manner of symbol theories) and the neural level (the level of the operation of individual neurons). He calls this level the subsymbolic level. The approach to cognitive modeling that takes the subsymbolic level as its preferred level of analysis Smolensky calls the subsymbolic paradigm This is the approach that Smolensky attributes to connectionism, properly treated. The traditional approach, which prefers the symbolic level, is the symbolic paradigm

The significance of the difference between the symbolic and subsymbolic paradigms (and their particular levels of analysis) lies in the syntactic and semantic status of symbols The symbol level view understands symbols to be simple; they are semantically and syntactically atomic. Not so for the subsymbolic view At this level, the symbols (of the symbolic level) are complex, both syntactically and semantically. Atomic symbols, on this view, are composed of subsymbols. The semantic role of subsymbols, as we explore below, involves the representation of "microfeatures"; entities typically represented by a single symbol at the symbolic level are typically represented by a large number of subsymbols.

The syntactic life of subsymbols is further unlike that of their larger cousins Subsymbols, as the operation of individual processing units, are manipulated in a numerical fashion

---

6 Fodor and Pylyshyn (1988. 49).

> [Subsymbols] participate in numerical - not symbolic - computation. Operations in the symbolic paradigm that consist of a single discrete operation (e.g. a memory fetch) are often achieved in the subsymbolic paradigm as the result of a large number of much finer-grained (numerical) operations.[7]

Before exploring further the nature of Smolensky's connectionist representations, consider again Fodor and Pylyshyn's view Their connectionist network had three nodes: one for "A&B", another for "A", and a third for "B" A network of this sort is not a distributed connectionist network; yet it is distributed connectionism where most interest is focused So the short way with Fodor and Pylyshyn is to say that their criticism is just misplaced; they are criticizing Feldman and Ballard style "localist" connectionism But Fodor and Pylyshyn see distributedness as a red herring. Their A&B network, they claim, could very well be distributed if their simple nodes were complex at a lower level. But a representation's being complex does not means that it has constituents. A representation has constituent structure only when its parts are semantically evaluable So distributedness is irrelevant. The compositionality of mental states is a "within level" issue (within the "representational level") while distributedness is a "between level" issue.[8]

Smolensky agrees that the distributedness (or not) of a representation is a "between level" issue, but he admonishes Fodor and Pylyshyn for thinking that this means that distributedness has no bearing on issues within the representational level. Smolensky thinks that the distributedness of representations has implications for their compositionality.

On Smolensky's analysis, while "subsymbols" in a connectionist device are the activity of individual processing units, "symbols" are the patterns (vectors) of activation of units. Subsymbols represent "microfeatures", which are at best explained as 'low level semantic details', for example, the symbol for "apple" might involve the subsymbols of "redness", "roundness", and so on. A more detailed example should help in explaining how Smolensky sees the relation between symbols and subsymbols, and how distributedness affects the composition of mental representations.

Smolensky gives this example in a number of places (including Smolensky 1987, 1988); it involves a hypothetical connectionist representation of coffee. The details of the example are rather close to the sensory level, and the microfeatures involved are not very micro, but suffice it to say that this is not an essential feature of the example Typically, microfeatures are meant to be somewhat less complex than those in the example

---

7 Smolensky (1988: 3).

8 Fodor and Pylyshyn (1988 19)

We are to consider first a distributed connectionist representation of cup with coffee. This would be a pattern of activation involving many different microfeatures: "brownness", "heat", "cylindrical container", "liquid contacting container", and so on. Consider a separate representation of cup (without coffee). This would involve the microfeatures "cylindrical container", "porcelain surface", etc.

Given this, how then are we to understand the connectionist representation of coffee? The preliminary answer is that the representation of coffee is just the representation of cup with coffee with all of the microfeatures involved in the representation of cup without coffee subtracted away. So we are left with a distributed representation of coffee, but it is of coffee in a particular context, that of being in a cup. Even without the microfeatures involved in cup without coffee, the context remains in the shape of microfeatures describing the interaction of cup and coffee. "liquid contacting porcelain", "liquid in cylindrical shape", and so on.

Since there is nothing sacred about coffee in cups, we could have easily started with can of coffee, or tree with coffee, and subtracted away features in order to produce a representation of coffee. But we would again find context creeping in; coffee in cans is generally a brown powder, on trees it is beans. The examples here are unnecessarily extreme, but the point should be emerging: There is no one distributed connectionist representation - no one symbol - of coffee, there is instead a family of related representations. Smolensky makes much of this claim concerning the context sensitivity of connectionist representations. His first moral from the coffee example is the claim that connectionist representations do have constituents.

Fodor and Pylyshyn have argued that a representation has constituent structure only when its constituents are themselves semantically evaluable. Smolensky and others in the connectionist camp have not disputed this view, and Smolensky himself seems to have laid a lot of groundwork toward the claim that connectionist symbols have constituent structure, since they have semantically evaluable subsymbols as their constituents. But that is not the claim that he makes. Smolensky does propose that connectionist representations have constituent structure, but the constituents he has in mind are vectors. The representation (vector) cup with coffee, for example, is composed of representations (vectors) of cup and of coffee. The relation between cup with coffee and cup is within the representational level. Smolensky concedes to Fodor and Pylyshyn that the relation between a vector and its individual elements (the relation between a symbol and a subsymbol) is a "between level" relation.[9] He ignores here the possibility of an interesting debate concerning the nature of the relation between symbols and subsymbols. It is also somewhat surprising to wade through

---

9 Smolensky (1987: 148).

accounts of the semantic content of subsymbols, only to find that the relation between symbols and subsymbols is not within the representational level. At any rate, the concession is made, and consequently most of Fodor and Pylyshyn's arguments against Smolensky are irrelevant, because they portray him as denying just the point that he concedes.[10]

So the vectors that are connectionist representations have smaller and equally representational vectors as their constituents This of course has to bottom out somewhere, vectors cannot be divided into smaller vectors ad infinitum. But recall also that the classical symbolic account has atomic symbols; that something has only one constituent does not mean that it lacks constituent structure Later we will explore the extent to which Smolensky's account satisfies the demand for constituent structure

Smolensky's account of connectionist representations serves to highlight some of the differences between connectionist and symbolic systems. Firstly, while the constituency of mental states is useful for the analysis of the behaviour of connectionist devices, it is not a part of their causal mechanism. The processes by which connectionist representations are manipulated are not defined over structure; it is the individual units that matter And whatever arbitrariness and ambiguity there is in what counts as a representation of cup or of coffee is an arbitrariness and ambiguity of analysis, not of processing. The real heart of the mysteries of connectionism, and what serves to separate it from classical symbol views, is that the entities manipulated by the equations that define connectionism (the activation levels of individual units) are not the same entities that get semantically evaluated (activation patterns) This leads to a further view of Smolensky's, that the decomposition of composite connectionist representations into their constituents is not precise, but approximate. Again one encounters ambiguity in the composition of connectionist representations, but again the ambiguity is one of analysis and not of function

Although Fodor and Pylyshyn misrepresent Smolensky on this topic, we should not too soon commit all of their arguments to the fire. In particular, they criticize Smolensky (as they understand him) for having a mistaken notion of constituency, they portray him as misusing the term "constituency" to refer to a semantic relation between predicates They argue that while "the florist" is a constituent of the sentence "Joan loves the florist", one would not normally say that "has a handle" is a constituent of "cup" any more than one might say that "the English phrase 'is an unmarried man' is part of the English phrase 'is a bachelor'".[11]

---

10 Fodor and Pylyshyn (1988. 19-28).

11 Fodor and Pylyshyn (1988: 21-2) In the interest of avoiding examples of the "John loves Mary" sort (the one actually used by Fodor and Pylyshyn), I have borrowed "Joan loves the florist" from Bechtel and Abrahamsen (1991).

Fodor and Pylyshyn are of course correct to demand that "real constituency" involve parts and wholes and not (just) semantic relations But what Smolensky takes to be constituents really are parts of a whole representation, and the relation between the two is within the representational level To the extent that we understand cups as being the sorts of things that have handles, there is a semantic relation between has a handle and cup; but on Smolensky's account of connectionist representations, the former representation really is a part of the latter. The complex representation cup is composed of different representations such as has a handle Fodor and Pylyshyn have again misunderstood Smolensky; he does not mean for semantic relations to constitute constituency; his interpretation of constituency is quite standard.

Returning to Fodor and Pylyshyn's analogy, while it is true that "is an unmarried man" is not part of the English phrase "is a bachelor", the analogy simply doesn't fit with Smolensky's representations. From the beginning, Smolensky and other connectionist advocates have emphasized the differences between connectionist and symbol style representations, and a natural language expression is an exemplar of the latter sort. That the connectionist representation cup has as a constituent the representation has a handle, while a parallel expression in English would not, is just the difference between the two styles of representing, rather than an indication of some error of Smolensky's

The misunderstanding here might arise also from worries about the processing of symbols, but we have seen already how Smolensky embraces the ambiguity involved in decomposing a complex representation into its constituents, and how connectionist processesing involves only the individual units. Smolensky's representations are more tools of analysis than items to be processed; they are more outcomes of processing than tools implicated in processing But while Fodor and Pylyshyn's criticisms of Smolensky's understanding of constituency may be wrongheaded, there is reason to believe that Smolensky has misunderstood the nature of the constituents involved in connectionist representations, as we will see below.

Smolensky concludes that his account of connectionism satisfies the demand for representations with combinatorial syntax and semantics and yet does not implement a language of thought. While his representations, he claims, meet all of the criteria demanded of them by the classical symbol view the context dependency and ambiguity involved in connectionist processing makes it clear that processing operates in a manner significantly different than it would if it were implementing a language of thought It is unclear exactly why Smolensky should want to devote space to establishing constituent structure for connectionist representations when he goes on to deny that they are not simply implementing a language of thought. He seems to want to claim for connectionism all of the advantages that

the classical view claims for constituent structure but, if that is his goal, he does so without making a clear connection between his understanding of connectionist representations and the supposed advantages

While Smolensky has made a good case for the position that connectionist representations have constituents, he has largely ignored the classical symbolic demand for a combinatorial syntax and semantics for representations. He is correct to say that his connectionist representations have constituents, but it is not at all clear that this is the sort of constituency demanded by the symbolic view. Since they are divisible into parts, connectionist representations do have constituents. But the symbolic view of cognition demands not simply constituency, but constituency of a certain type, it demands of representations that they have a combinatorial syntax and semantics

Smolensky and other connectionists make few attempts to convince us that connectionist representations have a combinatorial syntax and semantics. And there is ample reason to think that they do not. To begin with, connectionist representations are not ordered, at least not in the same manner as symbolic representations. The two symbolic expressions "Rab" and "Rba" are importantly different (as expressions in first order logic), the role of order in connectionist representations is not as clear. Smolensky's representations are conglomerates of different representations where the order is largely irrelevant to processing. Connectionist representations are processed by virtue of being composed of individual units, over which the equations governing processing are defined. Connectionist processing pays no direct attention to structural properties such as order. As vectors, connectionist representations are ordered; there is a difference between the vectors $\cdot 0,1 \cdot$ and $\cdot 1,0 \cdot$, but the difference between the two is manifested by the structure of connections between units in the network, not by processing alone.

The result of this feature of connectionist representations is that it becomes tricky to have representations of relational properties. We might reasonably demand of a means of representation that we be able to represent both "Joan loves the florist" and "The florist loves Joan" and be able to distinguish between them, but it is not immediately obvious how to do so if one clumps together all of "Loves", "Joan" and "The florist" without using order to indicate the direction of the relation.

One solution to this problem, and the solution employed by McClelland and Rumelhart in their past-tense acquisition network (McClelland and Rumelhart 1986), is to account for order by multiplying representations The proposal here is to have a different representation for each relationship; one for "Joan loves the florist" and another for "The florist loves Joan" A problem with this solution is the sheer gross tonnage of representations

required to represent the most basic relational properties. But this is not the only approach possible to the problem; the point here is that the problem is not insoluble

We have seen already some of the reasons that the symbolic view cites for demanding representations with constituent structure; but Smolensky's representations simply don't meet the demand They are not ordered in the way that the symbolic view demands. The symbolic view demands of representations that they have a particular sort of order, in order to account for features of cognition such as systematicity and productivity; the relation between connectionist representations and these features is not at all clear. To cry that connectionist representations are vectors and therefore ordered is to play on two different sorts of order rather than to save connectionism

Because connectionist representations are not ordered in the same manner as standard symbolic expressions, it is doubtful that they will prove to have the combinatorial syntax and semantics also demanded by the symbolic view. Different connectionist representations can be put together in different ways, just as cup can be combined with coffee or with tea, but we have seen how introducing relational properties muddies these waters significantly What role order plays in syntax and semantics will depend on what solution one uses to the problem of relational properties. The one solution hinted at above involves eliminating order altogether and employing only atomic symbols With only atomic symbols, there is little for a combinatorial syntax and semantics to do.

It is at best not obvious that connectionist representations have a combinatorial syntax and semantics. And no one, Smolensky included, has done very much to try to convince us that there is reason to think otherwise Smolensky has devoted a lot a effort and text to pulling constituents out of connectionist representations, but they simply are not the sort of constituents involved in the symbolic view. They don't do the things that the symbolic view demands of them so there is little basis for Smolensky to claim that they do

It should now be fairly easy to argue, as Smolensky does, that his connectionism does not operate by implementing a language of thought. To this end, Smolensky points to

> .the context dependency of the constituents, the interactions that must be accommodated when they are combined, the inability to uniquely, precisely identify constituents, the need to take seriously the notion that the representation of coffee is a collection of vectors knit together by family resemblance...

to argue that his otherwise quite symbolic connectionism does not (necessarily) implement a language of thought.[12] I will not evaluate what little argument Smolensky provides for this

---

12 Smolensky (1987: 149).

point, it should suffice to say here that his position becomes much more certain when we recognize that his constituents are not at all as he imagined.

§

Connectionist models, particularly those in the Parallel Distributed Processing vein, present the beginnings of a new and exciting approach to the understanding of cognition But the classical symbol - processing approach has successes of its own, and we should not be too quick to dismiss them In the final section of this chapter, I will argue that the greatest contribution connectionism can make is toward a better understanding of neuroscience. But we should first understand that many of the claims for the superiority of connectionism over symbol processing are not as well grounded as some might have us think The understanding here is that the debate between the two approaches concerns which of them provides the best account of actual mental representations.

There are in general three sorts of superiorities claimed for connectionism in the literature: (1) processing strategies intrinsic to connectionism that make it an appropriate choice for certain tasks (content addressable memory, default assignment, spontaneous generalization, satisfaction of "soft" constraints, etc.), (2) more general advantages not related to a particular processing strategy (graceful degradation and speed), and (3) biological plausibility. All three sorts of advantages emphasize the distance between the classical and connectionist approaches, yet all three have their difficulties

(1)

All of the first set of advantages ultimately concern the capacity to have a certain input-output profile. The details of the particular advantages are not important here (and I have listed only a few), but consider as a brief example the content addressability of memory In connectionist models of memory, memories are addressable by their contents, while this is often not the case with classical models. Content addressability is the ability to retrieve or otherwise "call up" stored information by virtue of its contents For example, my memory of "cat" might be retrieved (recalled, activated, or what have you) by asking me to name a pointed-eared furry fourlegged animal. Content addressable memory has the virtue of facilitating recall with incomplete or conflicting information Memory in classical models is usually addressed by assigning each set of information ("set", in the sense that my memory of "cat" is a set of information) and assigning it a numerical address in memory. Often this

address will be mapped on to one of the particular contents of memory, so that memory is content addressable for one of the contents of memory (This would be the case where, for example, all of the information on your bank account is accessed by your account number)

At any rate, despite the limitations that it might put on the processing necessary to obtain it, content addressability of memory is just the ability to output a set of information given one or more of its contents· so the content addressability of memory is a matter only of the ability to maintain a certain input/output profile. The same is true for all of the other features in (1)

This feature of all of the advantages in (1) - that they demand only a certain input-output profile - proves to be their undoing. We know, if the Church-Turing thesis is correct, that any computable input/output profile can be the profile of a Turing machine, and Turing machines are the archetype of classical architecture. So any input/output profile of a connectionist model can be the profile of a classical model. So the advantages in (1) are really not advantages at all

As is typical of the literature debating the merits of connectionist models, the fate of the features in (1) lies somewhere between the claims of connectionists (the features of (1) are reason to prefer connectionist models) and classicists (classical models can exhibit all of the features in (1)). If connectionists are guilty of ignoring the Church-Turing thesis, then it may be that classicists that make the sorts of claims I am attributing to them may be guilty of taking it too seriously.[13] The problem here is that while the Church-Turing thesis tells us about the computational capacities of Turing machines, it says nothing of the time required for computation If the thesis is correct, then there can be Turing machines (and therefore classical models) that exhibit the feature of content addressable memory. But such models will be inadequate as models of, for example, human memory, if they take an inordinate amount of time to compute the function demanded of them.

To be charitable, the motivation in the connectionist camp for making the claims that they do about (1) seems to be that the advantages seem to be natural features both of connectionist models and human cognition, but which are, at best, features of that have to be forced on classical models. But this is hardly reason enough to claim that a connectionist style account of cognition would be somehow better than a classical account There are various spheres of our cognitive performance for which classical models seem plausible, while connectionist models seem forced· this is true for almost anything involving linguistic behavior And it is fallacious to think that because some features of cognition seem to us to

---

13 I present Fodor and Pylyshyn (1988· 54ff) as classicists who seem to make this sort of claim, althought without explicity invoking the Church-Turing thesis

be natural and basic aspects of our cognitive life, that the best account of cognition will also have these capacities as "basic" and "natural" features.

The fate of the connectionist claims of advantage in (1) should therefore be cautious rejection There may be specific input/output profiles which are easier to realize in a connectionist machine (easier, for example, in terms of the time required for processing) But many of the alleged advantages that would qualify as type (1) advantages are just as easily realized in classical machines. (Which capacities these turn out to be is an empirical question, it's simply a matter of designing a classical machine that will maintain the appropriate input/output profile.) And further, if the advantages in (1) are meant to be reasons for preferring connectionist models over classical ones, it seems that there are many different advantages that classical models can claim over connectionist ones

(2)

There are a few advantages that connectionists claim for their models that involve more than the ability to maintain a certain input-output profile Two of the more conspicuous claims involve speed and graceful degradation

Graceful degradation involves the degradation of performance and is of two sorts degradation from noisy data and from physical damage A good case can be made for the position that degradation from noisy data involves only the ability to maintain a certain input/output profile - in that this sort of graceful degradation is the capacity to produce a certain sort of output given a certain sort of input - so I will ignore it here

Graceful degradation from physical damage involves the ability of a network or brain to maintain coherent performance given minimal physical damage As the amount of damage increases, so does the distance between the normal and damaged behaviors Some machines, like ordinary von Neumann computers, do not degrade gracefully. The most minimal physical damage to most areas of a von Neumann computer will result in a full-scale breakdown of performance, or no performance at all.

The connectionist claim is therefore this: Graceful degradation from physical damage is reason to prefer connectionist over classical models, since both brains and connectionist models degrade gracefully, while classical models do not

Again the truth about the wholesale advantages of connectionist models is not quite as simple as connectionist advocates would have us believe While it is true that connectionist models gracefully degrade and that classical models generally do not, the connectionist claims concerning graceful degradation are confused on two points First it is not essential to classical models that they degrade as poorly as they often do Second, the connection

between the graceful degradation of connectionist machines and that of brains is somewhat weak

Both problems involve the time-honoured functionalist distinction between mental states and the physical stuff in which they are implemented· the same algorithm ("same" in functional terms) can have different physical implementations Classical models, as presently implemented on von Neumann computers, are not very resistant to damage. But given a different implementation (eg: on a computing device with many, rather than one, processor), a classical model may degrade in a manner similar to that of brains and of connectionist devices.

The second and more interesting misunderstanding in the connectionist position as I have stated it involves the relationship between the degradation of brains and of connectionist machines Brains degrade gracefully after physical damage, and we might reasonably ask of a model of cognition that it exhibit this phenomenon. And in general, connectionist models do exhibit this phenomenon But here the status of the units in a connectionist model comes into question The recieved view among core connectionist advocates like Smolensky, Rumelhart and McClelland is that the units in a network are not meant to be identifiable as neurons, nor as some larger part of the nervous system. They play some unidentified functional role in processing. This view goes hand in hand with the position that connectionist models are meant to be accounts of the algorithms of cognition, rather than of the physical implementation of some algorithm.

We are concerned here with graceful degradation following physical damage But the status of "physical" is very different between brains and connectionist models Physical damage to brains is physical damage. But physical damage to the sorts of connectionist models we have been discussing is damage to an identifiable part of an algorithm, where no commitment is made to the implementation of the particular algorithm These two types of physical damage are very different, even though they both result in damage at the algorithmic level (alteration of the algorithm).The difference between them is just the difference between the implementational and algorithmic levels. Damage to physical brains is damage to implementational hardware. The cognitive level damage that may or may not occur depends on how the algorithm is implemented in the brain. Damage at the implementation level is often not easily mapped onto some algorithmic level damage. But damage in a connectionist model is directly mappable onto algorithmic damage because the damage done is exactly at the algorithmic, and not the implementational level.

Just as graceful degradation depends in part on the details of implementation, so does the speed of processing Since the same algorithm can have different implementations, the speed at which an algorithm is executed is a quite dependent on its implementation. So with

both graceful degradation and speed, connectionist models can claim superiority over classical ones, but there is nothing intrinsic to the classical symbol processing view that makes this so. It is only the present implementation of both sorts of models that makes this so.

(3)

What does it mean to say that a cognitive model is biologically plausible? The claim seems not to be that biologically implausible models are not biologically possible (whatever that would mean) but that, given the actual biology of brains, a particular model is more plausible as an account of how brains work By the use of the term "biological" here we seem to be taking for granted that the goal of cognitive models (and theories) is to provide an account of the operation of the nervous system. I have already expressed sympathy with this view, but we should not jump the gun on competing views; it will probably not matter to someone who does not mean to model the kinematics of the nervous system that their model is not biologically plausible.

We know so little about the biology of brains that it simply isn't clear what sorts of things make a model more plausible. Further, even if we had a more thorough understanding of the biology of brains, it still isn't clear what would make one model more biologically plausible than another. All of this is just to say that it isn't clear what biological plausibility is

If biological plausibility concerns plausibility given some specific hardware, then perhaps a biologically plausible cognitive model is just one that involves the same sort of hardware as actual brains. This seems to be something like what connectionist advocates have in mind. Brains use elementary processing units linked in parallel. Connectionist models, like brains and unlike classical models, use elementary processing units linked in parallel. Therefore connectionist models are more biologically plausible than classical models.

This would be an odd view to take of connectionism if one held, in accordance with the recieved view of connectionism, that connectionist units are not meant to correspond to neurons in the brain and are not meant to be understood as operating at the neural level If we instead view a connectionist unit as some sort of higher level functional construct with no commitments to its implementation, it is difficult to see in what sense we are to understand connectionist networks as being more biologically plausible than classical models (on this understanding of biological plausibility). If connectionist units do not have some identifiable correlate in the brain then any claim to biological plausibility needs to be supplemented with some account of how connectionist networks are to be implemented in brains The crux of

this point is that classical models can also be supplemented with an account of their implementation, and there is at the moment little reason to think that classical models plus implementation will be any less biologically plausible than connectionist models plus implementation.

On this basic understanding of what might be meant by biological plausibility, plausibility, like speed and graceful degradation, is a feature of the implementation of an algorithm, rather than a cognitive level feature. But if this is really what connectionists mean by biological plausibility, they are just contradicting themselves by saying that connectionism is both biologically plausible and is not concerned with the details of implementation. An alternative is to conceive of biological plausibility as a cognitive level feature, but it is unclear just what sorts of features would count as biologically plausible. It is also unclear, again, that connectionist models would fare any better than classical ones on this view of plausibility. However, this view of biological plausibility seems to be lurking behind some of the connectionist literature, at least when given a somewhat less charitable (yet defensible) reading One cannot help, upon hearing phrases like "brain style processing" that we are supposed to think of connectionism as more (biologically) plausible because connectionist networks, like brains, process in parallel. But it is simply a mistake to think that classical models ipso facto cannot employ parallel processing as well.

If we are to judge cognitive models (and therefore cognitive theories) by their empirical successes, then what room is there for claims of biological plausibility, which seems to be a distinctly non-empirical feature? The vague understanding of biological plausibility in the connectionist literature seems not to involve a superiority in empirical success Biological plausibility is not meant to be an empirical virtue of cognitive theories or models. So why then all of this talk about plausibility?

The short way to deal with this question is to point out all of the other non-empirical theoretical virtues that have arguably played a role in the practice of science Simplicity is often invoked (at least on some readings of the history of science) in the justification for preferring one theory over another. The capacity to unify predecessor theories is another non-empirical virtue that looms large in the history of science. And both of these virtues share with biological plausibility their degree of vagueness. (eg: When is one theory more simple than another?)

For some reason, theoretical simplicity seemed to some people to be a reason to think that when given two theories, approximately equal in their empirical adequacy, the simpler one had captured something that the more complex one missed, or the simpler one was at least to be preferred. Connectionists, for whatever reasons, have the same idea about biological plausibility. Given two cognitive theories that make approximately the same

predictions, the more biologically plausible one should be preferred and is likely to have captured something that the less plausible one missed.

It is then, nothing new or mysterious that someone might invoke some non-empirical quality of a theory as a reason to prefer it over its competitors. But claims about biological plausibility cry out for an articulation of just what is being claimed, and, given that, for some account of what good it will do us to choose "biologically plausible" theories over their competitors

It should be clear therefore that there is a significant gap between the symbolic and connectionist approaches, despite what one might think of the respective merits of the two This gap leads Smolensky to propose that an aim of connectionist research should be to "develop new formalizations of the fundamental computational notions that have been given one particular shape in the traditional symbolic paradigm".[14] Fodor and Pylyshyn, on the other hand, think that there is some place for work in connectionism but it is in support of, rather than in competition with, the symbolic approach

Given their dissatisfaction with connectionist models as models of psychological phenomena, Fodor and Pylyshyn propose that the proper role for connectionism is to provide an account of how classical cognitive architectures might be implemented in connectionist style devices, specifically the brain. Since connectionist devices are capable, at a functional level, of behaving in accordance with classical architecture, one can reject the possibility of a connectionist style psychology but employ connectionism (or something like it) to explain how the brain realizes a classical architecture In this way, one may still grant the virtues of some connectionist models and yet reject connectionism as a psychological hypothesis

§

The cognition - implementation distinction is grounded in the functionalist distinction between mental states and the particular physical stuff in which they are realized A functionalist mental state is delineated by the set of causal relations it bears to environmental input, bodily behavior and to other mental states. Mental states, like mousetraps, are functional kinds: they can be realized in different ways, using different materials The cognition half of the cognition - implementation distinction concerns theories that involve

---

14 Smolensky (1987. 137).

mental states, their manipulation, and so on, all without providing the details of their physical realization It is left to theories of implementation to give an account of the physical details of cognition

The significance of the distinction is found in the perceived aims of psychology Among its aims, we are to understand, is the elicitation of cognitive level theory Accounts of implementation are valued in the psychological literature, but usually only to the extent that they serve to distinguish correct cognitive level accounts from incorrect ones.

The majority of those working with distributed connectionist models are quite adamant that they are working toward an account of cognition; their models of particular cognitive abilities are meant to be interpreted at the cognitive level, despite occasional polemic to the effect that the boundary between cognition and implementation is not as clear as one might like it

> [T]he claim that our models address a fundamentally different level of description than other psychological models is based on a failure to acknowledge the primary level of description to which much psychological theorizing is directed. At this level, our models should be considered as competitors of other models as a means of explaining psychological data.[15]

We heard above an opinion from the classical camp on the potential for a connectionist style cognitive level theory. connectionism provides inadequate means for an account of cognition, and its best hope is in providing an account of the implementation of a classical cognitive theory. Two difficulties arise from this proposal for connectionism The first is less pressing: it involves a need to rewrite the literature of connectionism, editing out all claims concerning supposed cognitive level features of connectionism and of the potential for "connectionist psychology".

The second problem is more serious: if connectionism is to tell us something about the implementation of a classical (or any other) psychology, in what sorts of things are we to imagine the implementation? For what sorts of things would a connectionist account of implementation be appropriate? The immediate answer is that connectionism is best suited to give an account of implementation in connectionist devices. But surely this is of doubtful utility, there are simpler means for realizing a device that operates in a classical manner The hint here from Fodor and Pylyshyn is that connectionism may have something to tell us about how classical symbol-crunching goes on in the brain.

How much can a connectionist style account of implementation tell us about the brain? Mainstream connectionists (exemplified by McClelland, Rumelhart, Smolensky, Hinton, and

---

[15] Rumelhart and McClelland (1986b. 124).

so on) make much ado about taking "brain style processing" and the details of implementation seriously, but they do mean for their work in connectionism to be interpreted at the cognitive level; they mean, eventually at least, to be offering psychological hypotheses to compete with those from the classical camp. With regard to implementation, they claim only to be paying attention to it, in some sense, rather than providing an account of it. Thus, they come to say that they are engaged in "neurally inspired" modeling rather than the modeling of neurons

So it is perhaps the "neurally inspired" features of connectionism that might make it tempting as the means for an account of the implementation of some cognitive theory But one should recognize both that connectionism as practiced is meant to be interpreted at the cognitive level and that as a model of the nervous system, connectionism involves enormous simplifications For all of the doubts about the status of connectionist or classical accounts of cognition, the simplifying nature of connectionism can serve us well in at least one respect it can tell us something about the nature of networks of real neurons We have seen that the processing units involved in connectionism, while "neurally inspired", are not meant to be models of actual neurons. But we should not be deterred by the prevailing view of the aims of connectionism. Connectionism borrows many of its basic features from neuroscience, and it has the potential to return the favour by providing insights on how large groups of vaguely neuronlike processing units might operate We need not develop specific connectionist style neural models to learn the lessons connectionism might have for us; the vaguely neural nature of connectionist processing units means that many of the insights provided by work in connectionism may be applicable to neuroscience and, to an extent, vice versa

In drawing neuroscientific lessons from connectionism, we are to an extent viewing connectionist models as simplifying neural models. There are, in general, two strategies in modeling.[16] In the first, called "realistic" modeling, one incorporates as much of the known features of the object of modeling as is possible. One of course wants models that are as accurate as they can be, but the danger with realistic models is that, because of the breadth of variables and parameters included, their complexity and the computational demands of a computer simulation may leave the models as poorly understood as the things modeled There is a danger, in other words, of learning nothing. As well, the computational demands of realistic models insure that only the smallest networks of neurons can be modeled in this fashion

The other general strategy in modeling is to simplify Simplifying models ignore some features of the object of modeling, while emphasizing others In this way, one risks

16 Sejnowski et al (1990)

inaccuracy in order to overcome some of the difficulties and disadvantages of realistic models But with the risk of ignoring relevant features or incorporating irrelevant ones comes, with experimentation, the reward of the potential of simplifying models to distinguish relevant features from irrelevant ones.[17] It is probably only through models that simplify the operation of neurons that we can gain some understanding of the basic computational features and constraints of the nervous system.

There are many aspects of connectionism relevant to an understanding of the workings of real neurons The areas in which connectionism is most helpful are, not surprisingly, those in which it has borrowed some feature from neuroscience and built on it For example, one of the tenets of connectionist psychology is that learning is the modification of (synaptic) connection weights over time as a function of experience This idea goes back at least to Donald Hebb (1949) Hebb's basic idea was that when two neurons were simultaneously excited, the strength of the connection between them should be increased. In this way, a nervous system could modify itself as a result of the particular experiences it had (thus "learning") Early work in connectionism applied Hebb's ideas to the operation of its processing units, so that connectionist networks would modify themselves according to their input along the lines of Hebb's hunch about real neurons.

Connectionism borrowed the Hebb rule from neuroscience, but in return, connectionism is better suited for an investigation of its particular potential and limitations Toying with Hebbian connectionist networks has demonstrated, for example, that the Hebb rule is excellent for pattern association tasks but not so much for tasks unrelated to pattern association Work in connectionism has also lead to the development of more powerful "learning rules", as they came to be called; these are new formulae for the modification of connection weights: the understanding of learning as the modification of connection weights over time as a function of experience has not changed

It would be further possible, given the superficial similarities between networks of connectionist processing units and those of real neurons, to turn back to neuroscience to attempt to find real instances of the newer learning rules developed for work in connectionism As one might expect from the grand psychological aims of connectionist research, the emphasis in connectionism has not been on developing learning rules that might apply to real neurons. Some of the more popular rules employed in the training of connectionist networks involve a measurement of error, which requires a "teacher" to both

---

[17] For reasons of brevity, I discuss here only the "amount of information incorporated" conception of what makes a model realistic or simplifying. My immediate concern is with distinguishing two strategies of neural modeling and not with the epistemological committments of referring to one as "realistic" and another "simplifying".

know the desired result and measure the disparity between the actual and desired results. (Teachers are usually the experimenters themselves or the computer on which the network is simulated.) Although we often do learn things from our teachers, we cannot expect learning always to depend on a foreknowledge of the correct answer, especially at the most obscure corners of our nervous systems.

As it is with learning, so it goes for storage. Connectionist networks, among other things, give us the means to investigate the way in which a group of neurons might store information, and how they might come to retrieve and modify the information stored. Some of the more fruitful contributions to be made by connectionist theory involve an analysis of the operation of networks at the level of the entire network. This is not an area foreign to neuroscience, but does present it with the dual restrictions of space and complexity, neither of which presents much of a problem for connectionist networks.

There are two analyses of the global activity of connectionist networks, one involves an analysis of unit activity, the other the global configuration of connection weights. These are state-space analyses: the state of a particular network at a particular time is represented by a point in an abstract multidimensional space. Both these analyses, as it turns out, are important for learning the epistemological lessons that connectionism has to teach us. Those lessons, together with an extended discussion of both activation space and weight space, appear in the last chapter.

§

We have seen the doubtful nature of the claims made for the superiority of connectionist models over classical ones. The connectionist and symbol processing approaches are each well suited to different cognitive tasks. I have proposed that the real advantage connectionism may have over the classical approach lies in its potential as a tool for the advancement of neuroscience; but it is important to recognize the distance between this proposal and that of Fodor and Pylyshyn. Their proposal banishes connectionism from the cognitive realm and restricts it to providing implementation level accounts.

These two fates imagined for connectionism may seem identical, but in fact that is not the case. To propose that connectionism can play a role in a better understanding of the nervous system, at the level of neurons and groups of neurons, is not to restrict connectionism to the implementational level. Often neural level accounts are understood to be

ipso facto at the implementation level, but this is simply not the case; what counts as the level of implementation is a function of interest

Viewing the level of neurons or thereabouts as the level of implementation is to confuse levels of organization with levels of analysis. The nervous system has many different levels of organization, ranging from the system as a whole on down through circuits of neurons to the level of ion transport. There are as well different levels of analysis of the nervous system. Above we discussed the cognitive and implementation levels; there may also be pertinent subdivisions of these levels, if not ones in addition to these two, but it is primarily the functional - physical distinction that is important here.[18] It is a mistake to think that these two different ways of understanding the nervous system are somehow congruous Each level of organization of the nervous system can be viewed through functional or physical lenses.

Consider, for example, an action potential. If one is interested in communication between neurons, then the details of an action potential are implementational, since what matters to neurons is really only the presence of a binary event If the interest is instead at the level of ionic distribution, then an action potential is a functional sort of thing, the result of an integration of many sources of information. At any level of organization of the nervous system, one may ask functional or implementational questions.[19] To view the "neural level" as somehow more implementational than other levels of organization is to make a category mistake, as with other such levels, we can reasonably ask both functional and implementational questions.

Similarly, one should not confuse the cognition - implementation distinction with the distinction between simplifying and realistic models. Just as levels of analysis and of organization are orthogonal, so too are these two ways of understanding models. Realistic style models need be no more or less implementational than those that are more simplifying The standard view of connectionism, for example, is that it presents simplifying models meant to be interpreted at the functional (cognitive) level. But one can at least conceive of a

---

18 In what follows, as elsewhere, the discussion involves the cognitive-implementational distinction of neural level phenomena. I have chosen to flip between "cognitive" and "functional" for the quite benign reason that many of the functions carried out by small groups of neurons do not qualify as cognitive as that term is usually understood. The present significance of the distinction between cognition and implementation is that "cognition" is characterized functionally while implemenation is not. I have used the term "cognitive-implementational distinction" elsewhere in the discussion in the interest of conforming to standard terminology.

19 In this example, and in the point it is used to make, I am borrowing from Churchland, Koch and Sejnowski (1990). Both the point and the example reappear in Churchland and Sejnowski (1992).

cognitive level model that was meant to be a "realistic" model, or of models that were meant to be simplifying implementation level accounts, and so on.

Fodor and Pylyshyn properly understand connectionist models as simplifying, but propose that they be understood at the implementational level. I propose no such revision of connectionism; the work done in connectionist theory and on particular models is clearly has something to say at the cognitive level. The nature of connectionism is such that it may be able to provide some insights into the "cognitive" (ie, functional) level features of networks of real neurons, without being conceived of as accounts of implementation. As well, if we keep in mind the simplifying nature of connectionism (simplifying with regard to neural function), then we need not even view particular connectionist models as neural models in order to learn something of neuroscience from them. We can learn, as just noted, some of the strengths and weaknesses of the Hebb rule. It is the simplifying nature of connectionism in general from which we learn this lesson; we needn't have viewed any particular connectionist model as a neural model It is more the simplifying nature of the basic features of connectionism that matters here, rather than the intended domain of any particular model

Chapter two
Connectionism and folk psychology

Equipped with an understanding of the nature of connectionist models, we can now explore the possibilities for an epistemology consistent with eliminative materialism. Chapters three and four address the explicitly epistemological issues; the goal in this chapter is to establish a support for eliminative materialism from connectionism, by arguing for the inconsistency of connectionism and folk psychology. I will repeat here my position on the status of folk psychology discussed in the introduction: in order to ultimately escape elimination, folk psychology must be compatible with our best account of the activity of the nervous system.

Smolensky and others working in connectionism are clearly tempted by the view that the generalizations and categories of folk psychology ignore important features of cognition In the less philosophical connectionist literature, the debate involves neither folk psychology nor eliminative materialism What discussion there is on these matters concerns the extent to which classical and connectionist modeling differ, and how classical descriptions of connectionist devices may make some correct predictions about their behaviour, while at the same time being altogether wrong about the internal mechanisms it posits. Because they present no account of folk psychology, Smolensky and his allies are not committed to its rejection, but for those who see some connection between folk psychology and the classical symbolic view of cognition, they present a basis for argument for the elimination of folk psychology based on the success of connectionism. It is partly on this basis that William Ramsey, Stephen Stich, and Joseph Garon present their arguments, discussed below, for the inconsistency of connectionist and folk psychologies.

The first chapter discussed in detail the extent to which the distributed connectionism discussed by those such as Smolensky, Rumelhart, McClelland and others differed from the classical view of cognition and cognitive modeling. From the successes of connectionist modeling and with a great deal of optimism for its future, Smolensky and others conclude that connectionism is somehow more faithful to the details of cognition than the classical view They acknowledge the predictive successes of the classical approach, but they understand cognition as fundamentally connectionist To this end, this camp of connectionists sees the relation between classical and connectionist modeling as analogous to that between classical and quantum mechanics

> It might be argued that conventional symbol processing models are macroscopic accounts, analogous to Newtonian mechanics, whereas our models offer more microscopic accounts, analogous to quantum theory ...Through a thorough understanding of the relationship between the Newtonian mechanics and quantum theory we can understand that the macroscopic level of description may be only an approximation to the more microscopic theory. [1]

There are similar examples in the connectionist literature meant to present a similar view of the relation between classical and connectionist accounts. One finds also a number of lengthy analyses of connectionist models of particular cognitive tasks meant to demonstrate that such models nowhere employ the mechanisms that the classical view finds necessary. The point of the demonstration is not only to distance connectionism and classicism, but to show that the particular connectionist networks do not invoke the rules they seem to obey.

It is not entirely clear that we should understand classical mechanics as an idealization of a more realistic quantum theory, but it is best not to push the analogy too hard. The goal here seems to be to acknowledge the predictive success of classicism, while proposing that the processes it attributes to cognition are at base more complex than classicism itself recognizes. Some networks may behave as though they were manipulating sentence-like representations according to processes sensitive to their structure, but that does not mean that there are sentential representations in the network, or structure sensitive processes, or whatever else the classical view might see there.

> A system that has, at the microlevel, soft constraints satisfied in parallel, has at the macrolevel, under the right circumstances, to have hard constraints, satisfied serially. But it doesn't really, and if you go outside the Newtonian domain, you see that it's really been a quantum system all along. [2]

Frequently, the quantum theory analogy is accompanied by the proposal that connectionism provides a unifying account of behaviours that were previously divided into "rule governed" and "exceptional" cases. Connectionism is concerned with the underlying process that accounts for both sorts of behaviours.

It seems best to understand the quantum theory analogy as only an illustration of the relationship between the classical and connectionist approaches, rather than as the basis for preferring one over the other, although the analogy is clearly meant as part of an argument for

---

1 Rumelhart and McClelland (1986: 125). Rumelhart and McClelland credit the example to Smolensky, who first discusses it in his (1988).

2 Smolensky (1988: 20).

the superiority of connectionism over classicism. Chapter one discussed some of the ways in which connectionism recognized as complex some of the things the classical approach took to be simples But by being more detailed in this sense, it does not necessarily follow that connectionism is somehow unifying in a way that classicism is not. An argument of this sort will need more than analogy for support

But if it is unifying, underlying processes that we are looking for, why stop at the level of connectionism? Certainly a half decent physics will study the processes that underlie all cognitive behavior. What is needed here is some reason - putting aside "unification" and being more "fundamental" to the extent that we can - for preferring one cognitive theory over another It will not help connectionism to be somehow more fundamental and unifying if that means that it is therefore not an account of cognition

§

Since folk psychology has yet to appear in the debates between connectionism and classicism, whatever alliance there might exist between connectionism and eliminative materialism is all but hidden. But, as hinted above, it may turn out that a proper understanding of both connectionist and folk psychologies may uncover insoluble conflicts between them

Such a view of connectionism and folk psychology is advocated by William Ramsey, Stephen Stich, and Joseph Garon (1991). They claim that if several particular psychological hypotheses presented by connectionism turn out to be right, then the prospects for folk psychology seem rather poor. In their paper, Ramsey, Stich and Garon defend only this conditional claim; they defend neither connectionist psychology nor eliminative materialism Their arguments depend for the most part on the understanding of folk psychology that I gave in the introduction; we are to consider folk psychology and connectionism as competitors. While Ramsey et al do not explicitly discuss this view of folk psychology, it is evident that they hold it.

The conflict between connectionism and folk psychology occurs over an alleged feature of cognition that Stich calls "propositional modularity" (see below). Folk psychology, goes the argument, is committed to propositional modularity, while connectionism presents psychological models that do not employ it Connectionist psychology, we are to understand, does not involve propositional modularity, while folk psychology is committed to it. So if the

relevant aspects of connectionism turn out to be right, folk psychology must be wrong, and if folk psychology is correct, then it is connectionism that is doomed

Propositional modularity is a feature of a belief or memory store Such a storage system is modular to the extent that "there is some more or less isolatable part of the system which plays (or would play) the central role in a typical causal history leading to the utterance of a sentence."[3] Folk psychology seems to make claims of this sort concerning propositional attitudes: that they are functionally discrete states that play a causal role in the production both of behaviour and of other propositional attitudes

Propositional modularity springs up all over folk psychology Consider a typical belief-desire explanation: I opened the window because I desired to make the room cooler and I believed that opening the window would make the room cooler Accounts of inference are equally flavoured by modularity: If I believe "if your lights are on then you are home", and come to believe "your lights are on", then I will typically come to believe "you are home". Consider also the principle that people who sincerely assert "p" generally have "p" somewhere in their store of beliefs, or are at least capable of producing it sincere speakers of English who say "snow is white" generally have the belief that snow is white And so on for all of the different roles played by the propositional attitudes within folk psychology Note that the interrelatedness of beliefs is not at issue here, but rather their functional discreteness, propositional modularity is happily consistent with the view, for example, that acquiring one particular belief often leads to having a set of related beliefs

Ramsey, Stich and Garon's arguments depend also on a further, related feature of folk psychology. Folk psychological generalizations like the ones hinted at above (eg people who say "p" generally believe that p) are couched in terms of the semantic properties of the propositional attitudes they involve It is by virtue of being the belief that p that a particular belief has the profile of cause or effect that it does. So, for folk psychology, predicates of the form "believes that p" are projectable; they are predicates appropriate for use in nomological generalizations.

The view of connectionism that Ramsey et al present depends largely on the distributed nature of connectionist representations; their arguments do not apply to localist style connectionism. Their view that connectionism does not involve propositional modularity relies on two points First, they endorse the view that connectionist models properly understood, are cognitive level models and not accounts of implementation Second, they emphasize the distributed nature of connectionist representations

---

3 Stich (1983. 237-238) I shall relay here Stich's warning against confusing propositional modularity with the altogether different notion of modularity discussed by Fodor, notably in The Modularity of Mind.

Ramsey et al see distributedness as inconsistent with propositional modularity in two ways First, it is often impossible to localize a particular connectionist representation, at least beyond the input and output layers. Second, the mode of representation is such that it does not lend itself to modularity, there is no part of a network that can be comfortably seen as a symbol; "it is often plausible to view such networks as collectively or holistically encoding"[4] information The authors reinforce their position by discussing a simple connectionist model of their own They present a simple model of memory that judges the truth or falsity of a given set of propositions The propositions considered here are quite basic· "Dogs have fur", "Fish have scales", and so on The network takes the propositions as input (encoded across the sixteen input units) and outputs the judgment "true" or "false" by level of activity of the single output unit. A four unit hidden layer lies between the input and output layers

Trained up to give the appropriate responses to input, one can fairly say that the network stores information concerning the truth or falsity of the propositions that make up the input, insofar as it produces the correct response for the given propositions But there is no distinct part of the network that plays a role in producing the right output for any given proposition. The information stored by the network is distributed across all of its units, and each particular unit or connection weight is involved in storing information about many different propositions.

These are basically the same lessons about distributed representation and superpositional storage that we learned in the previous chapter. But notice how these features of connectionist models conflict with propositional modularity; there is no functionally discrete part of the network implicated in producing a response from a particular input proposition All of the units and all of the connection weights are involved in processing a single input vector As well, it is important to remember that, in accordance with the arguments of Ramsey et al, we are considering their network not as just a physical connectionist device, but as a cognitive model, making no claims as to its implementation To claim that a particular system is modular is to make a claim about its operation at a functional level, if the connectionist network under consideration were merely an account of the implementation of some cognitive model, then claims about its distributed nature would be of doubtful relevance Given the view of the status of folk psychology given in the introduction to the thesis, and given also the claim that folk psychology requires propositional modularity, the assumption here is that connectionist networks must have functionally discrete subparts that satisfy propositional modularity in order for connectionism to be compatible with folk psychology

---

[4] Ramsey, Stich and Garon (1991. 104).

As discussed earlier, connectionism invokes a distinction between representation and information storage; at least these are the terms that are commonly used to describe two distinct features of connectionist networks. It would seem that in this first argument for the inconsistency of connectionism and propositional modularity, Ramsey et al have glossed over this important, though perhaps for their purposes not crucial distinction. In making the above point about the non-modularity of their model, they have this to say:

> ...there is no distinct state or part of the network that serves to represent any particular proposition. The information encoded... is stored holistically and distributed throughout the network.[5]

The authors make a number of claims, as they seem to here, in which they apparently confound distributed representation and superpositional storage Indeed, the two claims in the above quote are quite distinct, although the argument proceeds as though they were equivalent. The second claim above is clearly about the nature of information storage, while the first is somewhat more ambiguously a claim about representation This quote is not unique in presenting this impression of the views of Ramsey et al; nowhere do they attempt to distinguish these two claims, nor do they leave the impression that any such clarification is necessary.

With regard to information storage, it should be fairly clear that nothing like propositional modularity is in operation in this case. All of the sixty-eight connection weights in the model are involved in producing the proper output from each particular input, the particular connections are differently weighted, but the weights do not change once the network is trained. The information necessary to produce the desired output is encoded in all of the connection weights; there is no isolatable subpart of all of the connection weights responsible for producing the proper output. It is not in the nature of connectionist networks to store information in specific locations, and so it is not in the nature of connectionist models to store information in a functionally discrete manner.

The status of connectionist representation is not as clear as that of information storage. Ramsey et al claim in the quote above that there is no distinct part of the network that serves to represent a particular proposition This claim is surely false, for that is exactly what the input layer of the network does It represents, in a vectorial code, the input proposition. The representation is of course distributed across the sixteen input units. But should that deter us from seeing the input layer as a functionally discrete subpart of the network that plays just the role that propositional modularity demands of it? The same units are involved in

---

5 Ramsey, Stich and Garon (1991, 108-9).

representing other propositions, but, unlike the superposition of information storage, the units represent only one proposition at a time, so appeals to distributedness will not keep propositional modularity from the door.

Surely Ramsey, Stich and Garon are aware of this feature of the input layer of their network No doubt they have ignored it because it is unclear that the nature of the input layer is enough for the network itself to satisfy the demands of propositional modularity. Since it is unclear, there is a burden of argument to demonstrate that the representational nature of the input layer, and it is a burden that Ramsey et al have not met. As one might expect, it is the role of the input layer of a network to encode the input to the network. If the network takes propositions as input, then the input vector (the pattern of activity of the input layer units) will represent a proposition. If not, not Apart from the nature of their input, there is no significant difference between connectionist networks that evaluate propositions and those that do not. So no significant conclusions about connectionist psychology are forthcoming from a network that takes propositions as its input.

The propositional nature of the input is, however, not really the issue here If the network we are considering took something else as its input, then questions about the status of the input layer would remain. The input layer (any therefore any representation therin) would still be functionally discrete, at least in the sense of being functionally distinguishable from the (representations in) the remainder of the network.

If there is a reason that the nature of the representation of the input does not justify judging the network as modular, it is perhaps that our concern with the network is with what happens after the input layer, and perhaps because there seems to be no alternative to representing input in a functionally discrete way. The input layer is just that: the area of the network reserved for representing the input. There is a sense in which it seems inevitable that the input to a system should be functionally discrete; the input has to be presented somehow, and there seems little option but to present it in a way that allows it to be functionally distinguished from information already in the network . So to the extent that input is both represented in this fashion and plays a role in the production of behaviour, propositional modularity is true - trivially true - of almost anything that represents and has an input-output profile.

It should be clear that propositional modularity is meant to be a substantive thesis about the nature of information storage, the functional structure of a system, and the etiology of behaviour The nature of a connectionist network is such that, beyond the input layer, it does not operate in a fashion that satisfies the demands of modularity. There is no functionally discrete part of the network of Ramsey et al in which the information needed for the proper operation of the network is stored. Apart from the trivial claims about input just

mentioned, there is no case in which an isolatable part of the system plays a central role in the production of the output of the network

Those hoping to find room for folk psychology in connectionism might want to argue here that the claims about the nature of the representation of input hinted at here are sufficient to viewing connectionist networks as modular But to the extent that we mean to have a substantive debate about the nature of connectionist processing, representation and storage, one is best advised to take a closer look, rather than ignoring wholesale the radically non modular features of connectionism

So Ramsey, Stich and Garon may be correct in saying that connectionism is non modular, but they reach their conclusion by ignoring some features of connectionist networks. Their second argument fares much better They introduce a second network, trained to judge the truth values of the same propositions as the first network, plus one more proposition Call the first network A and the second B Both A and B were trained up on the original sixteen propositions and produce the appropriate output for each, and both employ the same number of units with the same pattern of connectivity The difference between the two networks is that the training set for network B included one additional proposition; so it encodes information concerning seventeen propositions, although both networks are minimally capable of generalizing beyond their respective training sets

Networks A and B are quite similar, but where they differ, they differ substantially The two networks have the same pattern of connection of units, but no connection weight in B corresponds to any of the weights in A All of the connections in B are weighted differently than their counterparts in A, the two matrices of weights are not even similar. The two networks are given the same vectorial encodings of propositions as input, but it is safe to say that in the hidden layer - where we find the only units not involved in representing input or output - whatever representing is occurring is not done the same way in B as it is in A

In more traditional cognitive models, it is a fairly straightforward matter to identify what role is played, in a particular case, by the information about an added proposition But clearly this is not the case with the networks we are considering. We have seen how adding a proposition to the original training set makes a difference to the resultant network, but none of the differences between A and B qualifies as a functionally discrete or semantically interpretable state of the network Ramsey et al have already made this point about connectionist information storage in their first argument, but this feature of the difference between A and B permits a different strategy for arguing for the incompatibility of connectionism and folk psychology.

Both A and B store the information necessary to give the proper response "true" to the proposition "Dogs have fur" If we so desired, we could construct a third network and train it

on the original set of sixteen propositions, plus the one added to B, plus one more. This third network would also store the information necessary to judging true the proposition "Dogs have fur", and it would be as different from A or B as they are from each other. The point here is that there are indefinitely many connectionist networks that can store information regarding the truth value of "Dogs have fur". And while there are indefinitely many such networks, they have no projectable features in common that are recognized by connectionist theory The class of networks that (model an agent who) "believes that dogs have fur" or "remembers that dogs have fur" is, for connectionism, not a kind at all, but a disjunctive set Folk psychology, on the other hand treats the class of agents who believe that dogs have fur as a psychologically natural kind; it takes "believes that dogs have fur" as a projectable predicate Connectionism and folk psychology are incompatible, on our original, robust sense of compatibility.

It's not immediately certain that this difference between connectionist and folk psychologies will result in the two theories making different predictions about behaviour, even if it is clear that they will describe behaviour differently. But we are understanding folk psychology and connectionism as competitors, and the substance of the difference between connectionism and folk psychology makes it a safe bet that the entities and processes recognized by connectionism do not cohere with those recognized by folk psychology Further, as we discovered in the first chapter, the laws of connectionism are not intentional in character, again significantly unlike folk psychology.

It is again clear that Ramsey, Stich and Garon share the view of folk psychology that I advocated in the introduction to the thesis. The view of folk psychology that I advocated there proposed that, if we are to retain folk psychology, then the kinds (kind predicates) it recognizes must be roughly coextensive or smoothly reducible to the kinds recognized by the best account of cognition. So again my claims (as well as those of Ramsey et al) are conditional on this view of folk psychology.[6]

As they present both connectionism and folk psychology, the conditional claim that Ramsey, Stich, and Garon argue for seems well founded: if connectionism is correct, then

---

6 Those who find my conditions too restrictive may wish to side with Jaegwon Kim. Distinguishing between predicates and properties, Kim argues that psychological properties might be multiply realizable in eg, different species. In this way, Kim can grant that psychological states supervene on physical ones, but without making psychological properties reducible to physical ones (because of multiple realization). So one might want to employ Kim's view in order to claim that folk psychological kinds need not be reducible to kinds recognized by an account of cognition. However, Kim does grant that we might have "local reduction" of psychology, and eg, reduce human psychology to physical states, and that possibility might prove problematic for those who want to use Kim in this way. See Kim (1984)

folk psychology is not. This conditional aside, there are basically two broad areas in which Ramsey et al may have erred: in their characterizations of connectionism and of folk psychology.

I have already devoted much of the exegesis of Ramsey et al to constructing a coherent view of connectionism. But there is one feature of connectionism not mentioned above that we should not overlook in our search for propositional modularity. The proposal that Ramsey et al anticipate here is that the sorts of states that propositional modularity demands can be found in the hidden unit activation space of the network considered in their arguments.

Hidden unit activation space is an abstract multidimensional space used to analyze the manner in which the hidden layer organizes the information that it processes For the proposition judging network, each input vector produces a different hidden unit vector So on an analysis of the hidden unit activation space for the network, each input proposition given produces a hidden unit vector that picks out a point in weight space. So states of activation space are the functionally discrete, semantically interpretable and causally active states that we have been looking for."

It is here that Ramsey, Stich, and Garon finally exhibit an understanding of the distinction between connectionist representation and storage. They do not think, however, that those things normally referred to as connectionist representations are deserving of the name. Folk psychology, they argue, sees beliefs and propositional memories as things that endure; "...they are the sorts of things that cognitive agents generally have lots of, even when they are not using them."[8] An activation pattern, on the other hand, is a fleeting state of a network, and a particular hidden unit activation pattern will occur only when the network is given the appropriate input. Folk psychology understands people as having myriad beliefs, many of which may have been around for years. But one cannot say of a network that it has many activation patterns continuously over a long period; that is simply not how they operate. So if activation patterns are not the sorts of states that can count towards propositional modularity, then points in hidden unit activation space - which are merely activation patterns presented in a different manner - are just as bad off

The strategy Ramsey et al adopt here is to go beyond the hunt for functionally discrete, semantically evaluable and causally active states, and look further into a folk psychological account of belief in order to find other ways of keeping a safe distance between connectionism and folk psychology So Ramsey, Stich and Garon in fact leave open the

---

" For an extended discussion of activation space, and the related weight space, see chapter four.

8 Ramsey, Stich and Garon (1991: 114)

possibility that hidden unit activation patterns are the sort of states needed to satisfy propositional modularity, and this is a much larger concession than they admit. If they are correct about the features of connectionist representations versus features demanded by folk psychology, then their final conclusion regarding the incompatibility of connectionism and folk psychology is quite safe. But they specifically draw that conclusion because they can find no states in connectionist networks that satisfy propositional modularity.

We have already explored, in chapter one, the extent to which connectionist representations fail to satisfy the demands of classical cognitive modeling, one of folk psychology's ideological allies. But it may not be necessary to nitpick about what sorts of states, beyond those that satisfy propositional modularity, can satisfy the demands of folk psychology, because it is not immediately clear that hidden unit activation patterns are the sorts of states demanded by propositional modularity. Hidden unit activation patterns clearly play a causal role in the operation of a network, but their semantic evaluability and functional discreteness are not obvious.

Concerning semantic evaluability, the simplicity of the network we have been considering may be misleading. The hidden layer, in a very real sense, is the intermediary between the input and output layers. The hidden unit activation pattern, together with the weighted connections with adjoining layers, are responsible for the proper transformation from the input to output vectors. There is some sense in which one might view a hidden unit activation pattern (and therefore a point in hidden unit activation space) as a representation of, for example, "'Dogs have fur' is true", insofar as it is the only intermediate representation in a network that behaves as though it believed (or whatever) that claim about "Dogs have fur". That the network behaves in that manner, however, is not evidence enough to claim that the hidden layer activation patterns are representations of that sort.

If we look to more complex connectionist cognitive models, the suspicion that hidden layer vectors are functionally discrete or semantically interpretable should disappear. In models that have more than three layers, and therefore more than one hidden layer, it will be more difficult to conceive of a representational role for the hidden units. As well, some connectionist models, to varying degrees, lack the layered structure found in the network we have been considering, so whatever activation space analysis may be possible will be even more obscure than in more straightforward networks.

If these considerations are not conclusive, consider another. Many connectionist models take as input a somewhat wider range of perhaps more complex input than that of Ramsey et al. In these sorts of networks, the areas of hidden unit activation space relevant to understanding how the network proceeds from input to output will not be particular points, but rather broader regions. It is, for many networks, the partitions of hidden unit activation

space that are relevant to understanding what role the hidden units play between input and output.

It will require significant argument to claim that regions of hidden unit activation space play a functionally discrete role in processing There are, firstly, no guarantees that such regions will have distinct boundaries, so the role played by such regions may not always be clear. Given a hidden unit activation space with a number of partitions relevant to the processing done by the network, the complexity of the mathematics of attractors necessary for understanding the role of partitions in vector transformation may be such that it leaves little room or profit for talk of functional discreteness At the least, it is not at all obvious that regions of hidden unit activation space can count as functionally discrete states of a connectionist network, so someone who wished to defend the modularity of connectionist networks in this manner has considerable work to do The final chapter will consider further the functional role of activation space

As we saw above, Ramsey, Stich and Garon dismiss the proposal concerning hidden unit activation patterns by looking further into folk psychology and disqualifying the states considered as examples of propositional modularity. That brings us to the second possible area in which the authors may have erred: in their elicitation of commonsense psychology It may be, for example, that Ramsey et al are simply wrong when they claim that folk psychology is committed to propositional modularity. Perhaps it is possible to have a psychology that is faithful to the intuitions of the folk, but does not demand of the propositional attitudes that they be functionally discrete, semantically evaluable, and causally active.

One should note with caution the extent to which one must go in order to present an account of folk psychology consistent with the connectionism that Ramsey et al present There are many reasonable arguments that find propositional modularity underlying many of the folk's intuitions about belief, about the interactions of beliefs, inference, and the explanation of behaviour. Propositional modularity aside, remember also that connectionism does not recognize the same kinds as folk psychology, this is a genuine problem for their compatibility if we are to understand them as competitor theories None of the alleged inconsistencies between connectionism and folk psychology are problematic for the view that connectionism is good only for an account of the implementation of folk psychology, but that is because such a view does not consider connectionism as a cognitive level account, and so does not see connectionism and folk psychology as competitors.

In his rebuttal to Ramsey, Stich and Garon's paper, John Heil (1990) suggests that connectionism is quite consistent with a view of folk psychology in which beliefs and desires "function holistically". Apart from tipping his hat to Donald Davidson, Heil does little to

explain what functioning holistically means for the relation between propositional attitudes Here Heil has two problems First, it is not at all obvious that functioning holistically is inconsistent with propositional modularity. Sentences, propositions, and things in general can be interrelated and logically interdependent and yet still be functionally discrete in just the sense we have been discussing. So at the very least Heil has more argument to give on this issue

Heil's second problem here is that he seems to have equated "functioning holistically" with the global, superpositional encoding of information in connectionist networks. It is in this equivalence that he finds consistency between connectionism and folk psychology, but the relation between the two is entirely obscure. Whatever functioning holistically may entail, its relationship to connectionist information storage is especially strained if we consider connectionist networks that are concerned with input other than propositions Does a distributed connectionist network that globally stores the information necessary for distinguishing squares from circles and triangles therefore function holistically?

In closing, I should note some dissent here concerning the commitments of folk psychology. I noted in the introduction to the thesis that I was considering folk psychology to be in competition with other accounts of cognition. I noted also that if folk psychology was to ultimately escape elimination, it must prove to be compatible with our best account of the activity of the nervous system This is not the universal view of folk psychology and its fate There are some, for example Dennett, who acknowledge that our best account of cognition is unlikely to posit anything that looks like beliefs or desires, yet for various reasons think that it would be mistaken to therefore reject folk psychology

On such a view of folk psychology, the arguments of Ramsey, Stich and Garon are of doubtful relevance If you don't think that folk psychology has to be compatible with an account of cognition, then it should not really matter to you that folk psychology is incompatible with a connectionist understanding of cognition. I note this view without arguing against it: the gap that lies between my understanding of folk psychology and this view is altogether too large to fill here. Suffice it to say that there is substantial disagreement here about the nature of folk psychology and about the philosophy of science, and that on either side of the gap there need not be complete agreement on related issues.

Note however that we can have an understanding of cognitive modeling on which the aim of connectionist models is the description of behaviour, with no mention of the mechanics of cognition Recall that connectionism and folk psychology recognize different kinds; "believer", for example, is a disjunctive set for connectionism. As competitor explanations of behaviour, the conflict between connectionism and folk psychology remains However, this too involves sneaking in assumptions about the status of folk psychology and

of the philosophy of science. But it is worth noting that we need not restrict ourselves to discussing the details of the nervous system in order to find some difficulties for folk psychology in connectionism.

If connectionist psychology is inconsistent with folk psychology, then of course connectionism is consistent with eliminative materialism. As well, evidence for connectionist psychology is therefore evidence for the claims of eliminative materialism So whatever epistemological claims there are to be gained from connectionism, they are claims that eliminative materialists should pay attention to, even though the direction of implication is such that they need not think the claims are correct The mere prospect of an epistemology that owed nothing to the categories of folk psychology should be enough to arouse interest

Chapter three

Interlude.

Naturalized epistemology and the proper relation between epistemology and psychology

The goal of this chapter is to establish the relevance of psychology for epistemology, via a discussion of naturalized epistemology. Naturalized epistemology is the scientific study of cognition, perception, learning, and anything else that may tell us what it is to know, and how we come to know what we know. While many welcome the programme of naturalization, there is disagreement over the extent to which a naturalized epistemology can satisfy the demands we have always made of epistemology. The different views within the naturalistic approach are based on, among other things, different views of the relevance of psychology for epistemology. Quine, for example, thinks that psychological questions hold all of the content of epistemological questions, while Alvin Goldman thinks that there are epistemological questions that psychology cannot answer. Both Quine and Goldman, to the extent that they see an epistemological role for psychology, welcome naturalization. But Goldman does not think that epistemology can be completely naturalized; he thinks that there are epistemological questions that cannot be answered by empirical inquiry alone.

Since the goal in this chapter is to defend the relevance of psychology for epistemology, my emphasis will be on defending the entire programme of naturalization, rather than defending a particular position within the naturalistic camp. I will however reveal my Quinean biases and argue against the view that epistemology poses questions that psychology cannot answer.

§

Epistemology has been traditionally conceived of as an a priori enterprise. Further, the empirical details of psychology were thought to be largely irrelevant for epistemic inquiry. It was permissible for science to inquire about descriptive psychological details - they could study the psychology of knowledge acquisition and of belief perseverance and so on, but epistemology proper, viz, normative epistemology, was strictly the philosopher's domain. Psychology can tell us how we come to know things, but if we also want an account of how we <u>should</u> come to believe, what the <u>best</u> means of belief acquisition is, then we have

gone beyond the realm of the descriptive epistemology that psychology gives us Normative epistemology tells us how belief acquisition should proceed, and so it is here, and not to psychology, that we look for epistemic guidance This traditional view of epistemology leaves it in somewhat the same situation as ethics Just as knowing how people act will not alone tell us how people should act, so too is knowing how people acquire knowledge insufficient for knowing how people should acquire knowledge. Empirical inquiry simply can't tell us everything we want to know. We need to go beyond the empirical if we hope to have a normative epistemology, or a normative ethics, and so on Since actual practice and justified practice need not be the same thing, looking at actual practice will not tell us what justification is

Recently this traditional view of epistemology and its proper relation to empirical science has come into question Some have proposed that empirical science (specifically psychology) is relevant to epistemology, while others argue that psychology can answer all of the questions that we previously expected only from a priori epistemology. The relevance of the alleged a priori status of epistemology is the way in which it conflicts with the hope for an empirical epistemology, and the ways in which empirical and a priori investigation differ I take it that the difference between the two styles of investigation is just the difference between the laboratory and the armchair. Both approaches endeavor to answer the same basic epistemological questions, but in different ways. The present debate finds its beginnings, largely, in Quine's "Epistemology Naturalized" (1969). What follows is a brief exegesis of Quine's view, my argumentive emphasis in this chapter lies in the next section, which concerns the proper relation between psychology and epistemology.

In his (1969), Quine argues not just for the relevance of psychology for epistemology; his position was that epistemology could be replaced by psychology. On his view, all genuine epistemological questions turn out to be psychological questions Quine expresses the motivation for such a view like this: all of the evidence that anyone has for their understanding of the world is the stimulation of their own sense receptors So why should we not settle for psychology to answer our epistemological questions?

Quine is often described on this issue as advocating the replacement of epistemology with psychology, but this characterization is not entirely fair Quine does not advocate the death of epistemology. Epistemology is a field of inquiry, a series of questions; Quine's proposal is that psychology (and perhaps additional relevant sciences) can give us the sort of answers (and questions) that we have generally expected, or should expect, of epistemology It is not that epistemology is replaced, it is rather that we locate epistemology within empirical psychology:

> .epistemology still goes on, though in a new setting and a clarified status Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a natural phenomenon, viz, a physical human subject. This human subject is accorded a certain experimentally controlled input - certain patterns of irradiation in assorted frequencies, for instance - and in the fullness of time the subject delivers as output a description of the three-dimensional external world and its history. The relation between the meager input and the torrential output is a relation that we are prompted to study for somewhat the same reasons that always prompted epistemology; namely, in order to see how evidence relates to theory, and in what ways one's theory of nature transcends any available evidence.[1]

For Quine, our belief in the external world is a hypothesis, we posit the existence of physical objects based on the data of our senses The ordinary human situation is therefore very much like that of the scientist the difference between ordinary and scientific positing is only that the former is "archaic", "unconscious", and "shrouded in prehistory".[2] Just as atoms and molecules are posits, so too are tables and chairs. The epistemological question concerning the proper relation between our picture of the world and our sensory data is, properly understood, a psychological question, answerable by psychological means.

There are a number of perceived problems with a naturalistic epistemology that have previously kept psychology and epistemology in separate arenas The first concerns a problem of circularity: if epistemology is concerned with validating the foundations of science, then it cannot look back to the very same science to validate itself Quine's view on this point is that the alleged circularity should not worry us His argument to this end is somewhat nebulous; he paints the history of epistemology as largely that of the foundationalist programme. The problem Quine wishes us to see in the foundationalist programme is not simply that foundationalists did not have the correct epistemological answers, but that they were asking the wrong sorts of questions. He urges that, in recognizing the stagnation of the foundationalist programme, we should also recognize that the only legitimate epistemological questions are psychological ones. "Once we have stopped dreaming of deducing science from observations," says Quine, "such scruples against circularity have little point".[3] Quine is characteristically enigmatic on the reason for granting this point, but part of the problem is the lack of a sense data language Because we cannot characterize our observations in a theory neutral way, building science up from observation is inevitably circular. The point then seems to be that we should embrace the circularity that

---

1 Quine (1969: 24)

2 Quine (1960: 22).

3 Quine (1969. 19).

epistemology presents us; we get epistemology that is continuous with science, rather than building science up from neutral observation. We are not building pyramids, but rather repairing the same raft on which we stand, to recount Neurath's metaphor. Quine's naturalistic epistemology is contained within science, as a branch of psychology. But previous relationships between science and epistemology remain: it is still the task of epistemology to study the relation between theory and data, between science and observation. So the relation between epistemology and science is not one of circularity but rather one of mutual dependence. The goal is to understand the "institution" or "process" of science, and not to obtain an epistemology that is any better than the science that is its object."[4]

Another traditional barrier to naturalizing epistemology is the perceived distance between descriptive and normative epistemology alluded to previously. Here the problem is not simply the traditional distance between psychology and epistemology that we have been discussing, but further the difference between descriptive and normative claims. "Is cannot imply ought," goes the familiar slogan: the problem here is that things are often not as they should be, and that the way things are does not itself indicate how things should be.

Returning to the analogy with ethics, most will agree that people often do not act as they should. So if we want to know how people should act, we will need something more than an understanding of how people do in fact act. So for epistemology, the problem is that the way people do acquire knowledge does not by itself tell us how we should come to acquire knowledge. We expect of epistemology an account of the way our epistemic lives should be, not merely how they are: psychology can tell us only the about the latter.

The short way to deal with this objection is to note that epistemology must have some relation to our actual abilities to be of any use to us. An epistemology that makes recommendations that humans cannot follow is of little use to humans; we cannot expect to employ an epistemology that advises the use of cognitive equipment we do not possess. It is of benefit to normative epistemology, therefore, that it be kept in check with our descriptive understanding. Most in the naturalistic camp seem to go further than this in their understanding of the relation between descriptive and normative claims. The standard naturalistic approach to this problem, and the one Quine seems to be hinting at, is to argue for the view that the processes by which we ought to acquire beliefs are the processes by which we actually acquire beliefs. With Kornblith (1985b), I will refer to this view as "psychologism." If psychologism is correct, then the distance between is and ought disappears.

---

4 Quine (1969: 24).

It is important to note that believing psychologism, as Kornblith and I use the term, does not commit one to thinking that people always reason perfectly. We can make a distinction, as far as reasoning is concerned, between competence and performance. On this move, psychologism becomes a thesis about competence, and apparently non ideal reasoning can then be explained as errors of performance. Given this move, psychologism is quite consistent with errors in reasoning, so long as such errors can be believably cast as errors of performance

Quine's position concerning epistemology seems to commit him to psychologism Suppose psychologism were false. Then, ex hypothesis, there are epistemic processes that we ought to be using but are not (or there are processes that we are using but ought not to be) But if this were so, empirical psychology could not hope to replace epistemology outright, as Quine says it must, because psychology is impotent to discover processes that we do not in fact use. So Quine's insistence on replacement commits him to psychologism

We should note however that there are a number of different views possible within the naturalistic approach concerning the relation between psychology and epistemology While replacement does imply psychologism, the converse does not hold: one might argue for psychologism without therefore being committed to replacement. Psychologism allows for strong links between psychological and epistemological concerns, but it does not follow that if psychologism is correct then we can simply "read off" epistemology from psychology For psychologism says nothing about the nature of epistemological and psychological inquiry or of the content of the questions they ask, and it is over these issues that much of the debate within the naturalistic approach occurs. One might, for example, argue for psychologism but understand epistemology as an autonomous enterprise, with its own subject matter and its own questions distinct from psychology On this view psychology and epistemology are two different means of achieving the same goals. Psychology and epistemology study the same phenomena, but employ different approaches

In his taxonomy of positions within naturalistic epistemology, Kornblith (1985b) refers to this weaker view of the relation between epistemology and psychology as the "weak replacement thesis" But this view does not promote replacement; it grants psychologism and grants also that psychology and epistemology study the same processes, but it understands them as distinct fields. A correct epistemology will describe the same processes as a correct psychology, but in a different manner. One way in which they may differ, as Alvin Goldman argues, is in their level of description: it may be that some epistemological questions require answers that are at a different level of generality than empirical psychology can provide

What divides Quine's replacement view from this weaker view concerns the autonomy of epistemology Quine holds that empirical psychology can answer all legitimate

epistemological questions psychological questions hold all of the content to be found in epistemological questions. The weaker view we have been discussing holds that there are epistemological questions that in principle cannot be answered by psychology, which is to say that there are epistemological questions distinct in content from psychological ones. The difference between the two views concerns the autonomy of epistemology. The weaker view sees epistemology as an autonomous enterprise, while Quine and those of his ilk are radically opposed to autonomy. (In the literature covering this debate, epistemology is considered autonomous if it can pose one legitimate question distinct in content from the questions of empirical psychology.)

§

Let us now consider Goldman's arguments concerning the autonomy of epistemology and the proper relation between epistemology and psychology. My concern here is not with the entirety of Goldman's epistemological views, but rather with his particular arguments concerning the issues presently before us, as they appear in his 1985 article, "The relation between epistemology and psychology" (Similar points appear in his Epistemology and Cognition (1986).)

Goldman reasonably proposes that the correct rules of epistemology will be that total set of rules such that conforming with it would maximize the attainment of epistemically valued ends. He finds ample room for psychology in epistemological pursuits and finds much psychological content in epistemological questions. But he does not argue for replacement; he instead finds that there are epistemological questions for which psychology has neither an answer nor an equivalent question.

Goldman finds three reasons for thinking that epistemology will never be fully absorbed by empirical psychology, because the latter cannot address all of the questions asked of the former. First, psychology alone cannot tell us what the "right-making characteristic" is for epistemic rules, where we understand epistemic rules as involving cognitive processes. A right making characteristic is that feature of (a set of) rules that makes them the right ones to use. Psychology can tell us what sorts of cognitive processes we have access to, but it cannot tell us what feature is shared by the "right" set of rules involving those processes; it cannot tell us what it is that makes the right set of rules right. So even if psychology can tell us which processes or combination of processes are the right ones to use,

it cannot tell us what characteristic they (or rather the rules concerned with them) share that makes this so

Second, if - as many might suspect - the right making characteristic involves truth and falsity, then we have again gone outside the realm of psychology On its own, psychology cannot tell us which cognitive processes best promote ends involving these sorts of characteristics

Third, the correct set of cognitive operations will not be obvious even when given a complete account of elementary operations. Ingenuity will be needed to design optimal combinations of operations with an eye to promoting the ends desired Goldman's view here is that the complexity of this task - determining the right set of cognitive operations from a base set of available elementary ones - is too complex for psychology alone, and would need to involve a mix of disciplinary contributions, "both logical-philosophical and psychological".[5]

With regard to the third point, epistemology and psychology are indeed both complex, and certainly there may be relevant contributions to be had from outside the usual domain of psychology Linguistics and computer science may have things to tell us about cognition, along with other fields Logic will indeed be involved if we are to set about investigating inputs and outputs and combinations of processes. But none of this goes against the understanding of epistemology as an empirical pursuit, not even if we allow room for "philosophy". Quine himself made no demands that a naturalistic epistemology involve only that work now done in psychology departments. If Goldman's point here is that the complexity mentioned is such that not all of the old epistemological questions can be answered or recast by a naturalistic epistemology, then that is another matter. But there is no obvious link between the complexity of cognition and the autonomy of epistemology; if there is such a connection, Goldman does not present it for our consideration

The most serious concern here is Goldman's worry about right making characteristics Certainly if there is any sense to talk of the "right set of rules", it is a reasonable epistemological question to ask what it is that makes one set right and another not. And psychology, while it is suited for the investigation of cognitive processes, would seem to stumble if asked to go beyond an analysis of processes or of their outputs and explain what makes some outputs special and others not.

We should first note that Goldman does not seem to accept psychologism, judging from his discussions concerning how we are to pick out the cognitive processes that we should be using from among the ones available to us Certainly arguments against

---

5 Goldman (1985 56)

replacement will be much easier if we do not allow psychologism. But since Goldman's objection here concerns the nature of a particular type of epistemological question, and since he makes a common complaint about the relation between psychology and epistemology, let us see how his objections fare both with and without psychologism For, if epistemology is such that Quine's vision of naturalization cannot be realized given psychologism, it plainly cannot be realized without it.

It is easy to overemphasize the significance of psychologism in this context Recall that psychologism is a thesis about which psychological processes we should be using, and not a claim about the relation between the disciplines of epistemology and psychology One might see fit to draw significant conclusions about the relation between epistemology and psychology from the thesis of psychologism, but such conclusions are not forthcoming without suitable argument and appropriate additional claims. Psychologism is of course important, if psychology is to tell us anything about how which cognitive (epistemic) processes we should be using.

Let us turn now to Goldman's objection about right making characteristics, and grant psychologism. Ex hypothesis, the cognitive processes we are using are the ones we should be using. So psychology, since it means to tell us what cognitive processes we are using, can also tell us what processes we should be using. Goldman's epistemology consists of rules that involve cognitive processes, he is undecided as to whether or not all of the rules of epistemology are concerned with cognitive processes Given the nature of psychology, and understanding that it can tell us (ultimately) what cognitive processes we should be using, can it also tell us what the rules of epistemology are? All of the processes together will not determine what the rules are, particularly since the rules may mention more than cognitive processes We do not have to give ourselves the impossible task of building up "the right set" of rules from the cognitive processes that they involve, in order for psychology to answer all of our epistemological questions. Given a rule along the lines of "in this situation and given these ends, use these processes", it is not immediately obvious that we have somehow gone outside of the domain of psychology. If we are to determine a correct epistemology from psychology alone, certainly we are going to need psychologism or something like it But to claim here that we can learn all that there is to learn about epistemology by doing psychology would be to go beyond psychologism and make some assumptions about the nature of epistemology So we should avoid begging the question by first hearing Goldman out about the domain of epistemological questions versus psychological ones

We should note beforehand that it is unclear that there is a problem to be found here, if we grant psychologism. If epistemology is to outline the processes we should be using (by studying the ones that we do use) then it is at best unclear what role there is for epistemic

rules in addition to an account of the proper epistemic processes. That issue aside, however the problem of what counts as the right epistemology falls away With psychologism, the right epistemology is the one that we use, there is no mystery as to which epistemology is the right one to use

Determining a right making characteristic is necessary, Goldman tells us, for choosing the correct epistemic rules from a sea of possible ones One might also say, as Goldman does not, that aside from its utility in determining rules, we can reasonably ask of epistemology that it tell us what it is that makes epistemic rules the right ones to use It is here that psychologism will fail us; psychologism alone will not tell us what it is that makes the processes we use the right ones, psychologism is not a thesis about what makes the rules right

Recall that Goldman claimed that determining the right making characteristic is beyond the reach of psychological questions. Goldman's view here seems to be that psychology can ask different questions about the nature of cognitive processes, but questions that go beyond the domain of cognitive processes cannot be psychological questions Specifically, questions concerning the rules of epistemology, which mention cognitive processes, cannot be completely addressed by psychology The role of psychology is to determine which cognitive processes we use, but questions about epistemic rules are not entirely psychological in content. Further, questions about the right making characteristics for epistemic rules, since they are questions <u>about</u> epistemic rules, are twice removed from the concerns of psychology.

Goldman's claims regarding the limits of psychology are crucial to his views of the proper relation between epistemology and psychology, but he leaves them undefended Why exactly is determining the right making characteristics of epistemic rules too large a task for psychology? Goldman stipulates that such a task is beyond the reach of psychological inquiry, but why should we think that is so? We need here to understand just what the boundaries of psychology are, we need to know what makes particular questions psychological ones (or not) We need to know, in other words, what psychology is

We already know, for example, that psychology is concerned with human behaviour and not stellar evolution. There are clear cases of psychological and non-psychological questions, but our present problem is that some questions are not clearly psychological or non-psychological. In a sense, we already have a clear understanding of what psychology is: it's whatever is studied by people in psychology departments But with the aim of arguing against Goldman's claim regarding the limits of psychology, this ostensive understanding of psychology does not present us with a solution to our problem. Rather, we have a new name

for it how far can psychologists stray from the questions they presently (or usually) ask and remain psychologists?

It may help here to consider some other aspects of philosophy as they relate to psychology For example, the issues concerning the nature of mind, as understood by dualists, behaviourists, functionalists, and so on - is this not a psychological debate? It is a debate about psychology, no doubt, but what reason is there to say that this is a debate that psychologists cannot enter into without straying outside of their field, or without being cast as overly "philosophical"? The division between disciplines is at best fuzzy, particularly when philosophers are involved, because they often have their noses in other people's business, and since many have abandoned the view that there is no empirical content to philosophical talk

There are many different reasons for the present divisions between disciplines, understood as the divisions between university departments The world is not compartmentalized in the way that university departments are; it could be, therefore, that the divisions between psychology and epistemology are a result only of fuzzy borders and not of the content of their respective questions. It is here that Goldman faces a dilemma He needs an argument that proceeds from the proper division between disciplines (based on an understanding of the nature of psychology) to the autonomy of epistemology. But if such an argument is to be of any interest, it must find some tangible difference in the content of the two fields; and if it is difference in content that justifies the original premise regarding the divisions between disciplines, then we have come, viciously, full circle, if the intention is to argue for the autonomy of epistemology. It seems that Goldman has already decided the issues in the baggage he has brought to the debate. By imposing undefended limits on the domain of psychological inquiry, Goldman has already committed himself to the autonomy of epistemology. Of course, there may be a good argument to be had for the autonomy of epistemology from the nature of psychology. But what Goldman needs is some justification for his understanding of psychology, and perhaps through that there might be a non-circular argument for the autonomy of epistemology. He needs to do more than just stipulate that psychology cannot be normative in character. There may of course be an understanding of psychology on which it cannot play a normative role, but Goldman's makes no connection between some understanding of psychology and his largely negative claims concerning its domain.

The other horn of this dilemma is the possibility for an argument for the autonomy of epistemology based on some reason for the divisions between disciplines that does not involve the content of their respective questions But it is the content of questions that is central to the debate concerning the autonomy of epistemology If the divisions between

disciplines is a result only of, say, the need to divide up office space, it is hard to see how that could be relevant to the present debate

One should not think that this dilemma is uniquely Goldman's problem. It is faced by anyone who hopes to present a view of epistemology based on assumptions about psychology The debate concerning the autonomy of epistemology and its proper relation to psychology is as much a debate about psychology as it is about epistemology, so it will not do to simply stipulate what the limits of psychological inquiry are. The only way out of a vicious circle in this debate is to defend one's views about psychology

Granting psychologism, Goldman's negative conclusions about the relation between psychology and epistemology are cast into doubt. Goldman himself does not rely on psychologism, but I have involved it in the discussion of his views, for he makes a complaint common to the camp that disagrees with Quine but welcomes the relevance of psychology for epistemology the complaint that epistemology and psychology are at different levels of generality, and so epistemology asks some questions that psychology cannot answer. Within that camp one may find those who grant psychologism and those who do not. Without psychologism, as we have already seen above, the fate Quine imagines for epistemology cannot be realized.

§

As we can see, the problems raised by the naturalistic approach are not easily solved One of the questions raised by the debate concerning naturalizing epistemology involves the relevance of psychological data for present work in epistemology; but there are also questions concerning the relation between the final complete psychology and the final complete epistemology. Our present understanding of psychology and epistemology makes the second question much more difficult to answer than the first, no matter how interrelated they may be.

Given our present understanding of the two fields, there are reasons - discussed at the end of this chapter - for thinking that there is room for psychological data in epistemological theory, even given a traditional understanding of epistemology. But with regard to a completed epistemology and a completed psychology, we have nothing that looks remotely like either of them, how are we to decide the relation between them? There are, it seems, no candidates for a "complete" science of any kind, so what sense is there to talk of the proper relation between two completed sciences?

This is a caricature of sorts, but it is not without its merit. There is of course a long history of epistemological and psychological inquiry, and that there is any debate at all on the proper relation between them is evidence of some understanding of what we expect from both of them. And talk of a "complete" science is of course a tool meant to help in deciding the present status of disciplines. The point here is that a debate concerning completed psychology and epistemology threatens to exhaust our understanding and our capacity to decide the issues; if my intuitions regarding the domain of a complete psychology differ from yours, then there is little we can do but agree to disagree. We can be honest about the baggage that we bring to the debate, but there is only so much baggage that one can be rid of

If we are to decide against Quine concerning the non-autonomy of epistemology, we need to be shown an epistemological question distinct in content from one that can be asked by psychology, or at least be convinced that such a question exists. The work here is not so much finding such a question, but in establishing its non-psychological nature. Goldman purports, above, to have presented such a question (the one concerning right making characteristics); I hope to have demonstrated how his view depends on undefended assumptions about psychology. For every such question proposed, there is a debate to be had concerning its being distinct in content from any psychological question

With regard to psychologism, there is ample scientific and everyday evidence that people often reason in ways that are not as good as they could be; but one can still defend psychologism by casting errors in reasoning as performance errors, rather than evidence of an imperfect competence. This is a bigger programme than it sounds, and there is not room enough to begin it here; suffice it to say that it is not an unreasonable defense. It seems that the status of psychologism is a matter for further conceptual and empirical investigation. If the best account of psychology and epistemology includes psychologism, then so be it In this way, the most profitable direction to take in order to properly decide on the relation between epistemology and psychology is to keep on doing epistemology and psychology, and hope to pull ourselves up by our own bootstraps

It seems to me that Quine is correct on these matters, but I do not have the room to undertake an adequate defense of his views, which would, as with Goldman, involve defending a host of psychological assumptions. We have already seem some of Quine's assumptions at the beginning of this chapter. If, however, in the process of investigating the nature of psychology, we uncover a legitimate epistemological question that psychology cannot address on its own, then so much the worse for Quine Just as with the issues of autonomy and psychologism, the extent to which we can or should "naturalize" epistemology is a matter for further inquiry; and inquiry not only of the conceptual sort, but of the empirical, scientific sort as well.

Psychologism and the autonomy of epistemology are the two major issues dividing those who agree on taking a naturalistic approach to epistemology. Quine, as we have seen, is committed to psychologism, and fervently denies the autonomy of epistemology If Quine's arguments concerning epistemology are sound, then he gets psychologism for free. But one might argue the other way around, from psychologism to the non-autonomy of epistemology, and certainly there are those who have argued for psychologism (if never using that term) for reasons other than worries about the autonomy of epistemology.[6]

A second position within the naturalistic approach admits psychologism but understands epistemology as autonomous. This view was discussed above in considering a third position, the one inhabited by Goldman. This view denies psychologism and argues for the autonomy of epistemology. As defenders of the naturalistic approach, Goldman and others in this camp see much relevance for psychology in epistemological pursuits, but it should be noted that to deny psychologism and propose the autonomy of epistemology is also the preferred position the "traditional approach", which sees no place for psychological findings in epistemological work There is a fourth position possible in this spectrum of views - to deny both psychologism and the autonomy of epistemology - but that does not seem a viable position

A few closing considerations may serve to justify the project of subjecting epistemology to empirical analysis, no matter what one might think of the arguments above; there are significant reasons for recognizing the relevance of empirical data for epistemology even when it is understood in a traditional manner First, there is room for a healthy relationship between psychology and epistemology without psychologism. One might understandably suspect that not all of our epistemic processes are ones that we should be using. If we grant instead that the processes that we in fact use are roughly like the ones we should be using then there will be some useful contact between the disciplines. The contact possible here is clearly not as rich as that given psychologism, but it does allow for the possibility of psychologists and epistemologists each discovering something of significance to the other, and so for the mutual relevance of psychology and epistemology

Second, and more significantly, there is reason to believe that the traditional view of epistemology as a strictly a priori discipline is not inconsistent with recognizing the relevance of psychological test for epistemology The reason, as Kornblith (1985b) tells us, is quite simple: a priority does not imply obviousness. The disciplines often cited as examples of a priori knowledge (eg: mathematics) are often quite difficult. So subjecting a priori claims to testing can only help determine the correct account. How do we test a priori claims? By

---

[6] The concerns here include intentional ascription, translation, and evolution: those concerned include, for example, Dennett See Stich (1984)

conjoining them with relevant empirical and theoretical claims. If a desired logical result consistently fails to appear from such a conjunction, we might come to accept the a posteriori claims while rejecting the a priori claims

Kornblith's example to this end concerns the theory of probability, granting for the sake of argument that it is knowable a priori: Hilary wants to start a life insurance agency In order to determine how much he needs to make a profit, he needs to make the appropriate actuarial calculations. He does so by writing out the relevant part of the probability calculus and gathering data about mortality rates. Data in hand, Hilary goes out to sell policies, but eventually loses a great deal of money. He may have simply been unlucky, or he may have erred in determining mortality rates or made a trivial error in calculation. But it is also possible that he erred in his formulation of the theory of probability, certainly not an unusual mistake.

The moral of the story is that empirical test is uniquely suited for discovering errors, even errors made in the process of an a priori, non-empirical armchair investigation It is possible that further a priori investigation would have uncovered Hilary's mistake, but that is little reason not to subject it to a posteriori test. Some may consider it heretical to suggest that a priori principles can fall victim to empirical testing Kornblith's example is indeed not without epistemological baggage of its own, but no more than Duhem's point that an unexpected test result does not falsify any one particular claim but rather a group of theories "It is not that we propose a theory and Nature may shout 'no,'" said Lakatos, "rather, we propose a maze of theories, and Nature may shout 'inconsistent.'"[7]

So, on even the most traditional view of the nature of epistemology, there is room for empirical data. We will not, on this view, be able to "read off" our epistemology from our psychology, but that too is the case for many of the different understandings of a naturalized epistemology. Only in the most ideal, Quinean world can we hope to do anything of this sort Whatever the debate among those who advocate naturalizing epistemology, there is agreement amongst them all concerning the relevance of psychological data for epistemology And there is reason, as we have seen, to think that a more traditional view of epistemology is not necessarily inconsistent with this relationship between epistemology and psychology So let us turn now and see what a connectionist understanding of cognition can tell us about epistemology.

---

[7] Lakatos (1970. 130).

Chapter four

Conclusion Features of a connectionist epistemology

We should now have sufficient background to address the goal of the thesis, which is to explore, with the assistance of connectionism, the prospects for an epistemology consistent with eliminative materialism There are two related roles for connectionism in this drama The first has already been played out, it involves establishing a link between connectionism and eliminative materialism. As well, to the extent that such a link can be established, then evidence for connectionism is, ceteris paribus, evidence for eliminative materialism

The second role for connectionism in our inquiry is in presenting the means for an account of epistemology. Connectionism allows us to present a model of cognition quite different from the classical understanding, so it seems worthwhile to investigate its epistemological consequences If there is reason to think that connectionist models are good psychological models, then there is reason to think as well that connectionism can tell us something about the epistemology employed by humans. These two roles are of course related, the first concerning psychology and the second epistemology.

None of the epistemological morals outlined below are new The contribution made by connectionism here is not to present particularly new ideas about epistemology but rather to provide support for a particular epistemology by presenting a model of mind that employs it In that way, recalling the lessons of the previous chapter, evidence that connectionism presents a sound model of cognition is evidence for the epistemology it employs.

As well, to the extent that connectionism presents a means for investigating neural level representation and processing, it can further contribute to a proper understanding of the sort of epistemology employed by humans An adequate epistemology will no doubt prefer a level of description somewhat above that of the neural level, just as an adequate psychology will But a proper understanding of the goings on at the neural level will inform our understanding of epistemology, and serve as well to restrict the field of candidate theories to those epistemologies that can be implemented in extant nervous systems.

§

Since we have found that psychology and epistemology are intertwined to some extent, many of the epistemological features of connectionism have been thoroughly discussed in the first chapter. The centerpiece of that chapter was a discussion of what is perhaps the most significant epistemological aspect of connectionism, the nature of connectionist representations. A basic demand of epistemology is to account for how a being represents the world to itself; connectionist models are but the latest in a long line of answers that begins with wax tablets and aviaries.

Without revisiting the debate of chapter one, consider briefly the epistemological significance of connectionist representations. Of immediate interest is their non-propositional nature. This makes connectionism a likely ally for eliminative materialism. Whether or not one thinks folk psychology is committed to an account of the operation of the nervous system, eliminative materialism will certainly be in trouble if the structures and processes of the nervous system do in fact respect the generalizations of folk psychology, and having propositional representations is the easiest way to do just that.

A clarification regarding the relation between eliminative materialism and the nature of connectionist representations is in order. Recall that the aim of investigating connectionism is to find an epistemology consistent with eliminative materialism. The argument here is not that connectionism is non-propositional and therefore supports eliminative materialism. What matters is the extent to which connectionism and eliminative materialism are compatible; insofar as connectionism can tell us something about psychology, then it can tell us also about the fate of folk psychology. If that means that the sort of epistemology one gets after rejecting folk psychology involves non-propositional representations, then so be it. The issues here are more complex than simply the structure of representations, even if the mode of representation in connectionist networks is at the heart of many of the differences between the classical and connectionist approaches.

Chapter one dealt largely with the low level details of network operation, the nature of representations, the details of information storage, and so on. It will be important, for an understanding of connectionist epistemology, to have a grasp of the global activity and dynamics of connectionist networks. Two different mathematical analyses are useful for this purpose. Both represent states of a network in a multidimensional space. The first involves representing the all of the various connection weights in a network by a point in "weight space", while the second involves representing the activity of a particular layer as a point in "activation space".

Weight space is an abstract multidimensional space, it has one axis for each connection in a network, with an additional axis for a global error measure. The global configuration of connection weights at a particular time is represented by a point in weight

space the weight measure for each connection serves as a coordinate All of the connection weights, together with a measure of the error, determine an individual point in weight space. Weight space presents an interesting basis for understanding the global change a network undergoes as a result of learning As we learned in chapter one, "learning", in connectionism, is the modification of connection weights over time as a function of experience. The expected result of learning is a reduction in error In weight space, we can understand learning as a change in position in weight space: since each possible global configuration of connection weights determines, error aside, a unique point in weight space, then any change in the weight of the connections will appear as a change of position in weight space Since, over time, the result of learning is a reduction of error, we can understand learning as a descent in weight space, where the position of the network in weight space descends relative to the error axis.

The second analysis of interest here represents not connection weights but the activity of a group of units in a network, usually a hidden layer. Activation space has as many dimensions as there are units to be analyzed. If we are analyzing the hidden layer of a network, the hidden unit vector will determine a particular point in activation space. Since different input will often produce different hidden layer vectors, the hidden units will often occupy different positions in activation space. So, on the activation space analysis, there can be a change in position without the global change of learning.

The interest in activation space lies not so much in the location of the individual points determined by the hidden layer, but in the overall structure of the space. The different hidden unit vectors lie in various regions of activation space. During training, what the network is searching for is a way to partition its activation space in order to make the discriminations demanded of it. A network that makes a binary discrimination (discriminates two types of input) needs to partition its activation space into subvolumes in such a way that one type of input will fall on one side of the partition, while a second type falls on the other side

Connectionist networks produce graded responses to discrimination tasks The activation space of a successfully trained network will be organized in such a way that input that is unambiguous or prototypical will produce in hidden layer activation space a point in the central region of a subvolume, while atypical or problematic input will be found on or near the partitions between the subvolumes. The discrimination tasks demanded of connectionist networks need not be so simple as a binary discrimination. Activation space is quite large; only the complexity and size of the network serve to restrict the number of partitionings of activation space that a network can produce. The partitioning of activation space is determined by the global configuration of weights together with the structure of the

network. So for a particular network, its point in weight space determines the structure of its activation space.

We should note the intriguing nature of the structure of activation space before we enter the debate concerning the epistemological significance of these two analyses. The hidden units in a connectionist network allow it to make discriminations based on higher order statistical features of the input set, while a simple two layer network is sensitive only to its first order statistics.[1] The hidden layer, in a very real sense, is responsible for categorizing the input. As a result, the partitioning of hidden unit activation space often reflects substantive differences in the world that are only partially or implicitly present in the input.

A dramatic example of this feature of activation space involves the network NETtalk (Rosenberg and Sejnowski 1987)[2]. NETtalk outputs a string of phonemes given seven letter word segments as input, with the appropriate vectorial codings for each. The network is trained up to output the appropriate string of phonemes for a given word or word segment, it does not parse sentences or "understand" words. When properly trained, there is a hierarchy of partitions in the hidden unit activation space of NETtalk, with two major regions themselves divided into smaller regions, with subdivisions of subdivisions and so forth. In all, there are 79 subdivisions of the NETtalk's activation space. It is no coincidence that one must master 79 different letter-to-phoneme associations in order to properly pronounce English spelling; when the network is properly trained, it produces a distinct hidden unit activation pattern when making each of the 79 possible associations.

If NETtalk's activation space partitions seem of only minor significance, consider another feature of the partitions. If, in the course of experimenting with the network, one takes the time to determine which hidden unit vector is involved in each letter phoneme association, it is possible to map all the 79 subdivisions of activation space onto distinct letter-phoneme associations. Recall that there is a hierarchy of divisions of NETtalk's activation space; it turns out that, at the top of the hierarchy, the broadest division of activation space represents the division between vowels and consonants. As well, looking into the consonant region, there are subdivisions of the principal consonant types. The activation space is structured so that similar letter-to-phoneme associations are proximal in space, while dissimilar ones are more distant.

Surely this feature of NETtalk's activation space is of significance. Trained to produce the proper sequence of phonemes from the input word, the network has learned not only the intricacies of the phonological significance of English spelling, but as well the

---

1 Sejnowski et al. (1986).

2 My discussion of NETtalk follows in part that of Churchland (1989a).

complex organization of the phonetic structure of English. It has partitioned its activation space in a way that will allow it to make the different letter-to-phoneme associations, and it has organized its activation space in a way that reflects substantive real world differences When trained, NETtalk is well equipped to go beyond its training set and pronounce new words

NETtalk's activation space is a quality space or "similarity metric" of letter-to-phoneme associations, the network has developed categories that respect differences in the domain of its task Its task domain is far too small to say that the network has the same sorts of concepts that human speakers of English do. But, over the course of training, NETtalk develops a system of categories that allows it to deal with its input in a way that reduces the error substantially

If we can view the nervous system as an immense network of interconnected processing units, then we have before us the beginnings of a rich understanding of human conceptual frameworks. An individual's activation space for a particular layer will be partitioned into distinct categories in such a way that they can make sense of their sensory input - including noisy, incomplete or ambiguous input - and the partitioning is such that it keeps error to a minimum. Through an analysis of the similarity metric of activation space, the brain reveals its categories There is of course a substantial jump here from connectionism to neuroscience, but the jump is quite deliberate, we want connectionism to tell us about what humans are up to. But it is unclear how faithful the connection weights of connectionist models are to the synaptic weights of neuroscience The differences between synapses and the weighted connections between connectionist processing units are too numerous to mention.

Connectionist models are not usually intended as models of neurons, but that does not mean that the jump here should be disturbing. We have already discussed the merits of models that simplify neural function in the manner of connectionism; an understanding of the global storage of information in the brain will be overwhelmingly complex unless we are employing a model that simplifies neural activity to a great extent. It is just this sort of jump from connectionism to neuroscience that was recommended at the end of chapter one, where it was suggested that we set aside somewhat the aims of those working in connectionism and see what their work can tell us about neuroscience when we employ connectionism in constructing simplified neural models

Before dealing further with the above grand claims about activation space, we should consider first the views expressed in the only other openly epistemological discussion of connectionism, due to Paul Churchland Churchland wants to employ connectionism in an account of the nature of theories Since connectionism gives us a non-propositional account

of knowledge representation, then we may have the basis for a non-propositional account of theories The analysis of activation space reveals that, in a properly trained network, connectionist activation patterns (representations) are such that they respect real distinctions and structures, and they allow the network to make sense of their input in a way that keeps the error to a minimum. "These," notes Churchland, "are the functions typically ascribed to theories."[3] Churchland recognizes the potential for an account of human categorization in activation space analysis, but he is caught between the two different analyses Variously using the terms "conceptual framework" and "global theory", Churchland wonders out loud how which of the two abstract analyses we have been considering should be used to identify an individual's global "theory of the world"

The similarity metric of neuron activation space and the manner in which it is responsible for categorization makes it the more obvious choice for global theory But the partitioning of activation space is determined by the brain's location in synaptic weight space, so should we not ultimately identify global theory with the brain's position is synaptic weight space? So begins this rather odd debate about the possible implications of connectionism.

The case for activation space lies largely in the transparent manner in which it does its categorization. In his (1989c), Churchland makes this defense of activation space:

> People react to the world in similar ways not because their underlying weight configurations are closely similar on a synapse-by-synapse comparison, but because their activation spaces are similarly partitioned.[4]

If we use the partitioning of activation space as our understanding of an individual's global theory, then we retain a similarity measure that would be lost if we identified global theory with a point in synaptic weight space. Certainly we should never expect two individuals to ever occupy identical points in weight space; indeed, there is no reason to think that an individual's point in weight space will be the same from one day to the next Further, the structure of weight space is such that proximity between two points is of little significance, one has to look to the activation spaces that they would respectively determine in order to understand in what ways they represent a similar understanding of the world

Churchland's case against activation space and for identifying global theory with a point in weight space goes like this: It is the point in weight space that determines the partitioning of activation space. The laws that govern cognitive evolution ("learning") do not recognize the partitioning of activation space, they mention only connection weights

---

3 Churchland (1989a 177).

4 Churchland (1989c: 234).

Learning will of course affect the partitioning of activation space, by virtue of changing the brain's position in synaptic weight space. But because different points in weight space can partition their respective activation spaces quite similarly, knowing where the brain is in synaptic weight space presents a fuller understanding than a knowledge of activation space partitions. To return for a moment to explicitly discussing connectionist networks, consider two networks that have similar activation space partitions but occupy different points in weight space. These two networks will behave in a similar fashion given similar input. But, so long as the networks are undergoing learning, given a large enough set of input with a sufficient amount of problematic or atypical input, these two networks may come to behave differently. Knowing the original points in weight space for these two networks, we are better equipped to understand and predict the dynamics of their behaviour. To this end Churchland has this to say:

> Accordingly, if we want our "unit of cognition" to figure in the laws of cognitive development, the point in weight space seems the wiser choice... We need only concede that different global theories can occasionally produce identical short-term behaviour.[5]

Is it really that important that accounts of cognition and of cognitive development be defined over the same entities? It's not at all clear that this is crucial. Cognition and cognitive development certainly involve some of the same sorts of things, but no one need deny that, and it does not count against understanding global theory as the partitioning of activation space. The best case for weight space points has already been made: the point in weight space both determines the activation space partitions and gives a better understanding of network dynamics and so of long-term behaviour.

The complexity and sheer size of brains may bring into question the relevance of this debate. But let us first try to resolve the dispute that Churchland seems to be having with himself. In a sense, he is quite right to flip from one view to the other, since the point in weight space determines the partitioning of activation space. The problem is that the relation does not hold in the opposite direction: a particular partitioning of activation space does not determine a unique point in activation space. Churchland thinks that we can overcome this by acknowledging that different global theories can produce the same short-term behaviour. While the synaptic weight space analysis does indeed present a richer understanding of cognitive dynamics, identifying global theory with a point in weight space is of doubtful utility. If we are to identify an individual's global theory or conceptual framework in either of these analogues of connectionist analyses, it should be in the partitioning of activation space.

---

[5] Churchland (1989a: 177-8).

The best case for points in weight space was that they both determined the partitioning of activation space and offered a better understanding of cognitive dynamics and so of long term behaviour. But the point about long term behaviour has no relevance for the decision between the two candidates for global theory Reverting again to talk of connectionist networks, consider again two networks that have similar activation space partitions but which occupy different points in weight space. The reason that these two networks will come to diverge in their behaviours, when given enough time and problematic enough input, is because both of their global theories will come to change They will come to partition their respective activation spaces differently, simply because their respective points in weight space change The point of the original example was that by viewing global theory as activation space partitions one will miss subtleties to which the "point in weight space" view is sensitive. But on either analysis of global theory, this supposed divergence of short term behaviour is a result of a change in global theory. It is still correct to say that the weight space analysis presents a better understanding of cognitive dynamics, change in theory results in a shift in weight space but a global restructuring of activation space But in terms of a given global theory, there are no subtleties recognized by the weight space analysis that the activation space analysis misses

Viewing global theory as a point in weight space also leads to a few problems To judge any similarity between global theories (either the theories of different individuals or of one individual at different times) one will have to appeal to the activation space partition analysis. If we employ only the weight space view, no two individuals will ever share the same global theory. Some may not find this point disturbing, so I will add two rather more ominous ones. Viewing global theory as a point in weight space means that one particular individual can never have the same global theory at two different times, not even in two consecutive seconds. Synaptic weight change occurs all of the time in the brain, and a change in the weight of one synapse is enough for a change in global theory, on this analysis As well, the structure of weight space, unlike that of activation space, is such that, except along the error axis, the distance between points (theories) is of little relevance The distance is a measure of how much global weight change must occur to get from one theory to the other Of what use is an analysis of global theory in which everyone, including one's one past and future time stages, has a different theory and there is no basis for recognizing similarities between theories?

The activation space analysis gives us some hope for an ability to recognize individuals who have similar global theories The demand here for a recognition of similarity is quite weak; if you and I recognize all of the same categories and have all of the same beliefs and so on, but yet one of us thinks that tomatoes are fruit while the other takes them to

be a vegetable, should that mean that our global theories are forever incomparable? One hopes not, although it depends on how one understands "global theory"; perhaps the fairest understanding of global theory committs one to saying that you and I (and all of my previous time stages) employ incomparable global theories I am not sure, however, that that would be a particularly useful concept To be fair, one can understand global theory to be a point in weight space and employ activation space in an analysis of similarity But it seems there is little reason left to prefer the view of global theory as a point in weight space over its competitor

How much substance is there to this debate? The jump here, from talk of connectionist networks to talk of brains, is no small one. Connectionist networks are typically simulated on ordinary serial computers, so knowing the weight of all of the various connections is a simple matter of having the proper programming. Determining the point in weight space for a particular network is a fairly simple matter. Ordinary brains, however, are not so forthcoming with the details of their synapses On a quite conservative estimate, the human brain has on the order of 10^{14} synapses. A corresponding synaptic weight space would be immense: it would have 10^{14} axes (plus an error axis) and on each axis would be as many possible positions as there are functionally distinct weightings of synapses. We will never be able to determine the point in synaptic weight space for a human brain. Nor, at least with present neuroscientific knowledge, will we be able to narrow down to any significant degree the region of weight space that the brain may occupy Synaptic weight space is altogether too large a space to deal with when one cannot even narrow down the possibilities

Similar but not as dire problems hold for the view of global theory as the partitioning of neural activation space. There simply are too many neurons, even if we acknowledge that activation space is concerned only with a particular layer of a network. But therein lies another problem; although the brain, in many areas, is structured in layers, the divisions between layers are not as distinct as the divisions usually found in connectionist networks And it is important, at least in connectionist analyses, that there be agreement on which units are in which layer As well, to the extent that the brain is layered, it has many layers So if we are to identify an individual's conceptual framework or global theory with the partitioning of their activation space, there are several, rather than just one, activation spaces to be considered This is of great help to the brain, but complicates somewhat the view of global theory as the partitioning of activation space.

These problems are real but not insurmountable The differences between connectionist networks and biological brains are many and well worth remembering, but they should not deter us from finding some neurological significance in these two analyses of connectionist processing The importance of these two analyses is the manner in which they

aid the understanding of the global behaviour of a network, and the global change that results from learning. It is this sort of understanding that one hopes to have of brains.

The "too many neurons" problem may be less of a problem for activation space than for weight space. In the activation space analysis, the concern is less with particular points in the space than with the overall structure of the space, although the way to determine the overall structure is to see where various individual vectors are located in the space. On the other side of the debate, while it is true that we will never determine the point in synaptic weight space for a particular brain, that does not mean that there is no utility in talk of such space. An investigation of the learning rules employed by the brain may tell us something about the mathematics of learning driven changes in weight space. So there are things to be learned about synaptic weight space, simply because there are things to be learned about synaptic weights.

To the extent that these connectionist analyses are applicable to brains, it would seem that the activation space analysis presents the most profitable conception of an individual's global theory. But if our interest is in learning and the global change involved in it, then we may find it more illuminating to think of global theory as a point in synaptic weight space.

§

My own concern with connectionism is for an investigation of quite basic features of epistemology: the nature of representation, the relation between the world and representations, and so forth. Before exploring further these aspects of connectionism, we should consider some conclusions, drawn by Paul Churchland, that are more in the realm of the philosophy of science.

We have already seen that Churchland thinks that connectionism presents the basis for a new, non sentential understanding of theories. However, if connectionism has anything to reveal about the nature of theories, we have so far only learned about the nature of global theory. At any rate, Churchland thinks that an understanding of connectionism can tell us something about a number of issues in the philosophy of science. I consider three of Churchland's conclusions below. They concern the nature of simplicity, of conceptual unification, and the potential for a vindication of Thomas Kuhn's views on the philosophy of science.

Churchland thinks that a connectionist understanding of the nature of theories allows us to have a better understanding of simplicity, and how it might count as a genuinely

epistemic virtue of theories, rather than a merely pragmatic or aesthetic virtue The explanation involves the partitioning of activation space and the role of hidden units in processing How well a network generalizes depends partly on how many hidden units it uses to solve the problems that it does. For any given task, there is an optimal number of units for the processing required. Below the optimum, a network never learns to respond properly to the training input. Above the optimum, a network will respond well to training, but will perform poorly with new input.

It is during training that a network comes to organize its hidden unit activation space in a way that, when successful, allows it to recognize relevant features of the input. A network with too many hidden units develops too many partitions in its activation space Instead of recognizing the categories required for the task at hand, this sort of network will recognize too many categories, it will develop a distinct activation space subvolume for each input vector (or a small group of such vectors) from the training set At the end of training, a network of this sort has learned to associate each input vector with the appropriate output, but without developing the activation space partitions that allow it to deal with input outside of the training set The network has organized its activation space as a result of learning, but in an "ad hoc, unprojectable" way.[6] It has too many partitions for the task at hand, and as a result fails to recognize features of the input suitable for generalization.

So as long as a network develops the minimum partitions suitable for its task, then the fewer partitions (and the fewer hidden units) the better "Ceteris paribus," interjects Churchland at this point, "the simpler hypotheses generalize better." Simplicity is a genuinely epistemic virtue because it facilitates superior generalization.[7] Lurking not too far in the background of this argument is an account of simplicity. It seems fairly obvious that Churchland wants us to think that simplicity is a matter of the number of activation space partitions. Or rather, it is the number of subvolumes produced by the partitions that matter, because that determines the number of different features of the input set (or kinds) recognized by the network Churchland makes no claims to have provided an account of simplicity, but clearly his conclusion that simplicity counts as an epistemic virtue depends on an account of simplicity, specifically the one alluded to in the previous paragraph But by using activation space partitions as a measure of simplicity, I'm not sure that we have an account of simplicity rather than just a new name for it. I'm not sure, in other words, that we have learned anything

The conclusions that Churchland draws about the virtue of simplicity may not apply in the places where we most need them To this end, I'm not sure of the relevance of the

---

6 Churchland (1989a: 180)

7 Churchland (1989a: 181).

story, recounted above, about learning with too many hidden units. The reason that the network with the optimum number of units is better suited to generalize over input outside of the training set is because it recognizes projectible features of the input. The network with too many units never latched on to the relevant features, it just memorized the right output for each input in the training set. So the optimal network generalizes better to new input than the non-optimal one. But why should this tell us anything about simplicity? If there is any sense to talk of these networks having theories of their task domain, then what we have here is just a bad theory versus a good one.[8]

The real curiosity with simplicity arises when we have the opportunity to give up a reasonably successful theory for a simpler one that describes the same phenomena. Should we prefer the more complex theory to the simpler one? If so, why? Churchland's story about learning and the optimum number of hidden units sheds little light on these questions. If he is correct in saying that, ceteris paribus, simple theories generalize better, then, ceteris paribus, we should never get into a situation where we even have to consider the simplicity of a theory when choosing between theories, because the simpler theories will generalize better. My remarks here may be somewhat unfair. But the point is that Churchland's story about learning is itself unfair. In a contest between a good, simple theory and a complex, ad hoc one, the former is bound to win. We will need more of a fair contest if we are going to learn anything about simplicity.

Churchland's second, related finding from connectionism concerns conceptual unification. He has us suppose a lowly creature that must employ whatever understanding of the world will allow it to survive to the next meal. What matters to it is getting its nervous system output at least roughly right, so the creature may have to be satisfied with employing a distinct similarity space for each sort of situation it meets. Where it is possible, we are to suppose, it is advantageous to generate a single similarity space that unifies the two distinct spaces. In this way, the creature has conserved its conceptual resources and obtains the means for dealing with phenomena that fell in between the two previous spaces and was dealt with poorly by both of them.

Conceptual unification and simplicity are related, and Churchland thinks that they share the same virtue of superior generalization. But his story here is rigged in the same way as the discussion of simplicity. Apart from the pragmatic point about doing more with the same resources, we are presented with a situation where we have no choice but to prefer the unified theory over the two distinct theories, because ex hypothesis the unified theory is more

---

[8] I should emphasize here that nowhere in this chapter is it claimed that connectionist networks have or use theories. One should not conclude from the connectionist flavour of the account of theories discussed in this chapter that there exists a commitment to such a claim.

successful. Again I am being somewhat unfair, because Churchland means to show that conceptual unification results in superior generalization, so it is in a way inevitable that the story will be rigged in this way But there is little in the story to demonstrate that conceptual unification will result in superior generalization in other cases.

Theories that are simpler than their competitors, or which unify their predecessors, may certainly be more successful. And the relationship between the simplicity or unification on the one hand and superior generalization of the other need not be accidental But we should not be convinced that simplicity and conceptual unification need always be empirical virtues After all, it is fairly easy to have a theory that is too simple and fails to deal adequately with the phenomena with which it is concerned. No one has denied that, but it serves to demonstrate that we come out of this debate knowing as little about simplicity as when we came in If simplicity and conceptual unification always produced superior generalization, then they need never have been the object of debate in the history of science (both the discipline and the actual past); theories could have been judged on predictive success alone

The third of Churchland's findings regarding connectionism and the philosophy of science involves a defense of Thomas Kuhn. In a brief passage[9], Churchland proposes that connectionism gives us the means for explicating Kuhn's notion of paradigms.[10] Kuhn is notoriously vague about what he means by "paradigm". Masterman (1970) finds no less than twenty-one different senses of the word in Kuhn's book, from "scientific achievement" and "analogy" to "a successful metaphysical speculation" and "a set of political institutions". I will not attempt to interpret Kuhn here; for his purposes Churchland understands a paradigm as a prototypical application of some set of resources, be they mathematical, conceptual, or instrumental.

Churchland finds a very strong connection between Kuhn's prototypical applications and the prototypes that are found in the centre of activation space subvolumes:

> For a brain to command a paradigm is for it to have settled into a weight configuration that produces some well-structured similarity space whose central hypervolume locates the prototypical application(s).[11]

This understanding of paradigms explains, for example, why even the most reflective and self aware persons are unable to fully articulate the relevant factors for applying a paradigm to a particular case.

---

9 Churchland (1989a. 191-2)

10 Kuhn (1962).

11 Churchland (1989a: 191).

This is very queer stuff indeed, but there are perhaps reasons better than queerness to reject Churchland's understanding of Kuhn Previously, I noted the many parallels between activation space subvolumes and kinds. While Kuhn's paradigms do not seem to be kinds in any useful sense, that alone should not deter us from Churchland's reading. For humans have several vast neural activation spaces, which allow for activation space subvolumes for many fiendishly complex categories, as well as wholly unnatural kinds, such as speech, thermal equilibrium, glaciation, political corruption and space flight.

Churchland's view allows us to explain some features of paradigms; it captures some of the ways in which paradigms are supposed to guide or govern one's perception or understanding of the world But paradigms play many roles, and Churchland's proposal seems ill suited to many of them. To this end, consider another parallel that Churchland draws between paradigms and prototype applications. Much of Kuhn's discussion of paradigms concerns scientific revolution and the resistance to change or displacement of a paradigm. Churchland proposes that this resistance is the result of the way in which networks (brains) learn. In terms of weight space, the goal of learning is to reduce the error to a minimum At times, a network will get "stuck" in a local error minimum, where the error is high, or at least not as low as it could be, yet any small change in the network results in more error than remaining in the minimum The network gets stuck because it is, overall, to its advantage not to increase the error

From this story, Churchland concludes that the resistance to an increase in error is at the heart of the resistance to paradigm shifts. I think Churchland is way off the mark here I just don't see that all of the socio-epistemological stories that Kuhn tells about the response to crises can all be explained as the result of this feature of learning Were those who advocated the shift from Ptolemaic to Copernican astronomy simply better learners than those who did not? A tendency to avoid error can, in a variety of situations, lead to conservatism of this sort. But there are a variety of reasons for the affiliations that scientists and ordinary folk draw during a crisis of Kuhn's sort. Doubtless there are many reasons that have little to do with error avoidance, and I do not see that Kuhn recognizes anything special about those that do.

§

In the remainder of this chapter, I want to explore further the implications of the connectionist inspired account of Weltanschauung developed earlier

Recall that the activation space analysis proved to be a more profitable analysis of global theory than that provided by an analysis of weight space. Essentially the same partitionings of activation space can be achieved by quite different global configurations of connection weights. So note, as mentioned previously, that this implies that the same global theory or conceptual framework can be shared by individuals with different global configurations of synaptic weights. Also, to make basically the same point, it means that people will not suffer from global conceptual shifts at every negligible synaptic weight change.

This is basically a point about implementing an epistemology, rather than a claim about some feature of a particular epistemology. There is something of the functional/implementational distinction in distinguishing activation space from weight space. The difference between the two analyses is not that one is a functional analysis while the other physical; the difference lies in the fact that for many of the possible organizations of activation space, there are a number of points in weight space that can achieve that organization. The global configuration of synaptic weights has important functional and developmental consequences but, at any particular time, the difference between two brains with the same activation space partitions but different global weight configurations is epistemologically uninteresting.

There is a feature of weight space, only hinted at previously, that is of enormous epistemological interest. I am thinking here of the error surface in weight space: this feature is of interest because it gives us some understanding of the relation between the organization of networks (brains) and the world. The error surface is in fact a multi-dimensional hypersurface in weight space. For each possible global configuration of connection weights for a particular network, there will be a particular global error value. Adding together all of the points representing global weight configurations and global error, the sheet one winds up with is the error surface. The error surface has one fewer dimensions than the weight space in which it resides.

With some altogether minor caveats, we can view the error surface as a powerful and intriguing analysis of the relation between different global theories and the world. The first caveat involves the conception of global theory considered earlier. In order to view the error surface in that manner proposed, we will have to understand an individual's global theory as a point in their synaptic weight space. Previously it was decided that the organization of activation space analysis was, for many reasons, the preferable analysis of global theory. However, at the end of that discussion I did leave room to revert to an understanding of global theory as a point in weight space, should such a change be appropriate given one's interests. A discussion of the error surface will lead us to considerations of learning and

weight change, just the sorts of considerations that make the analysis of global theory as a point in weight space a useful one

The second caveat here involves the role of the world I am assuming that, for humans (but perhaps also for networks), the world often plays an important role in error I erred in reaching out to my cup because my actual hand fell short of the actual cup When I say "Windsor is north of Detroit", I err because the world is in fact not as I describe it I am promoting here a realist flavoured conception of error, but that is what we are lead to if we want to discuss the relation between global theory and world external to our minds Those queasy with this view of error might otherwise think of error as simply empirical adequacy inverted. For the moment, I leave this (the first) view of error as an undefended caveat, I will discuss later the appropriateness of this and the related connectionist conception of error

Caveats in place, we can now consider the error surface as a representation of the varying adequacy of the universe of global theories that humans can possibly implement This allows us to ask a thoroughly epistemological question: what is the shape of our error surface? Some global theories are better than others, so there will be variations in the height of the surface (relative to the error axis): it will not be flat, in other words

So the error surface is bound to be bumpy. A simple thought experiment should prove that a large part of the error surface should have a high altitude. Recall that there are a vast number of synapses in the human nervous system, at least 10^{14}, but perhaps even greater than that in magnitude. Now consider how well an ordinary brain will perform with its synapses all randomly weighted. There is an extraordinarily good chance that a brain frazzled in this way will perform dismally, producing only noise as output To say that there is a very bad chance of getting a working brain by a random configuration of weights is just to say that most of the points on our error surface are quite high in altitude Choosing a random configuration of weights is, after all, to choose a particular point in weight space So if we are correct in thinking that a random configuration of weights is likely to produce a great deal of error, then a randomly chosen point in weight space is likely to be high in error; that is, it will be high up on the error surface.

We do know from our own experience that the global theory we use is successful, at least to the extent that it allows us to survive. For the moment, I want only to suggest that the sort of survival that humans enjoy implies that our nervous system output is low in error, at least when compared to the alternatives on the error surface I do not intent here to make a connection between success and truth. We know as well that for survival value (error avoidance), a number of different global weight configurations will be similarly successful So not all of the error surface will have a high altitude

Error surfaces for connectionist networks (or rather their three dimensional analouges) often look something like valleys or caverns. They have a high region, a low region (the global error minimum) with sloping sides between Often, regions of the error surface will have local minima regions that are less errorful than the surrounding alternatives, but yet are higher than the global energy minimum Local minima are of interest because, to escape them, a network's weight space point must first climb the grade of the error surface. So a network must temporarily increase its global error in order to eventually reduce its global error to an effective minimum. An adequate learning rule does much more than simply force a weight space point downwards. It must in a sense "shake" a network out of local minima in order for it to function as best it can.

As it turns out, the more dimensions there are to a particular weight space, the smaller the probability of there being troublesome local minima of this sort. But apart from probabilities of this sort, there is no guarantee that our brains are not presently stuck in a local minimum that is both low enough in altitude to permit the survival that we enjoy, and deep enough that no amount of "shaking" would allow us to escape. Further, our error surfaces may fail to have a unique global error minimum. There may be several equally or at least similarly errorful minima in our error spaces. There may be, in other words, ways of seeing the world that are different from our own, yet are equally successful

At the outset of this chapter, I warned that I would endeavor to put new wine into old bottles. There is a longstanding debate concerning the possibility, and indeed the very coherence, of there being alternative conceptual frameworks or "conceptual schemes". I do not intend to enter this debate, except to point to all of the previous argument of the thesis My aim in the thesis is, from the perspective of naturalized epistemology, to see what connectionism has to say about epistemology. If it says that alternative conceptual schemes are possible, then so be it. My epistemological claims are largely the consequent of a conditional, with an assertion of the adequacy of connectionism as the antecedent. To the extent that I do in fact make claims about epistemology, I can and do point to the evidence for the adequacy of connectionism elsewhere in the thesis. Insofar as I am approaching epistemology from the naturalistic camp, I have already made all of my arguments for the coherence and possibility of alternative conceptual schemes, in demonstrating that connectionism finds nothing incoherent or (logically) impossible in the possiblility of alternatative conceptual schemes

Churchland notes briefly the two possible features of error mentioned above, and tosses the wand to Stich (1990) in the hope for a pluralistic form of pragmatism While I share his hopes, I do not think, with regard to pluralism, that our present understanding of neuroscience has quite got us to this point The connectionist inspired understanding of

global theory and error discussed does give us some way of understanding how it might be possible for there to exist conceptual schemes different from our own But at the moment, possibility is all we have. We got into this discussion by wondering out loud about the shape of our error surface. Throughout this thesis I have posed empirical questions in the hopes of finding epistemological answers, and this is just another such question The shape of our error space is a matter for empirical investigation, rather than conceptual speculation It may turn out that there is in fact a unique global error minimum in our error space, and so we are left free to judge different conceptual schemes as sub-optimal, depending of course on our confidence that ours is the optimal viewpoint. Whether or not we should think that there is a uniquely good way of seeing the world - whether or not we should be epistemic pluralists - is an empirical question. Note further, however, that different individuals may in fact have different error surfaces; so whatever danger there is of pluralism in the views being discussed, it appears twice over

The potential for pragmatism here is more clear. We are lead to something like pragmatism, not because of inevitabilities, but because it allows us to make the most sense of what connectionism seems to have told us. Epistemic pragmatism is somewhat nebulous a view, my immediate interests are in tying pragmatism to the claims already made rather than specifying exactly what pragmatism should look like Pragmatism is first of all a consequentialist view. Epistemological (ie. cognitive) processes are judged as we might judge other processes by their ability to help us achieve valued ends Thus, pragmatism is relativistic, in that we judge systems of reasoning (conceptual frameworks, cognitive processes) relative to the ends desired: should different people or different cultures desire different ends, then they may not find the same system to be equally valuable.[12]

This last point leads us to the connection between pragmatism and epistemic pluralism Should different individuals value different things, then it is possible that they might employ, with equal success, two wildly different conceptual frameworks It is this potential for pluralism that produces much of the distaste some have for pragmatism, because it seems to lead to skepticism If there are equally good ways of reasoning that would lead to different beliefs given the same evidence, then there seems to be little connection between good reasoning and reasoning that leads to the truth Or so goes the argument

With regard to truth, Stich thinks that the conclusion that we should draw from pragmatism is that truth is not something that should really concern us This conflicts

---

12 Stich (1990) argues that any consequentialist accounts of reasoning, even ones appealing to truth, lead to relativism in a second way The output of a cognitive process he argues, depends on the social environment in which it functions. So evaluations of the output of such a process are bound to be environment relative (pp 136-8)

somewhat with the views of other pragmatists, for example Rescher (1977), who hope to give a pragmatist account of truth. Stich's argument - the argument of his whole book (1990) - has two parts. The first part is a defense of consequentialism in epistemology, while the second involves a series of stories meant to demonstrate that truth or the generation of true beliefs is in many cases not a valued end. Stich's argument is too lengthy to evaluate here. Suffice it to say that he thinks that we should prefer cognitive processes that best allow us to achieve ends that we value, and on a close enough examination, truth is actually not something that we value. In essence, therefore, Sitch's response to the skeptical challenge above is not that it is wrong, but that it should not worry us. Stich is quite happy to say that we should not worry about a connection between good reasoning and reasoning that leads to the truth.[13]

Returning to the connectionist style account of epistemology developed above, the best sense we can make of connectionist learning and connectionist processes is judge them by their consequences. So a connectionist epistemology is at least a consequentialist one. The avoidance of error in output has throughout been understood to be the principal virtue of connectionist processes. What then might this account tell us about the desired relation between head and world, between theory and data? Again, this is an empirical matter, and an unresolved one at that. It may be that there are equally successful global theories dotting our error surface, but it is also possible that there is only one global error minimum. It should be clear that the existence of alternative conceptual schemes is inconsistent with the "One final true theory" conception of truth. But it is not new to suggest that this is an outmoded conception of truth.

What does connectionism tell us about our epistemically valued ends? Throughout the above discussion, and throughout actual work in connectionism, output has been judged by its distance from the ideal output, and this measure is called "error". I have already discussed the general nature of the sort of epistemology that connectionism gives us, but now there is sufficient background for a statement of it: The activation space partition (or weight space point) to be preferred is the one that has the lowest point on the error surface.

This proposal has been discussed at length, but the pragmatist contribution regarding the relevance of epistemically valued ends is recent to the discussion. There are two foibles in

---

13 There is an illuminating parallel between Stich's pragmatism and the views of the ancient sceptic Pyrrho. Pyrrho, finding that certain, genuine knowledge could not be attained, reccomended that one act on what seems most plausible or probable. For Pyrrho, this is a matter of working with what we have: a second best version of knowledge. Recognizing much the same distinction and a similar epistemic situation, Stich embraces the "second best" and finds that it is the only conception of knowledge that we need ask for. The "proto-pragmatist" reading of Pyrrho is due to Rescher (1977).

the above statement with regard to viewing error avoidance as an epistemically valued end First, there is a sense in which it borders on the tautological to say that error avoidance is an epistemically valued end. Or perhaps the problem is really that error avoidance isn't an epistemic end at all, error is the distance between actual output and output that best serves ones epistemic ends So error avoidance is everywhere desirable If it is an epistemic end at all, it is one that everyone values End or not, this means some trouble for the proposal to judge global theories by their location on an error surface The second foible with error involves a response to the first As noted earlier, the connectionist conception of error is a wholly realistic one: measuring error involves measuring the distance between the actual output of a network and the right answers to the problem the network is given The low level sorts of problems usually considered by connectionist networks are usually such that the right answers are not in doubt. If, because of its realist character, there are situations in which we cannot measure our error, then an epistemology that depends on such a measure will be of no use to us

There is a sense, alluded to before, in which this realist understanding of error is hard to avoid, if we accept that there is a world external to our minds If you really are standing on a precipice, the actual error in your nervous system output will be crucially important The importance of the first foible of error should be noted It may be that even the measurement of error requires one to have some (other) epistemically valued end But this is not a real difficulty for connectionism. It is only a sign that there is more work to be done; we need to investigate what sorts of epistemic ends people value.

With regard to what a connectionist epistemology can tell us about truth, I have suggested that this is a matter for further investigation. It may be, for whatever reason, that human error surfaces inevitably have a unique global error minimum If that were so, then there would be no need to worry about a conflict between connectionism and the naive realist conception of truth. There is however the possibility for a such a conflict, because there is nothing in connectionism that dictates that error spaces must have a unique minimum Churchland and Stich, both being prone to the elimination of troublesome concepts, seem to think that the solution to a possible conflict is to remove one of the aggressors, namely truth They seem to ignore the possibility of finding a middle ground. Stich, for his part, has gone to lengths to try to demonstrate the advantages of his pragmatism, but sees, unlike many others, little room for truth in pragmatism.

While Stich sees truth as unimportant, Churchland, even the early preconnectionist Churchland, seems to think that truth is just a bad concept Truth, he tells us, is bound up in folk psychology, folk epistemology, folk semantics, and so on, just as there has been progress in the domain of these theories, so to can we come to better understand what truth is

invoked to explain These are not new proposals, but Churchland thinks that we can still keep a brand of scientific realism without truth, because we can still assert the existence of a mind independent world, and we can still assert the rationality of commitment to the ontology of the best available theory [14]

I mention Stich and Churchland's views of truth in part to take some parting shots at their haste, but mostly to note the feature of connectionism most significant to for the fate of truth That there might be many equally good ways of viewing the world is not the worst obstacle to finding a place for truth in a connectionist epistemology, if indeed that is an obstacle The real twist to connectionism in this respect is that truth and falsity are features of propositions If states of a connectionist device cannot be mapped onto propositions in any useful way then it will be a mystery how brain states might come to have truth conditions. If connectionism leads us to non-propositional knowledge and non-propositional theory, then we will have to have a non-propositional understanding of truth Whether such an account of the relation between the world and our understanding of it would deserve the name "truth" remains to be seen Unlike Stich and Churchland, I am somewhat willing to wait

§

In closing, I want to note again the use of connectionism that I promoted at the end of the first chapter I recommended there, for a number of reasons, that the most fruitful application of connectionist models was to limit them to modeling phenomena at a level such that findings in connectionism were applicable to neuroscience This is not the only good use of connectionist models, but it produces a useful tool for investigation, and avoids many of the overblown claims of central members of the connectionist camp

Throughout this chapter, I have employed this view of connectionism. But I need not have There would still be epistemological lessons to be learned from connectionism on its standard interpretation But, for reasons discussed in the first chapter, the lessons are far less mysterious on the recommended view If we keep in mind the simplifying nature of connectionism with regard to neuroscience, then it is far easier to transplant findings of connectionism to our understanding of biological brains. We can, for example, hope to draw conclusions about synaptic weight space from the findings concerning network connection weight space Without such a view of connectionism, it is far less obvious what findings

---

14 Churchland (1985 151)

about connection weight space tell us about actual brains, other features of connectionism will prove equally troubling.

In establishing, with the help of Ramsey, Stich and Garon, a connection between eliminative materialism and the adequacy of connectionist models, the recommendations of the first chapter were ignored But the matter is not at all crucial, my recommendation concerns only the level of organization (or complexity) at which connectionism aims to model the functional activity of the nervous system As for those who find the difference between the connectionism of this chapter and that of chapter two disturbing, I can only direct them back to the discussion of the possible misunderstandings of my recommendation at the end of the first chapter.

As noted previously, my epistemological claims are for the most part conditional, based on the adequacy of connectionism. In the first chapter I expressed some misgivings about the present use of connectionist models, and of their alleged advantages over classical models. But overall I hope to have given some reason for preferring a connectionist style understanding of cognition over the classical symbolic view Thus, wherever there are reasons for preferring connectionism, there are also, via modus ponens, reasons for thinking that epistemology is the way that connectionism portrays it.

References

Bechtel, W , and A Abrahamsen (1991) Connectionism and the Mind· An introduction to Parallel Processing in Networks. Oxford. Basil Blackwell.

Churchland, P. M (1981) Eliminative Materialism and the Propositional Attitudes The Journal of Philosophy 78· 67-90 Reprinted in Churchland (1989), 1-22.

Churchland, P M. (1985) The Ontological Status of Observables: In Praise of Superempirical Virtues. Reprinted in Churchland (1989), 139-149.

Churchland, P. M (1989a) On the Nature of Theories Reprinted in Churchland (1989a) 153-196

Churchland, P M. (1989b) A Neurocomputational Perspective. Cambridge, MA· MIT Press

Churchland, P M (1989c) Learning and Conceptual Change. In Churchland (1989), 231-254

Churchland, P S (1986). Neurophilosophy Cambridge, MA· MIT Press

Churchland, P S , and T J. Sejnowski (1992). The Computational Brain Cambridge, MA MIT Press

Churchland, P S., C Koch, and T J Sejnowski (1990) What Is Computational Neuroscience? In Schwartz, ed. (1990), 46-55.

Feldman, J A., and D H Ballard (1982) Connectionist models and their properties Cognitive Science 6 205-254.

Fodor, J A (1975). The Language of Thought New York. Crowell

Fodor, J A and Z. Pylyshyn (1988) Connectionism and cognitive architecture A critical analysis Cognition 28· 3-71.

Goldman, A (1985) The relation between epistemology and psychology. Synthese 64 29-68

Goldman, A (1986) Epistemology and Cognition Cambridge, MA: Harvard University Press

Greenwood, J D , ed (1991) The future of folk psychology Cambridge· Cambridge University Press.

Hanson, S J , and C R Olson, eds (1990) Connectionist Modeling and Brain Function Cambridge, MA. MIT Press.

Hebb, D O (1949) The Organization of Behavior New York· Wiley and Sons

Heil, J (1991) Being Indiscrete In Greenwood, ed. (1991), 120-134.

Hinton, G E , J L McClelland, and D E. Rumelhart (1986). Distributed Representations In Rumelhart, D. E., J L. McClelland and the PDP Research Group (1986), 77-109

Kim, J (1984) Concepts of Supervenience Philosophy and Phenomenological Research 45. 153-176.

Kornblith, H , ed (1985a) Naturalizing Epistemology Cambridge, MA: MIT Press

Kornblith, H (1985b). Introduction: What is Naturalistic Epistemology? In Kornblith, ed (1985a), 1-14.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions Chicago University of Chicago Press (2nd edition 1970.)

Lakatos, I (1970). Falsification and the methodology of scientific research programmes In Lakatos and Musgrave (1970). 91-195

Lakatos, I. and A Musgrave, eds (1970) Criticism and the Growth of Knowledge Cambridge: Cambridge University Press

Masterman M. (1970). The Nature of a Paradigm. In Lakatos and Musgrave (1970)

McClelland, J. L. and D E Rumelhart (1986). A Distributed Model of Human Learning and Memory In McClelland, J. L., D. E. Rumelhart, and the PDP Research Group (1986b). 170-215.

McClelland, J L., D. E Rumelhart, and G E. Hinton (1986) The Appeal of Parallel Distributed Processing In Rumelhart, D. E., J. L. McClelland and the PDP Research Group (1986), 3-44

McClelland, J. L., D E. Rumelhart, and the PDP Research Group (1986b) Parallel Distributed Processing, Vol II: Psychological and Biological Models Cambridge, MA MIT Press.

Minsky, M , and S. Papert (1969) Perceptrons Cambridge, MA MIT Press.

Quine, W V O (1960) Word and Object Cambridge, MA: MIT Press

Quine, W. V O (1969). Epistemology Naturalized Reprinted in Kornblith ed (1985), 15 29.

Ramsey, W , S P. Stich, and J Garon (1991). Connectionism, eliminativism, and the future of folk psychology. In Greenwood, ed (1991), 93-119

Rescher, N (1977). Methodological Pragmatism Oxford: Basil Blackwell

Rosenberg, C. R., and T J. Sejnowski (1987) Parallel Networks that Learn to Pronounce English text Complex Systems 1 145-168.

Rosenblatt, F (1962) Principles of Neurodynamics New York Spartan Books

Rumelhart, D E and J L McClelland (1986a) On Learning the Past Tenses of English Verbs In McClelland, J L., D E. Rumelhart, and the PDP Research Group (1986), 216-271

Rumelhart, D E and J L McClelland (1986b) PDP Models and General Issues in Cognitive Science. In Rumelhart, D. E., J L. McClelland and the PDP Research Group (1986), 110-146

Rumelhart, D E , J L McClelland and the PDP Research Group (1986) Parallel Distributed Processing, Vol 1: Foundations. Cambridge, MA: MIT Press.

Schwartz, E L , ed. (1990) Computational Neuroscience. Cambridge, MA: MIT Press.

Sejnowski, T J , P K Kienker and G. E. Hinton (1986). Learning Symmetry Groups with Hidden Units Beyond the Perceptron. Physica D 22D: 260-275.

Sejnowski, T J , C Koch and P S. Churchland (1990). Computational Neuroscience In Hanson and Olson, eds. (1990), 5-35.

Smolensky, P (1987). The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn. Southern Journal of Philosophy 26 (supplement): 137-161

Smolensky, P. (1988) On the proper treatment of connectionism. Behavioral and Brain Sciences 11 1 23.

Stich, S P (1983) From Folk Psychology to Cognitive Science Cambridge, MA· MIT Press.

Stich, S P (1984). Could Man be an Irrational Animal? Some notes on the Epistemology of Rationality Reprinted in Kornbith, ed. (1985a): 249-267.

Stich, S P. (1990) The Fragmentation of Reason Cambridge, MA MIT Press.