Mind Design II

Philosophy
Psychology
Artificial Intelligence

Revised and enlarged edition

edited by
John Haugeland

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

To say this is not to disparage the brilliance of the connectionist achievement. For what it shows us is nothing less than how we might begin to put sentential epistemology together with the organic brain, and that is well worth celebrating. As is so often the case on such occasions, however, some of the revelers have a tendency to celebrate to excess. They claim that what connectionism shows us is an epistemology, that is not sentential, and that the organic brain might use instead. But that, I have argued, is just Enthusiasm—and in philosophy, too, Enthusiasm is still a sin.

Notes

1. See, in this connection, Sellars 1981, to which the present discussion is, and will continue to be, deeply indebted.
2. The priority of the notion of a judgment to that of a concept is also, of course, a central principle of Frege's philosophy. Frege's strategy, recall, is precisely to replace a "bottom up" account of judgments in terms of the composition of concepts by a "top down" analysis of the notion of "conceptual content" in terms of intersubstitutivity of and in judgments *salva* correct inferences.
3. This fact, in turn, is explained by the (temporal) *linearity* of speech (and the spatial linearity of script). As Wittgenstein was the first to see—in the *Tractatus* (1922/74)—the predicate expressions of such a linear representational system are, from the *functional* point of view, auxiliary signs, serving only to guarantee a stock of characteristics of and relations among referring expressions [*names*] (that is, "being concatenated with a 'red'", "standing respectively to the left and right of a 'taller than'") adequate for representing possible characteristics of and relations among the objects to which those expressions refer.
4. *Monadology*, #26 (Leibniz 1714/1977), cited in Sellars 1981, p. 342.
5. It is this sense of 'rational', I think, that Jonathan Bennett proposes to isolate and examine in his delightful and insightful little book, *Rationality* (1964).

Connectionism and Cognitive Architecture: A Critical Analysis

12

Jerry A. Fodor
Zenon W. Pylyshyn
1988

1 Introduction

Connectionist or *PDP* models are catching on. There are conferences and new books nearly every day, and the popular science press hails this new wave of theorizing as a breakthrough in understanding the mind. There are also, inevitably, descriptions of the emergence of connectionism as a Kuhnian "paradigm shift". (See Schneider 1987, for an example of this and for further evidence of the tendency to view connectionism as the "new wave" of cognitive science.) The fan club includes the most unlikely collection of people. Almost everyone who is discontent with contemporary cognitive psychology and current "information processing" models of the mind has rushed to embrace "the connectionist alternative".

When taken as a way of modeling *cognitive architecture*, connectionism really does represent an approach that is quite different from that of the classical cognitive science that it seeks to replace. Classical models of the mind were derived from the structure of Turing and Von Neumann machines. They are not, of course, committed to the details of these machines as exemplified in Turing's original formulation or in typical commercial computers—only to the basic idea that the kind of computing that is relevant to understanding cognition involves operations on symbols (see Newell 1980, 1982; Fodor 1976, 1987; and Pylyshyn 1980, 1984). In contrast, connectionists propose to design systems that can exhibit intelligent behavior without storing, retrieving, or otherwise operating on structured symbolic expressions. The style of processing carried out in such models is thus strikingly unlike what goes on when conventional machines are computing some function.

Connectionist systems are networks consisting of very large numbers of simple but highly interconnected "units". Certain assumptions are generally made both about the units and the connections. Each unit is assumed to receive real-valued activity (either excitatory or inhibitory or both) along its input lines. Typically the units do little more than sum this activity and change their state as a function (usually a threshold function) of this sum. Each connection is allowed to modulate the activity it transmits as a function of an intrinsic (but modifiable) property called its "weight". Hence the activity on an input line is typically some non-linear function of the state of activity of its sources. The behavior of the network as a whole is a function of the initial state of activation of the units and of the weights on its connections, which serve as its only form of memory.

Numerous elaborations of this basic connectionist architecture are possible. For example, connectionist models often have stochastic mechanisms for determining the level of activity or the state of a unit. Moreover, units may be connected to outside environments. In this case the units are sometimes assumed to respond to a narrow range of combinations of parameter values and are said to have a certain "receptive field" in parameter-space. These are called "value units" (Ballard 1986). In some versions of connectionist architecture, environmental properties are encoded by the pattern of states of entire populations of units. Such "coarse coding" techniques are among the ways of achieving what connectionists call "distributed representation".[1] The term 'connectionist model' (like 'Turing Machine' or 'Von Neumann machine') is thus applied to a family of mechanisms that differ in details but share a galaxy of architectural commitments. We shall return to the characterization of these commitments below.

Connectionist networks have been analyzed extensively—in some cases using advanced mathematical techniques. They have also been simulated on computers and shown to exhibit interesting aggregate properties. For example, they can be "wired" to recognize patterns, to exhibit rule-like behavioral regularities, and to realize virtually any mapping from patterns of (input) parameters to patterns of (output) parameters—though in most cases multi-parameter, multi-valued mappings require very large numbers of units. Of even greater interest is the fact that such networks can be made to learn; this is achieved by modifying the weights on the connections as a function of certain kinds of feedback (the exact way in which this is done constitutes a

preoccupation of connectionist research and has lead to the development of such important techniques as "back propagation").

In short, the study of connectionist machines has led to a number of striking and unanticipated findings; it's surprising how much computing can be done with a uniform network of simple interconnected elements. Moreover, these models have an appearance of neural plausibility that classical architectures are sometimes said to lack. Perhaps, then, a new cognitive science based on connectionist networks should replace the old cognitive science based on classical computers. Surely this is a proposal that ought to be taken seriously; if it is warranted, it implies a major redirection of research.

Unfortunately, however, discussions of the relative merits of the two architectures have thus far been marked by a variety of confusions and irrelevances. It's our view that when you clear away these misconceptions, what's left is a real disagreement about the nature of mental processes and mental representations. But it seems to us that it is a matter that was substantially put to rest about thirty years ago; and the arguments that then appeared to militate decisively in favor of the classical view appear to us to do so still.

In the present paper we will proceed as follows. First, we discuss some methodological questions about levels of explanation that have become enmeshed in the substantive controversy over connectionism. Second, we try to say what it is that makes connectionist and classical theories of mental structure incompatible. Third, we review and extend some of the traditional arguments for the classical architecture. Though these arguments have been somewhat recast, very little that we'll have to say here is entirely new. But we hope to make it clear how various aspects of the classical doctrine cohere and why rejecting the classical picture of reasoning leads connectionists to say the very implausible things they do about logic and semantics. In section 4, we return to the question what makes the connectionist approach appear attractive to so many people. In doing so we'll consider some arguments that have been offered in favor of connectionist networks as general models of cognitive processing.

1.1 Levels of explanation

There are two major traditions in modern theorizing about the mind, one that we'll call 'representationalist' and one that we'll call 'eliminativist'. Representationalists hold that postulating representational (or 'intentional' or 'semantic') states is essential to a theory of cognition;

according to representationalists, there are states of the mind which function to encode states of the world. Eliminativists, by contrast, think that psychological theories can dispense with such semantic notions as representation. According to eliminativists the appropriate vocabulary for psychological theorizing is neurological or, perhaps behavioral, or perhaps syntactic; in any event, not a vocabulary that characterizes mental states in terms of what they represent. (For a neurological version of eliminativism, see Patricia Churchland 1986; for a behavioral version, see Watson 1930; for a syntactic version, see Stich 1983).

Connectionists are on the representationalist side of this issue. As Rumelhart and McClelland say, PDPs "are explicitly concerned with the problem of internal representation" (1986, p. 121). Correspondingly, the specification of what the states of a network *represent* is an essential part of a connectionist model. Consider, for example, the well-known connectionist account of the bi-stability of the Necker cube (Feldman and Ballard 1982). "Simple units representing the visual features of the two alternatives are arranged in competing coalitions, with inhibitory ... links between rival features and positive links within each coalition... The result is a network that has two dominant stable states" (see figure 12.1). Notice that, in this as in all other such

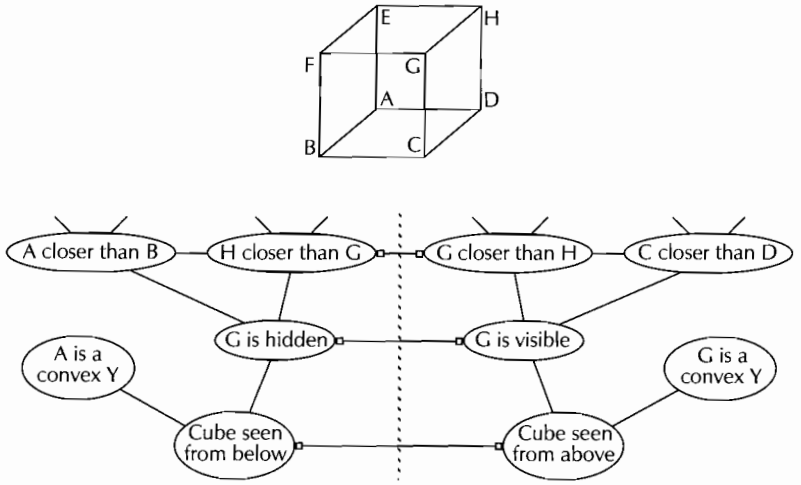


Figure 12.1: A connectionist network illustrating the two stable states of the Necker cube. Plain links are mutually supportive; links with circles are inhibitive. The dashed line separates the two stable states. (Adapted from Feldman and Ballard 1982.)

connectionist models, the commitment to mental representation is explicit: the label of a node is taken to express the representational content of the state that the device is in when the node is excited, and there are nodes corresponding to monadic and to relational properties of the reversible cube when it is seen in one way or the other.

There are, to be sure, times when connectionists appear to vacillate between representationalism and the claim that the "cognitive level" is dispensable in favor of a more precise and biologically-motivated level of theory. In particular, there is a lot of talk in the connectionist literature about processes that are "subsymbolic"—and therefore presumably *not* representational. But this is misleading: connectionist modeling is consistently representationalist in practice, and representationalism is generally endorsed by the very theorists who also like the idea of cognition "emerging from the subsymbolic". Thus, Rumelhart and McClelland (1986, p. 121) insist that PDP models are "strongly committed to the study of representation and process". Similarly, though Smolensky (1988, p. 2) takes connectionism to articulate regularities at the "subsymbolic level" of analysis, it turns out that subsymbolic states do have a semantics—though it's not the semantics of representations at the "conceptual level". According to Smolensky, the semantical distinction between symbolic and subsymbolic theories is just that "entities that are typically represented in the symbolic paradigm by [single] symbols are typically represented in the subsymbolic paradigm by a large number of subsymbols". Both the conceptual and the subsymbolic levels thus postulate representational states, but subsymbolic theories slice them thinner.

We are stressing the representationalist character of connectionist theorizing because much connectionist methodological writing has been preoccupied with the question: What level of explanation is appropriate for theories of cognitive architecture? (See, for example, the exchange between Broadbent 1985 and Rumelhart and McClelland 1985.) And, as we're about to see, what one says about the levels question depends a lot on what stand one takes about whether there are representational states.

It seems certain that the world has causal structure at very many different levels of analysis, with the individuals recognized at the lowest levels being, in general, very small and the individuals recognized at the highest levels being, in general, very large. Thus there is a scientific story to be told about quarks; and a scientific story to be told about atoms; and a scientific story to be told about molecules, ... ditto rocks

and stones and rivers, ... ditto galaxies. And the story that scientists tell about the causal structure that the world has at any one of these levels may be quite different from the story that they tell about its causal structure at the next level up or down. The methodological implication for psychology is this: if you want to have an argument about *cognitive* architecture, you have to specify the level of analysis that's supposed to be at issue.

If you're *not* a representationalist, this is quite tricky since it is then not obvious what makes a phenomenon cognitive. But specifying the level of analysis relevant for theories of cognitive architecture is no problem for either classicists or connectionists. Since classicists and connectionists are both representationalists, for them any level at which states of the system are taken to encode properties of the world counts as a *cognitive* level; and no other levels do. (Representations of "the world" include of course, representations of symbols; for example, the concept WORD is a construct at the cognitive level because it represents something, namely words.) Correspondingly, it's the architecture of representational states and processes that discussions of *cognitive architecture* are about. Put differently, the architecture of the cognitive system consists of the set of basic operations, resources, functions, principles, and so on, (generally the sorts of properties that would be described in a "user's manual" for that architecture if it were available on a computer) whose domain and range are the *representational states* of the organism.

It follows that, if you want to make good the connectionist theory *as a theory of cognitive architecture*, you have to show that the processes which operate on *the representational states* of an organism are those which are specified by a connectionist architecture. It is, for example, *no use at all*, from the cognitive psychologist's point of view, to show that the *non*representational (for instance, neurological, or molecular, or quantum mechanical) states of an organism constitute a connectionist network, because that would *leave open* the question whether the mind is a such a network *at the psychological level*. It is, in particular, perfectly possible that nonrepresentational neurological states are interconnected in the ways described by connectionist models *but that the representational states themselves are not.* This is because, just as it is possible to implement a *connectionist* cognitive architecture in a network of causally interacting nonrepresentational elements, so too it is perfectly possible to implement a *classical* cognitive architecture in such a network. In fact, the question whether connectionist networks

should be treated as models at the implementation level is moot, and will be discussed at some length in section 4.

It is important to be clear about this matter of levels on pain of simply trivializing the issues about cognitive architecture. Consider, for example, the following remark of Rumelhart's (1984, p. 60).

> It has seemed to me for some years now that there must be a unified account in which the so-called rule-governed and [the] exceptional cases were dealt with by a unified underlying process—a process which produces rule-like and rule-exception behavior through the application of a single process ... [In this process] both the rule-like and non-rule-like behavior is a product of the interaction of a very large number of "subsymbolic" processes.

It's clear from the context that Rumelhart takes this idea to be very tendentious; one of the connectionist claims that classical theories are required to deny.

But in fact it's not. For, *of course* there are "subsymbolic" interactions that implement both rule-like and rule-violating behavior; for example, quantum mechanical processes do. *That's* not what classical theorists deny; indeed, it's not denied by anybody who is even vaguely a materialist. Nor does a classical theorist deny that rule-following and rule-violating behaviors are both implemented by the very same neurological machinery. For a classical theorist, neurons implement *all* cognitive processes in precisely the same way: namely, by supporting the basic operations that are required for symbol processing.

What *would* be an interesting and tendentious claim is that there's no distinction between rule-following and rule-violating mentation *at the cognitive or representational or symbolic level*; specifically, that it is not the case that the etiology of rule-following behavior is mediated by the representation of explicit rules. We will consider this idea in section 4, where we will argue that it too is *not* what divides classical from connectionist architecture; classical models *permit* a principled distinction between the etiologies of mental processes that are explicitly rule governed and mental processes that aren't; but they don't demand one.

In short, the issue between classical and connectionist architecture is not about the explicitness of rules; as we'll presently see, classical architecture is not, per se, committed to the idea that explicit rules mediate the etiology of behavior. And it is not about the reality of representational states; classicists and connectionists are all representational realists. And it is not about nonrepresentational architecture; a

connectionist neural network can perfectly well implement a classical architecture at the cognitive level.

So, then, what *is* the disagreement between classical and connectionist architecture about?

2 The nature of the dispute

Classicists and connectionists all assign semantic content to *something*. Roughly, connectionists assign semantic content to 'nodes' (that is, to units or aggregates of units; see footnote 1)—to the sorts of things that are typically labeled in connectionist diagrams; whereas classicists assign semantic content to *expressions*—to the sorts of things that get written on the tapes of Turing machines and stored at addresses in Von Neumann machines. But classical theories disagree with connectionist theories about what primitive relations hold among these content-bearing entities. Connectionist theories acknowledge *only causal connectedness* as a primitive relation among nodes; when you know how activation and inhibition flow among them, you know everything there is to know about how the nodes in a network are related. By contrast, classical theories acknowledge not only causal relations among the semantically evaluable objects that they posit, but also a range of structural relations, of which constituency is paradigmatic.

This difference has far reaching consequences for the ways that the two kinds of theories treat a variety of cognitive phenomena, some of which we will presently examine at length. But, underlying the disagreements about details are two architectural differences between the theories.

(1) **Combinatorial syntax and semantics form mental representations.** Classical theories—but not connectionist theories—postulate a "language of thought" (see, for example, Fodor 1975); they take mental representations to have *a combinatorial syntax and semantics*, in which (a) there is a distinction between structurally-atomic and structurally-molecular representations; (b) structurally-molecular representations have syntactic constituents that are themselves either structurally-molecular or are structurally-atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure. For purposes of convenience, we'll sometime abbreviate (a)–(c) by speaking of classical theories as committed to "complex" mental representations or to "symbol structures".

(2) **Structure sensitivity of processes.** In classical models, the principles by which mental states are transformed, or by which an input selects the corresponding output, are defined over structural properties of mental representations. Because classical mental *representations* have combinatorial structure, it is possible for classical mental *operations* to apply to them by reference to their *form*. The result is that a paradigmatic classical mental process operates upon any mental representation that satisfies a given structural description, and transforms it into a mental representation that satisfies another structural description. (So, for example, in a model of inference, one might recognize an operation that applies to any representation of the form $P \& Q$ and transforms it into a representation of the form P.) Notice that, since formal properties can be defined at a variety of levels of abstraction, such an operation can apply equally to representations that differ widely in their structural complexity. The operation that applies to representations of the form $P \& Q$ to produce P is satisfied by, for example, an expression like '$(A \vee B \vee C) \& (D \vee E \vee F)$', from which it derives the expression '$(A \vee B \vee C)$'.

We take (1) and (2) as the claims that define classical models, and we take these claims quite literally; they constrain the physical realizations of symbol structures. In particular, the symbol structures in a classical model are assumed to correspond to real physical structures in the brain and the *combinatorial structure* of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. For example, the relation *part-of*, which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states. This is why Newell (1980) speaks of computational systems such as brains and classical computers as "*physical* symbols systems".

This bears emphasis because the classical theory is committed not only to there being a system of physically instantiated symbols, but also to the claim that the physical properties onto which the structure of the symbols is mapped *are the very properties that cause the system to behave as it does*. In other words, the physical counterparts of the symbols, and their structural properties, *cause* the system's behavior. A system which has symbolic expressions, but whose operation does not depend upon the structure of these expressions, does not qualify as a classical machine since it fails to satisfy condition (2). In this respect, a classical model is very different from one in which behavior is caused

by mechanisms, such as energy minimization, that are not responsive to the physical encoding of the structure of representations.

From now on, when we speak of "classical" models, we will have in mind *any* model that has complex mental representations, as characterized in (1) and structure-sensitive mental processes, as characterized in (2). Our account of classical architecture is therefore neutral with respect to such issues as whether or not there is a separate executive. For example, classical machines can have an "object-oriented" architecture, like that of the computer language *Smalltalk*, or a "message-passing" architecture, like that of Hewett's (1977) *Actors*—so long as the objects or the messages have a combinatorial structure which is causally implicated in the processing. Classical architecture is also neutral on the question whether the operations on the symbols are constrained to occur one at a time or whether many operations can occur at the same time.

Here, then, is the plan for what follows. In the rest of this section, we will sketch the connectionist proposal for a computational architecture that does away with complex mental representations and structure-sensitive operations. (Although our purpose here is merely expository, it turns out that describing exactly what connectionists are committed to requires substantial reconstruction of their remarks and practices. Since there is a great variety of points of view within the connectionist community, we are prepared to find that some connectionists in good standing may not fully endorse the program when it is laid out in what we take to be its bare essentials.) Following this general expository (or reconstructive) discussion, section 3 provides a series of arguments favoring the classical story. Then the remainder of the paper considers some of the reasons why connectionism appears attractive to many people and offers further general comments on the relation between the classical and the connectionist enterprise.

2.1 Complex mental representations

To begin with, consider a case of the most trivial sort; two machines, one classical in spirit and one connectionist. Here is how the connectionist machine might reason. There is a network of labelled nodes as in figure 12.2. Paths between the nodes indicate the routes along which activation can spread (that is, they indicate the consequences that exciting one of the nodes has for determining the level of excitation of the others). Drawing an inference from **A & B** to **A** thus corresponds to an excitation of node 2 being caused by an excitation of

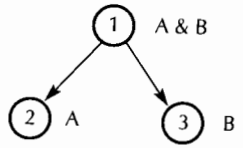


Figure 12.2: A possible connectionist network for drawing inferences from A & B to A or to B.

node 1 (alternatively, if the system is in a state in which node 1 is excited, it eventually settles into a state in which node 2 is excited).

Now consider a classical machine. This machine has a tape on which it writes expressions. Among the expressions that can appear on this tape are: '**A**', '**B**', '**A & B**', '**C**', '**D**', '**C & D**', '**A & C & D**', and so on. The machine's causal constitution is as follows: whenever a token of the form *P & Q* appears on the tape, the machine writes a token of the form *P*. An inference from **A & B** to **A** thus corresponds to a tokening of type '**A & B**' on the tape causing a tokening of type '**A**'.

So then, what does the architectural difference between the machines consist in? In the classical machine, the objects to which the content **A & B** is ascribed (namely, tokens of the expression '**A & B**') literally contain, as proper parts, objects to which the content **A** is ascribed (namely, tokens of the expression '**A**'.) Moreover, the semantics (the satisfaction condition, say) of the expression '**A & B**' is determined in a uniform way by the semantics of its constituents. By contrast, in the connectionist machine none of this true; the object to which the content **A & B** is ascribed (node 1) is *causally* connected to the object to which the content **A** is ascribed (node 2); but there is no structural (for instance, no part/whole) relation that holds between them. In short, it is characteristic of classical systems, but not of connectionist systems, to exploit arrays of symbols, some of which are atomic (such as '**A**'), but indefinitely many of which have other symbols as syntactic and semantic parts (as does '**A & B**').

2.1.4* Representations as "distributed" over microfeatures

Many connectionists hold that the mental representations that correspond to common-sense concepts (CHAIR, JOHN, CUP, and so on) are "distributed" over galaxies of lower-level units which themselves

* *Editor's note:* Section numbers have been retained from the original, and hence are not always sequential in this abridged edition.

have representational content. To use common connectionist terminology (see Smolensky 1988), the higher or "conceptual-level" units correspond to vectors in a "subconceptual" space of microfeatures. The model here is something like the relation between a defined expression and its defining feature analysis; thus, the concept BACHELOR might be thought to correspond to a vector in a space of features that includes ADULT, HUMAN, MALE, and MARRIED—in particular, as an assignment of the value + to the first three features, and of − to the last. Notice that distribution over microfeatures (unlike distribution over neural units) is a relation among representations, hence a relation at the cognitive level.

On the most frequent connectionist accounts, theories articulated in terms of microfeature vectors are supposed to show how concepts are *actually* encoded, hence the feature vectors are intended to *replace* "less precise" specifications of macrolevel concepts. For example, where a classical theorist might recognize a psychological state of entertaining the concept CUP, a connectionist may acknowledge only a *roughly analogous* state of tokening the corresponding feature vector. (One reason that the analogy is only rough is that which feature vector "corresponds" to a given concept may be viewed as heavily context dependent.) The generalizations that "concept-level" theories frame are thus taken to be only approximately true, the exact truth being stateable only in the vocabulary of the microfeatures. Smolensky, for example, is explicit in endorsing this picture: "Precise, formal descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level."[2] This treatment of the relation between common-sense concepts and microfeatures is exactly analogous to the standard connectionist treatment of rules; in both cases, macrolevel theory is said to provide a vocabulary adequate for formulating generalizations that roughly approximate the facts about behavioral regularities. But the constructs of the macrotheory do *not* correspond to the causal mechanisms that generate these regularities. If you want a theory of these mechanisms, you need to replace talk about rules and concepts with talk about nodes, connections, microfeatures, vectors, and the like.

Now, it is among the major misfortunes of the connectionist literature that the issue about whether common-sense concepts should be represented by sets of microfeatures has gotten thoroughly mixed up with the issue about combinatorial structure in mental representations. The crux of the mixup is the fact that sets of microfeatures can overlap,

so that, for example, if a microfeature corresponding to '+has-a-handle' is part of the array of nodes over which the common-sense concept CUP is distributed, then you might think of the theory as representing '+has-a-handle' as a *constituent* of the concept CUP; from which you might conclude that connectionists have a notion of constituency after all—contrary to the claim that connectionism is not a language-of-thought architecture. (See Smolensky 1988).

A moment's consideration will make it clear, however, that even on the assumption that concepts are distributed over microfeatures, '+has-a-handle' is not a constituent of CUP in anything like the sense that 'Mary' (the word) is a constituent of (the sentence) 'John loves Mary'. In the former case, "constituency" is being (mis)used to refer to a semantic relation between predicates; roughly, the idea is that macrolevel predicates like CUP are defined by sets of microfeatures like 'has-a-handle', so that it's some sort of semantic truth that CUP applies to a subset of what 'has-a-handle' applies to. Notice that while the extensions of these predicates are in a set/subset relation, the predicates themselves are not in any sort of part-to-whole relation. The expression 'has-a-handle' isn't *part of* the expression CUP any more than the English phrase 'is an unmarried man' is part of the English phrase 'is a bachelor'.

So far as we know, there are no worked-out attempts in the connectionist literature to deal with the syntactic and semantical issues raised by relations of real constituency. There is, however, a proposal that comes up from time to time: namely, that what are traditionally treated as complex symbols should actually be viewed as just sets of units, with the role relations that traditionally get coded by constituent structure represented by units belonging to these sets. So, for example, the mental representation corresponding to the belief that John loves Mary might be the feature vector ⟨*+John-subject; +loves; +Mary-object*⟩. Here 'John-subject' 'Mary-object' and the like, are the labels of units; that is, they are primitive (or: micro-) features, whose status is analogous to 'has-a-handle'. In particular, they have no internal syntactic or semantic structure, and there is no relation (except the orthographic one) between the feature 'Mary-object' that occurs in the set ⟨John-subject; loves; Mary-object'⟩ and the feature 'Mary-subject' that occurs in the set ⟨Mary-subject; loves; John-object⟩.

As we understand it, the proposal really has two parts. On the one hand, it is suggested that, although connectionist representations cannot exhibit real constituency, nevertheless the classical distinction

between complex symbols and their constituents can be replaced by the distinction between feature sets and their subsets; and, on the other hand, it is suggested that role relations can be captured by features. We'll consider these ideas in turn.

(1) Instead of having sentences like "John loves Mary" in the representational system, you have feature sets like ⟨*+John-subject; +loves; +Mary-object*⟩. Since this set has ⟨*+John-subject*⟩, ⟨*+loves; +Mary-object*⟩, and so forth, as subsets, it may be supposed that the force of the constituency relation has been captured by employing the subset relation.

However, it's clear that this idea won't work since not all subsets of features correspond to genuine constituents. For example, among the subsets of ⟨*+John-subject; +loves; +Mary-object*⟩ are the sets ⟨*+John-subject; +Mary-object*⟩ and the set ⟨*+John-subject; +loves*⟩ which do not, of course, correspond to constituents of the complex symbol "John loves Mary".

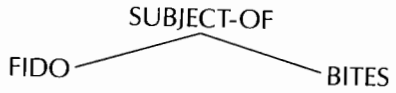
(2) Instead of defining roles in terms of relations among constituents, as one does in classical architecture, introduce them as primitive features.

Consider a system in which the mental representation that is entertained when one believes that John loves Mary is the feature set ⟨*+John-subject; +loves; +Mary-object*⟩. What representation corresponds to the belief that John loves Mary and Bill hates Sally? Suppose, pursuant to the present proposal, that it's the set ⟨*+John-subject; +loves; +Mary-object; +Bill-subject; +hates; +Sally-object*⟩. We now have the problem of distinguishing that belief from the belief that John loves Sally and Bill hates Mary, and from the belief that John hates Mary and Bill loves Sally, and from the belief John hates Mary and Sally and Bill loves Mary, and so on, since these other beliefs will all correspond to precisely the same set of features. The problem is, of course, that nothing in the representation of Mary as *+Mary-object* specifies whether it's the loving or the hating that she is the object of; similarly, mutatis mutandis, with the representation of John as *+John-subject.*

It's important to see that this problem arises precisely because the theory is trying to use sets of atomic representations to do a job that you really need complex representations for. Thus, the question we're wanting to answer is: Given the total set of nodes active at a time, what distinguishes the subvectors that correspond to propositions from the subvectors that don't? This question has a straightforward answer if,

contrary to the present proposal, complex representations are assumed. When representations express concepts that belong to the same proposition, they are not merely simultaneously active, but also *in construction with each other.* By contrast, representations that express concepts that don't belong to the same proposition may be simultaneously active; but they are ipso facto *not* in construction with each other. In short, you need *two* degrees of freedom to specify the thoughts that an intentional system is entertaining at a time: one parameter (active versus inactive) picks out the nodes that express concepts that the system has in mind; the other (in construction versus not) determines how the concepts that the system has in mind are distributed in the propositions that it entertains. For symbols to be "in construction" in this sense is just for them to be constituents of a single complex symbol. Representations that are in construction form parts of a geometrical whole, *where the geometrical relations are themselves semantically significant.* The representation that corresponds to the thought that John loves Fido is not a *set* of concepts but something like a *tree* of concepts, and it's the geometrical relations in this tree that mark (for example) the difference between the thought that John loves Fido and the thought that Fido loves John.

We've occasionally heard it suggested that you could solve the present problem consonant with the restriction against complex representations if you allow networks like this:



The intended interpretation is that the thought that Fido bites corresponds to the simultaneous activation of these nodes; that is, to the vector ⟨+FIDO, +SUBJECT-OF, +BITES⟩—with similar though longer vectors for more complex role relations.

But, on second thought, this proposal merely begs the question that it set out to solve. For, if there's a problem about what justifies assigning the proposition *John loves Fido* as the content of the set ⟨JOHN, LOVES, FIDO⟩, there is surely the same problem about what justifies assigning the proposition *Fido is the subject of bites* to the set ⟨FIDO, SUBJECT-OF, BITES⟩. If this is not immediately clear, consider the case where the simultaneously active nodes are ⟨FIDO,

SUBJECT-OF, BITES, JOHN⟩. Is it the propositional content that Fido bites or that John does?

There are, to reiterate, two questions that you need to answer to specify the content of a mental state: "Which concepts are 'active'?" and "Which of the active concepts are in construction with which others?" Identifying mental states with sets of active nodes provides resources to answer the first of these questions but not the second. That's why the version of network theory that acknowledges sets of atomic representations but no complex representations fails, in indefinitely many cases, to distinguish mental states that are in fact distinct.

But we are *not* claiming that you *can't* reconcile a connectionist architecture with a combinatorial syntax and semantics for mental representations. On the contrary, of course you can. All that's required is that you use your network to implement a Turing machine, and specify a combinatorial structure for its computational language. What it appears that you can't do, however, is have both a combinatorial representational system and a connectionist architecture *at the cognitive level.*

2.2 Structure-sensitive operations

Classicists and connectionists both offer accounts of mental processes, but their theories differ sharply. In particular, the classical theory relies heavily on the notion of the logico/syntactic form of mental representations to define the ranges and domains of mental operations. This notion is, however, unavailable to orthodox connectionists since it presupposes that there are nonatomic mental representations.

The classical treatment of mental processes rests on *two ideas*, each of which corresponds to an aspect of the classical theory of computation. Together they explain why the classical view postulates at least three distinct levels of organization in computational systems: not just a physical level and a semantic (or "knowledge") level, but a syntactic level as well.

The first idea is that it is possible to construct languages in which certain features of the syntactic structures of formulas correspond systematically to certain of their semantic features. Intuitively, the idea is that in such languages the syntax of a formula encodes its meaning—most especially, those aspects of its meaning that determine its role in inference. All the artificial languages that are used for logic have this property, and English has it more or less. Classicists believe that it is a crucial property of the language of thought.

A simple example of how a language can use syntactic structure to encode inferential roles and relations among meanings may help to illustrate this point. Thus, consider the relation between the following two sentences:

(1) John went to the store and Mary went to the store.

(2) Mary went to the store.

On the one hand, from the semantic point of view, (1) entails (2) (so, of course, inferences from (1) to (2) are truth-preserving). On the other hand, from the syntactic point of view, (2) is a constituent of (1). These two facts can be brought into phase by exploiting the principle that sentences with the *syntactic* structure '(S1 and S2)$_S$' entail their sentential constituents. Notice that this principle connects the syntax of these sentences with their inferential roles. Notice too that the trick relies on facts about the grammar of English; it wouldn't work in a language where the formula that expresses the conjunctive content *John went to the store and Mary went to the store* is *syntactically* atomic.

The second main idea underlying the classical treatment of mental processes is that it is possible to devise machines whose function is the transformation of symbols, and whose operations are sensitive to the syntactical structure of the symbols they operate on. This is the classical conception of a computer; it's what the various architectures that derive from Turing and Von Neumann machines all have in common.

Perhaps it's obvious how the two "main ideas" fit together. If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a *syntactically* driven machine whose state transitions satisfy *semantical* criteria of coherence. Such a machine would be just what's required for a mechanical model of the semantical coherence of thought; correspondingly, the idea that the brain *is* such a machine is the foundational hypothesis of classical cognitive science.

So much for the classical story about mental processes. The connectionist story must, of course, be quite different. Since connectionists eschew postulating mental representations with combinatorial syntactic/semantic structure, they are precluded from postulating mental processes that operate on mental representations in a way that is sensitive to their structure. The sorts of operations that connectionist models do have are of two sorts, depending on whether the process under examination is learning or reasoning.

2.2.1 Learning

If a connectionist model is intended to learn, there will be processes that determine the weights of the connections among its units as a function of the character of its training. Typically in a connectionist machine (such as a Boltzmann Machine) the weights among connections are adjusted until the system's behavior comes to model the statistical properties of its inputs. In the limit, the stochastic relations among machine states recapitulate the stochastic relations among the environmental events that they represent.

This should bring to mind the old associationist principle that the strength of association between "ideas" is a function of the frequency with which they are paired "in experience" and the learning-theoretic idea that the strength of a stimulus-response connection is a function of the frequency with which the response is rewarded in the presence of the stimulus. But though connectionists, like other associationists, are committed to learning processes that model statistical properties of inputs and outputs, the simple mechanisms based on co-occurrence statistics that were the hallmarks of old-fashioned associationism have been augmented in connectionist models by a number of technical devices. (Hence the 'new' in 'new connectionism'). For example, some of the earlier limitations of associative mechanisms are overcome by allowing the network to contain "hidden" units (or aggregates) that are not directly connected to the environment, and whose purpose is, in effect, to detect statistical patterns in the activity of the "visible" units including, perhaps, patterns that are more abstract or more "global" than the ones that could be detected by old-fashioned perceptrons.

In short, sophisticated versions of the associative principles for weight setting are on offer in the connectionist literature. The point of present concern, however, is what all versions of these principles have in common with one another and with older kinds of associationism: namely, that these processes are all *frequency*-sensitive. To return to the example discussed above: if a connectionist learning machine converges on a state where it is prepared to infer **A** from **A & B** (that is, to a state in which, when the '**A & B**' node is excited, it tends to settle into a state in which the '**A**' node is excited), the convergence will typically be caused by statistical properties of the machine's training experience (for instance, by correlations between firings of the '**A & B**' node and firings of the '**A**' node, or by correlations of the firings of both with some feedback signal). Like traditional associationism, connectionism treats learning as basically a sort of statistical modeling.

2.2.2 Reasoning

Association operates to alter the structure of a network *diachronically* as a function of its training. Connectionist models also contain a variety of types of 'relaxation' processes which determine the *synchronic* behavior of a network; specifically, they determine what output the device provides for a given pattern of inputs. In this respect, one can think of a connectionist model as a species of analog machine constructed to realize a certain function. The inputs to the function are (i) a specification of the connectedness of the machine (of which nodes are connected to which); (ii) a specification of the weights along the connections; (iii) a specification of the values of a variety of idiosyncratic parameters of the nodes (such as intrinsic thresholds, time since last firing, and the like); (iv) a specification of a pattern of excitation over the input nodes. The output of the function is a specification of a pattern of excitation over the output nodes; intuitively, the machine chooses the output pattern that is most highly associated to its input.

Much of the mathematical sophistication of connectionist theorizing has been devoted to devising analog solutions to this problem of finding a "most highly associated" output corresponding to an arbitrary input; but, once again, the details needn't concern us. What is important, for our purposes, is another property that connectionist theories share with other forms of associationism. In traditional associationism, the probability that one idea will elicit another is sensitive to the strength of the association between them (including "mediating" associations, if any). And the strength of this association is in turn sensitive to the extent to which the ideas have previously been correlated. Associative strength was not, however, presumed to be sensitive to features of the content or the structure of representations per se. Similarly, in connectionist models, the selection of an output corresponding to a given input is a function of properties of the paths that connect them (including the weights, the states of intermediate units, and so on). And the weights, in turn, are a function of the statistical properties of events in the environment (or, perhaps, of relations between patterns of events in the environment and implicit "predictions" made by the network). But the syntactic/semantic structure of the representation of an input is *not* presumed to be a factor in determining the selection of a corresponding output since, as we have seen, syntactic/semantic structure is not defined for the sorts of representations that connectionist models acknowledge.

To summarize: classical and connectionist theories disagree about the nature of mental representation; for the former, but not for the latter, mental representations characteristically exhibit a combinatorial constituent structure and a combinatorial semantics. Classical and connectionist theories also disagree about the nature of mental processes; for the former, but not for the latter, mental processes are characteristically sensitive to the combinatorial structure of the representations on which they operate.

We take it that these two issues define the present dispute about the nature of cognitive architecture. We now propose to argue that the connectionists are on the wrong side of both.

3 The need for symbol systems: productivity, systematicity, and inferential coherence

Classical psychological theories appeal to the constituent structure of mental representations to explain three closely related features of cognition: its productivity, its systematicity, and its inferential coherence. The traditional argument has been that these features of cognition are, on the one hand, pervasive and, on the other hand, explicable only on the assumption that mental representations have internal structure. This argument—familiar in more or less explicit versions for the last thirty years or so—is still intact, so far as we can tell. It appears to offer something close to a demonstration that an empirically adequate cognitive theory must recognize not just causal relations among representational states but also relations of syntactic and semantic constituency; hence that the mind cannot be, in its general structure, a connectionist network.

3.1 Productivity of thought

There is a classical productivity argument for the existence of combinatorial structure in any rich representational system (including natural languages and the language of thought). The representational capacities of such a system are, by assumption, unbounded under appropriate idealization; in particular, there are indefinitely many propositions which the system can encode. However, this unbounded expressive power must presumably be achieved by finite means. The way to do this is to treat the system of representations as consisting of expressions belonging to a generated set. More precisely, the correspondence between a representation and the proposition it expresses is,

in arbitrarily many cases, built up recursively out of correspondences between parts of the expression and parts of the proposition. But, of course, this strategy can operate only when an unbounded number of the expressions are nonatomic. So linguistic (and mental) representations must constitute *symbol systems*. So the mind cannot be a PDP.

Very often, when people reject this sort of reasoning, it is because they doubt that human cognitive capacities are correctly viewed as productive. In the long run, there can be no a priori arguments for (or against) idealizing to productive capacities; whether you accept the idealization depends on whether you believe that the inference from finite performance to finite capacity is justified, or whether you think that finite performance is typically a result of the interaction of an unbounded competence with resource constraints.

In the meantime, however, we propose to view the status of productivity arguments for classical architectures as moot; we're about to present a different sort of argument for the claim that mental representations need an articulated internal structure. It is closely related to the productivity argument, but it doesn't require the idealization to unbounded competence. Its assumptions should thus be acceptable even to theorists who—like connectionists—hold that the finitistic character of cognitive capacities is intrinsic to their architecture.

3.2 Systematicity of cognitive representation

The form of the argument is this. Whether or not cognitive capacities are really *productive*, it seems indubitable that they are what we shall call *systematic*. And we'll see that the systematicity of cognition provides as good a reason for postulating combinatorial structure in mental representation as the productivity of cognition does. You get, in effect, the same conclusion, but from a weaker premise.

The easiest way to understand what the systematicity of cognitive capacities amounts to is to focus on the systematicity of language comprehension and production. In fact, the systematicity argument for combinatorial structure in *thought* exactly recapitulates the traditional structuralist argument for constituent structure in *sentences*. But we pause to remark upon a point that we'll reemphasize later: linguistic capacity is a paradigm of systematic cognition, but it's wildly unlikely that it's the only example. On the contrary, there's every reason to believe that systematicity is a thoroughly pervasive feature of human and infrahuman mentation.

What we mean when we say that linguistic capacities are *systematic* is that the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others. You can see the force of this if you compare learning languages the way we really do learn them, with learning a language by memorizing an enormous phrase book. The point isn't that phrase books are finite and can therefore exhaustively specify only *non*productive languages; that's true, but we've agreed not to rely on productivity arguments for our present purposes. Our point is rather that you can learn *any part of a phrase book without learning the rest.* Hence, on the phrase book model, it would be perfectly possible to learn that uttering the form of words "Granny's cat is on Uncle Arthur's mat" is the way to say (in English) that Granny's cat is on Uncle Arthur's mat, and yet have no idea at all how to say that it's raining (or, for that matter, how to say that Uncle Arthur's cat is on Granny's mat.) Perhaps it's self-evident that the phrase-book story must be wrong about language acquisition because a speaker's knowledge of his native language is never like that. You don't, for example, find native speakers who know how to say in English that John loves the girl but don't know how to say in English that the girl loves John.

Notice, in passing, that systematicity is a property of the mastery of the syntax of a language, not of its lexicon. The phrase-book model really does fit what it's like to learn the *vocabulary* of English, since when you learn English vocabulary you acquire a lot of basically *independent* capacities. So you might perfectly well learn that using the expression 'cat' is the way to refer to cats and yet have no idea that using the expression 'deciduous conifer' is the way to refer to deciduous conifers. Systematicity, like productivity, is the sort of property of cognitive capacities that you're likely to miss if you concentrate on the psychology of learning and searching lists.

There is, as we remarked, a straightforward (and quite traditional) argument from the systematicity of language capacity to the conclusion that sentences must have syntactic and semantic structure. If you assume that sentences are constructed out of words and phrases, and that many different sequences of words can be phrases of the same type, the very fact that one formula is a sentence of the language will often imply that other formulas must be too. In effect, systematicity follows from the postulation of constituent structure.

Suppose, for example, that it's a fact about English that formulas with the constituent analysis 'NP Vt NP' are well formed; and suppose

that 'John' and 'the girl' are NPs and 'loves' is a Vt. It follows from these assumptions that 'John loves the girl', 'John loves John', 'the girl loves the girl', and 'the girl loves John' must *all* be sentences. It follows too that anybody who has mastered the grammar of English must have linguistic capacities that are systematic in respect of these sentences; he *can't but* assume that all of them are sentences if he assumes that any of them are. Compare the situation on the view that the sentences of English are all atomic. There is, then, no structural analogy between 'John loves the girl' and 'the girl loves John' and hence no reason why understanding one sentence should imply understanding the other—no more than understanding 'rabbit' implies understanding 'tree'.

On the view that the sentences are atomic, the systematicity of linguistic capacities is a mystery; on the view that they have constituent structure, the systematicity of linguistic capacities is what you would predict. So we should prefer the latter view to the former.

We can now, finally, come to the point. The argument from the systematicity of linguistic capacities to constituent structure in *sentences* is quite clear. *But thought is systematic too,* so there is a precisely parallel argument from the systematicity of thought to syntactic and semantic structure in *mental* representations.

What does it mean to say that thought is systematic? Well, just as you don't find people who can understand the sentence 'John loves the girl' but not the sentence 'the girl loves John', so too you don't find people who can *think the thought* that John loves the girl but can't think the thought that the girl loves John. Indeed, in the case of verbal organisms the systematicity of thought *follows from* the systematicity of language if you assume—as most psychologists do—that understanding a sentence involves entertaining the thought that it expresses; on that assumption, nobody *could* understand both of the sentences about John and the girl unless he were able to think both of the thoughts about John and the girl.

But now, if the ability to think that John loves the girl is intrinsically connected to the ability to think that the girl loves John, that fact will somehow have to be explained. For a representationalist (which, as we have seen, connectionists are), the explanation is obvious. Entertaining thoughts requires being in representational states (that is, it requires tokening mental representations). And, just as the systematicity of language shows that there must be structural relations between the sentence 'John loves the girl' and the sentence 'the girl loves John,' so the systematicity of thought shows that there must be structural

relations between the mental representation that corresponds to the thought that John loves the girl and the mental representation that corresponds to the thought that the girl loves John;[3] namely, the two mental representations, like the two sentences, *must be made of the same parts.* But if this explanation is right (and there don't seem to be any others on offer), then mental representations have internal structure and there is a language of thought. So the architecture of the mind is not a connectionist network.

To summarize the discussion so far: productivity arguments infer the internal structure of mental representations from the presumed fact that nobody has a *finite* intellectual competence. By contrast, systematicity arguments infer the internal structure of mental representations from the patent fact that nobody has a *punctate* intellectual competence. Just as you don't find linguistic capacities that consist of the ability to understand sixty-seven unrelated sentences, so too you don't find cognitive capacities that consist of the ability to think seventy-four unrelated thoughts. Our claim is that this isn't, in either case, an accident. A linguistic theory that allowed for the possibility of punctate languages would have gone not *just* wrong, but *very profoundly* wrong. And similarly for a cognitive theory that allowed for the possibility of punctate minds.

3.4 The systematicity of inference

In section 2, we saw that, according to classical theories, the syntax of mental representations mediates between their semantic properties and their causal roles in mental processes. Take a simple case: it's a "logical" principle that conjunctions entail their constituents (so the argument from $P \& Q$ to P and to Q is valid). Correspondingly, it's a psychological law that thoughts that $P \& Q$ tend to cause thoughts that P and thoughts that Q, all else being equal. Classical theory exploits the constituent structure of mental representations to account for both these facts, the first by assuming that the combinatorial semantics of mental representations is sensitive to their syntax and the second by assuming that mental processes apply to mental representations in virtue of their constituent structure.

A consequence of these assumptions is that classical theories are committed to the following striking prediction: inferences that are of similar logical type ought, pretty generally,[4] to elicit correspondingly similar cognitive capacities. You shouldn't, for example, find a kind of mental life in which you get inferences from $P \& Q \& R$ to P but you

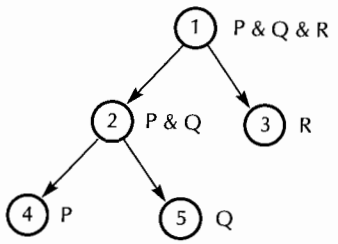


Figure 12.3 A possible connectionist network which draws inferences from $P \& Q \& R$ to P and also draws inferences from $P \& Q$ to P.

don't get inferences from $P \& Q$ to P. This is because, according to the classical account, this logically homogeneous class of inferences is carried out by a correspondingly homogeneous class of psychological mechanisms. The premises of both inferences are expressed by mental representations that satisfy the same syntactic analysis (namely: $S_1 \& S_2 \& S_3 \& \ldots S_n$); and the process of drawing the inference corresponds, in both cases, to the same formal operation of detaching the constituent that expresses the conclusion.

A connectionist can certainly model a mental life in which, if you can reason from $P \& Q \& R$ to P, then you can also reason from $P \& Q$ to P. For example, the network in figure 12.3 would do. But notice that *a connectionist can equally-well model a mental life in which you get one of these inferences and not the other.* In the present case, since there is no structural relation between the $P \& Q \& R$ node and the $P \& Q$ node (remember, all nodes are atomic—don't be misled by the node *labels*) there's no reason why a mind that contains the first should also contain the second, or vice versa. [Thus, figure 12.3 does not contain a $Q \& R$ node. (*Ed.*)] So, the connectionist architecture tolerates gaps in cognitive capacities; it has no mechanism to enforce the requirement that logically homogeneous inferences should be executed by correspondingly homogeneous computational processes.

But, we claim, you don't find cognitive capacities that have these sorts of gaps. You don't, for example, get minds that are prepared to infer *John went to the store* from *John and Mary and Susan and Sally went to the store* and from *John and Mary went to the store* but not from *John and Mary and Susan went to the store.* Given a notion of logical syntax—the very notion that the classical theory of mentation requires to get its account of mental processes off the ground—it is a *truism* that you don't get such minds. Lacking a notion of logical syntax, it is a *mystery* that you don't.

3.5 Summary

What's deeply wrong with connectionist architecture is this. Because it acknowledges neither syntactic nor semantic structure in mental representations, it perforce treats them not as a generated set but as a list. 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*. Cognitive capacities come in structurally related clusters; their systematicity is pervasive. All the evidence suggests that *punctate minds can't happen*. This argument seemed conclusive against the connectionism of Hebb, Osgood and Hull twenty or thirty years ago. So far as we can tell, nothing of any importance has happened to change the situation in the meantime.[5]

A final comment to round off this part of the discussion. It's possible to imagine a connectionist being prepared to admit that, while systematicity doesn't *follow from*—and hence is not explained by—connectionist architecture, it is nonetheless *compatible* with that architecture. It is, after all, perfectly possible to follow a policy of building networks that have *a***R***b* nodes only if they have *b***R***a* nodes, and so on. There is therefore nothing to stop a connectionist from stipulating—as an independent postulate of his theory of mind—that all biologically instantiated networks are, de facto, systematic.

But this misses a crucial point. It's not enough just to stipulate systematicity; one is also required to specify a mechanism that is able to enforce the stipulation. To put it another way, it's not enough for a connectionist to agree that all minds are systematic; he must also explain *how nature contrives to produce only systematic minds*. Presumably there would have to be some sort of mechanism, over and above the ones that connectionism per se posits, the functioning of which insures the systematicity of biologically instantiated networks; a mechanism such that, in virtue of its operation, every network that has an *a***R***b* node also has a *b***R***a* node, and so forth. There are, however, no proposals for such a mechanism. Or, rather, there is just one. The only mechanism that is known to be able to produce pervasive systematicity is classical architecture. And, as we have seen, classical architecture is not compatible with connectionism since it requires internally structured representations.

4 The allure of connectionism

The current popularity of the connectionist approach among psychologists and philosophers is puzzling in view of the sorts of problems raised above—problems which were largely responsible for the development of a syntax-based (proof-theoretic) notion of computation and a Turing-style, symbol-processing notion of cognitive architecture in the first place. There are, however, a number of apparently plausible arguments, repeatedly encountered in the literature, that stress certain limitations of conventional computers as models of brains. These may be seen as favoring the connectionist alternative. Below we will sketch a number of these before discussing the general problems which they appear to raise.

- RAPIDITY OF COGNITIVE PROCESSES RELATIVE TO NEURAL SPEEDS: THE "HUNDRED-STEP" CONSTRAINT. It has been observed (Feldman and Ballard 1982) that neurons take tens of milliseconds to fire. Consequently, in the time it takes people to carry out many of the tasks at which they are fluent (like recognizing a word or a picture, either of which may require considerably less than a second) a *serial* neurally-instantiated program would only be able to carry out about 100 instructions. Yet such tasks might typically require many thousands—or even millions—of instructions in present-day computers (if they can be done at all). Thus, it is argued, the brain must operate quite differently from computers. In fact, the argument goes, the brain must be organized in a highly parallel manner ("massively parallel" is the preferred term of art).
- DIFFICULTY OF ACHIEVING LARGE-CAPACITY PATTERN RECOGNITION AND CONTENT-BASED RETRIEVAL IN CONVENTIONAL ARCHITECTURES. Closely related to the issues about time constraints is the fact that humans can store and make use of an enormous amount of information—apparently without effort (Fahlman and Hinton 1987). One particularly dramatic skill that people exhibit is the ability to recognize patterns from among tens or even hundreds of thousands of alternatives (for instance, word or face recognition). In fact, there is reason to believe that many expert skills may be based on large, fast recognition memories (see Simon and Chase 1973). If one had to search through one's memory serially, the way conventional computers do, the complexity would overwhelm any machine. Thus, the knowledge that people have must be stored and retrieved differently from the way conventional computers do it.

- **Lack of progress in dealing with processes that are nonverbal or intuitive.** Most of our fluent cognitive skills do not consist in accessing verbal knowledge or carrying out deliberate conscious reasoning (Fahlman and Hinton 1987; Smolensky 1988). We appear to know many things that we would have great difficulty in describing verbally, such as how to ride a bicycle, what our close friends look like, and how to recall the name of the President. Such knowledge, it is argued, must not be stored in linguistic form, but in some other "implicit" form. The fact that conventional computers typically operate in a "linguistic mode", inasmuch as they process information by operating on syntactically structured expressions, may explain why there has been relatively little success in modeling implicit knowledge.
- **Acute sensitivity of conventional architectures to damage and noise.** Unlike digital circuits, brain circuits must tolerate noise arising from spontaneous neural activity. Moreover, they must tolerate a moderate degree of damage without failing completely. With a few notable exceptions, if a part of the brain is damaged, the degradation in performance is usually not catastrophic but varies more or less gradually with the extent of the damage. This is especially true of memory. Damage to the temporal cortex (usually thought to house memory traces) does not result in selective loss of particular facts and memories. This and similar facts about brain damaged patients suggest that human memory representations, and perhaps many other cognitive skills as well, are *distributed* spatially, rather than being neurally localized. This appears to contrast with conventional computers, where hierarchical-style control keeps the crucial decisions highly localized and where memory storage consists of an array of location-addressable registers.
- **Conventional rule-based systems depict cognition as "all-or-none".** But cognitive skills appear to be characterized by various kinds of continuities. For example:
  - **Continuous variation in degree of applicability of different principles, or in the degree of relevance of different constraints, "rules", or procedures.** There are frequent cases (especially in perception and memory retrieval), in which it appears that a variety of different constraints are brought to bear on a problem simultaneously and the outcome is a combined effect of all the different factors (see, for example, the informal discussion by McClelland, Rumelhart and Hinton 1986, pp. 3-9). That's why "constraint propagation" techniques are receiving a great deal of attention in artificial intelligence (see Mackworth 1987).

  - **Nondeterminism of human behavior.** Cognitive processes are never rigidly determined or precisely replicable. Rather, they appear to have a significant random or stochastic component. Perhaps that's because there is randomness at a microscopic level, caused by irrelevant biochemical or electrical activity, or perhaps even by quantum mechanical events. To model this activity by rigid deterministic rules can only lead to poor predictions because it ignores the fundamentally stochastic nature of the underlying mechanisms. Moreover, deterministic, all-or-none models will be unable to account for the gradual aspect of learning and skill acquisition.
  - **Failure to display graceful degradation.** When humans are unable to do a task perfectly, they nonetheless do something reasonable. If the particular task does not fit exactly into some known pattern, or if it is only partly understood, a person will not give up or produce nonsensical behavior. By contrast, if a classical rule-based computer program fails to recognize the task, or fails to match a pattern to its stored representations or rules, it usually will be unable to do anything at all. This suggests that, in order to display graceful degradation, we must be able to represent prototypes, match patterns, recognize problems, and so on, in various *degrees.*
- **Conventional models are dictated by current technical features of computers and take little or no account of the facts of neuroscience.** Classical symbol processing systems provide no indication of how the kinds of processes that they postulate could be realized by a brain. The fact that this gap between high-level systems and brain architecture is so large might be an indication that these models are on the wrong track. Whereas the architecture of the mind has evolved under the pressures of natural selection, some of the classical assumptions about the mind may derive from features that computers have only because they are explicitly designed for the convenience of programmers. Perhaps this includes even the assumption that the description of mental processes at the cognitive level can be divorced from the description of their physical realization. At a minimum, by building our models to take account of what is known about neural structures we may reduce the risk of being misled by metaphors based on contemporary computer architectures.

4.1 Replies: why the usual reasons given for preferring a connectionist architecture are invalid

It seems to us that, as arguments against classical cognitive architecture, all these points suffer from one or other of the following two defects.

(1) The objections depend on properties that are not in fact intrinsic to classical architectures, since there can be perfectly natural classical models that don't exhibit the objectionable features. (We believe this to be true, for example, of the arguments that classical rules are explicit and classical operations are "all or none".)

(2) The objections are true of classical architectures insofar as they are implemented on current computers, but need not be true of such architectures when differently (for instance, neurally) implemented. They are, in other words, directed at the implementation level rather than the cognitive level, as these were distinguished in our earlier discussion. (We believe that this is true, for example, of the arguments about speed and resistance to damage and noise.)

In the remainder of this section we will expand on these two points and relate them to some of the arguments presented above. Following this analysis, we will present what we believe may be the most tenable view of connectionism—namely that it is a theory of how (classical) cognitive systems might be implemented, either in real brains or in some "abstract neurology".

4.1.1 Parallel computation and the issue of speed

Consider the argument that cognitive processes must involve large-scale parallel computation. In the form that it takes in typical connectionist discussions, this issue is irrelevant to the adequacy of classical *cognitive* architecture. The "hundred-step constraint", for example, is clearly directed at the implementation level. All it rules out is the (absurd) hypothesis that cognitive architectures are implemented in the brain in the same way as they are implemented on electronic computers.

The absolute speed of a process is a property *par excellence* of its implementation. (By contrast, the *relative* speed with which a system responds to different inputs is diagnostic of distinct processes; but this has always been a prime empirical basis for deciding among alternative

algorithms in information-processing psychology.) Thus, the fact that individual neurons require tens of milliseconds to fire can have no bearing on the predicted speed at which an algorithm will run *unless there is at least a partial, independently motivated, theory of how the operations of the functional architecture are implemented in neurons.* Since, in the case of the brain, it is not even certain that the firing of neurons is invariably the relevant implementation property (at least for higher-level cognitive processes like learning and memory) the hundred-step "constraint" excludes nothing.

Finally, absolute constraints on the number of serial steps that a mental process can require, or on the time that can be required to execute them, provide weak arguments against classical architecture because classical architecture in no way excludes parallel execution of multiple symbolic processes. Indeed, it seems extremely likely that many classical symbolic processes are going on in parallel in cognition, and that these processes interact with one another (for instance, they may be involved in some sort of symbolic constraint propagation). Operating on symbols can even involve "massively parallel" organizations; that might indeed imply new architectures, but they are all *classical* in our sense, since they all share the classical conception of computation as symbol processing. (For examples of serious and interesting proposals on organizing classical processors into large parallel networks, see Hewett's (1977) "Actor" system, Hillis's (1985) "Connection Machine", as well as various recent commercial multi-processor machines.) The point here is that an argument for a network of parallel computers is not in and of itself either an argument against a classical architecture or an argument for a connectionist architecture.

4.1.2 Resistance to noise and physical damage (and the argument for distributed representation)

Some of the other advantages claimed for connectionist architectures over classical ones are just as clearly aimed at the implementation level. For example, the "resistance to physical damage" criterion is so obviously a matter of implementation that it should hardly arise in discussions of cognitive-level theories.

It is true that a certain kind of damage resistance appears to be incompatible with localization, and it is also true that representations in PDPs are distributed over groups of units (at least when "coarse coding" is used). But distribution over units achieves damage resistance only if it entails that representations are also *neurally* distributed.

However, neural distribution of representations is just as compatible with classical architectures as it is with connectionist networks. In the classical case, all you need are memory registers that distribute their contents over physical space. You can get that with fancy storage systems like optical ones, or chemical ones, or even with registers made of connectionist nets.

The physical requirements of a classical symbol-processing system are easily misunderstood. For example, conventional architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent. Since such expressions often have a structure consisting of concatenated parts, the adjacency relation must be instantiated by *some* physical relation when the architecture is implemented. However, since the relation to be physically realized is *functional* adjacency, there is no necessity that physical instantiations of adjacent symbols be *spatially* adjacent. Similarly, although complex expressions are made out of atomic elements, and the distinction between atomic and complex symbols must somehow be physically instantiated, there is no necessity that a token of an atomic symbol be assigned a smaller region in space than a token of a complex symbol—even a token of a complex symbol of which it is a constituent. In classical architectures, as in connectionist networks, functional elements can be physically distributed or localized to any extent whatever.

4.1.3 "Soft" constraints, continuous magnitudes, and stochastic mechanisms

The notion that "soft" constraints, which can vary continuously (as degree of activation does), are incompatible with classical rule-based symbolic systems is another example of the failure to keep the psychological (or symbol-processing) and the implementation levels separate. One can have a classical rule system in which the decision concerning which rule will fire resides in the functional architecture and depends on continuously varying magnitudes. Indeed, this is typically how it is done in practical "expert systems" which, for example, use a Bayesian mechanism in their production-system rule interpreter. The soft or stochastic nature of rule-based processes arises from the interaction of deterministic rules with real-valued properties of the implementation, or with noisy inputs or noisy information transmission.

It should also be noted that rule applications need not issue in "all-or-none" behaviors, since several rules may be activated at once and can have interactive effects on the outcome. Or, alternatively, each of the activated rules can generate independent parallel effects, which might get sorted out later—depending, say, on which of the parallel streams reaches a goal first. An important, though sometimes neglected, point about such aggregate properties of overt behavior as continuity, "fuzziness", randomness, and the like, is that they need not arise from underlying mechanisms that are themselves fuzzy, continuous or random. It is not only possible in principle, but often quite reasonable in practice, to assume that apparently variable or nondeterministic behavior arises from the interaction of multiple deterministic sources.

A similar point can be made about the issue of "graceful degradation". Classical architecture does not require that when the conditions for applying the available rules aren't precisely met, the process should simply fail to do anything at all. As noted above, rules could be activated in some measure depending upon how close their conditions are to holding. Exactly what happens in these cases may depend on how the rule-system is implemented. On the other hand, it could be that the failure to display "graceful degradation" really is an intrinsic limit of the current class of models or even of current approaches to designing intelligent systems. It seems clear that the psychological models now available are inadequate over a broad spectrum of measures, so their problems with graceful degradation may be a special case of their general unintelligence. They may simply not be smart enough to know what to do when a limited stock of methods fails to apply. But this needn't be a principled limitation of classical architectures.

4.1.4 Explicitness of rules

According to McClelland, Feldman, Adelson, Bower and McDermott (1986, p. 6),

> connectionist models are leading to a reconceptualization of key psychological issues, such as the nature of the representation of knowledge ... One traditional approach to such issues treats knowledge as a body of rules that are consulted by processing mechanisms in the course of processing; in connectionist models, such knowledge is represented, often in widely distributed form, in the connections among the processing units.

As we remarked in the Introduction, we think that the claim that most psychological processes are rule-implicit, and the corresponding claim that divergent and compliant behaviors result from the same cognitive

mechanisms, are both interesting and tendentious. We regard these matters as entirely empirical and, in many cases, open. In any case, however, one should not confuse the rule-implicit/rule-explicit distinction with the distinction between classical and connectionist architecture.

This confusion is just ubiquitous in the connectionist literature. It is universally assumed by connectionists that classical models are committed to claiming that regular behaviors must arise from explicitly encoded rules. But this is simply untrue. Not only is there no reason why classical models are required to be rule-explicit but—as a matter of fact—arguments over which, *if any*, rules are explicitly mentally represented have raged for decades *within* the classicist camp. (See, for relatively recent examples, the discussion of the explicitness of grammatical rules in Stabler 1985, and replies; for a philosophical discussion, see Cummins 1983). The one thing that classical theorists do agree about is that it *can't* be that *all* behavioral regularities are determined by explicit rules; at least some of the causal determinants of compliant behavior *must* be *im*plicit. (The arguments for this parallel Lewis Carroll's observations in "What the Tortoise Said to Achilles"; see Carroll 1956). All other questions of the explicitness of rules are viewed by classicists as moot; and every shade of opinion on the issue can be found in the classicist camp.

The basic point is this: not all the functions of a classical computer can be encoded in the form of an explicit program—some of them must be wired in. In fact, the entire program can be hard-wired in cases where it does not need to modify or otherwise examine itself. In such cases, classical machines can be *rule implicit* with respect to their programs, and the mechanism of their state transitions is entirely subcomputational (that is, subsymbolic).

What *does* need to be explicit in a classical machine is not its program but the symbols that it writes on its tapes (or stores in its registers). These, however, correspond not to the machine's rules of state transition but to its data structures. Data structures are *the objects that the machine transforms, not the rules of transformation.* In the case of programs that parse natural language, for example, classical architecture requires the explicit representation of the structural descriptions of sentences, but is entirely neutral on the explicitness of grammars, contrary to what many connectionists believe.

So, then, you can't attack classical theories of cognitive architecture by showing that a cognitive process is rule-implicit; classical architecture *permits* rule-explicit processes but does *not* require them. However, you *can* attack connectionist architectures by showing that a cognitive process is rule-*ex*plicit since, by definition, connectionist architecture precludes the sorts of logico-syntactic capacities that are required to encode rules and the sorts of executive mechanisms that are required to apply them.

4.1.5 On "brain-style" modeling

The relation of connectionist models to neuroscience is open to many interpretations. On the one hand, people like Ballard (1986), and Sejnowski (1981), are explicitly attempting to build models based on properties of neurons and neural organizations, even though the neuronal units in question are idealized (some would say more than a little idealized; see, for example the commentaries following Ballard 1986). On the other hand, Smolensky (1988) views connectionist units as mathematical objects which can be given an interpretation in either neural or psychological terms. Most connectionists find themselves somewhere in between, frequently referring to their approach as "brain-style" theorizing.[6]

Understanding both psychological principles *and* the way that they are neurophysiologically implemented is much better (and, indeed, more empirically secure) than only understanding one or the other. That is not at issue. The question is whether there is anything to be gained by designing "brain-style" models that are uncommitted about how the models map onto brains.

Presumably the point of "brain style" modeling is that theories of cognitive processing should be influenced by the facts of biology (especially neuroscience). The biological facts that influence connectionist models appear to include the following: neuronal connections are important to the patterns of brain activity; the memory "engram" does not appear to be spatially local; to a first approximation, neurons appear to be threshold elements which sum the activity arriving at their dendrites; many of the neurons in the cortex have multidimensional "receptive fields" that are sensitive to a narrow range of values of a number of parameters; the tendency for activity at a synapse to cause a neuron to "fire" is modulated by the frequency and recency of past firings.

Let us suppose that these and similar claims are both true and relevant to the way the brain functions—an assumption that is by no means unproblematic. The question we might then ask is: What

follows from such facts that is relevant to inferring the nature of the cognitive architecture? The unavoidable answer appears to be: very little. That's not an a priori claim. The degree of relationship between facts at different levels of organization of a system is an empirical matter. However, there is reason to be skeptical about whether the sorts of properties listed above are reflected in any more-or-less direct way in the structure of the system that carries out reasoning.

The point is that the structure of "higher levels" of a system are rarely isomorphic, or even similar, to the structure of "lower levels" of a system. No one expects the theory of protons to look very much like the theory of rocks and rivers, even though, to be sure, it is protons and the like that rocks and rivers are "implemented in". Lucretius got into trouble precisely by assuming that there must be a simple correspondence between the structure of macrolevel and microlevel theories. He thought, for example, that hooks and eyes hold the atoms together. He was wrong, as it turns out.

The moral seems to be that one should be deeply suspicious of the heroic sort of brain modeling that purports to address the problems of cognition. We sympathize with the craving for biologically respectable theories that many psychologists seem to feel. But, given a choice, truth is more important than respectability.

4.2 Concluding comments: connectionism as a theory of implementation

A recurring theme in the previous discussion is that many of the arguments for connectionism are best construed as claiming that cognitive architecture is *implemented* in a certain kind of network (of abstract "units"). Understood this way, these arguments are neutral on the question of what the cognitive architecture is. In these concluding remarks we'll briefly consider connectionism from this point of view.

Almost every student who enters a course on computational or information-processing models of cognition must be disabused of a very general misunderstanding concerning the role of the physical computer in such models. Students are almost always skeptical about "the computer as a model of cognition" on such grounds as that "computers don't forget or make mistakes", "computers function by exhaustive search", "computers are too logical and unmotivated", "computers can't learn by themselves, they can only do what they're told", or "computers are too fast (or too slow)", or "computers never get tired or bored", and so on. If we add to this list such relatively more

sophisticated complaints as that "computers don't exhibit graceful degradation" or "computers are too sensitive to physical damage" this list will begin to look much like the arguments put forward by connectionists.

The answer to all these complaints has always been that the *implementation*, and all properties associated with the particular realization of the algorithm that the theorist happens to use in a particular case, is irrelevant to the psychological theory; only the algorithm and the representations on which it operates are intended as a psychological hypothesis. Students are taught the notion of a "virtual machine" and shown that *some* virtual machines *can* learn, forget, get bored, make mistakes, and whatever else one likes, providing one has a theory of the origins of each of the empirical phenomena in question.

Given this principled distinction between a model and its implementation, a theorist who is impressed by the virtues of connectionism has the option of proposing PDPs as theories of implementation. But then, far from providing a revolutionary new basis for cognitive science, these models are in principle neutral about the nature of cognitive processes. In fact, they might be viewed as advancing the goals of classical information-processing psychology by attempting to explain how the brain (or perhaps some idealized brain-like network) might realize the types of processes that conventional cognitive science has hypothesized.

Connectionists do sometimes explicitly take their models to be theories of implementation. Ballard (1986) even refers to connectionism as "the implementational approach". Touretzky (1986) clearly views his BoltzCONS model this way; he uses connectionist techniques to implement conventional symbol processing mechanisms such as push-down stacks and other LISP facilities. Rumelhart and McClelland (1986, p. 117), who are convinced that connectionism signals a radical departure from the conventional symbol processing approach, nonetheless refer to "PDP implementations" of various mechanisms such as attention. Later in the same essay, Rumelhart and McClelland make their position explicit: unlike "reductionists", they believe "that new and useful concepts emerge at different levels of organization". Although they then defend the claim that one should understand the higher levels "... through the study of the interactions among lower level units", the basic idea that there *are* autonomous levels seems implicit everywhere in the essay.

But once one admits that there really are cognitive-level principles distinct from the (putative) architectural principles that connectionism articulates, there seems to be little left to argue about. Clearly it is pointless to ask whether one should or shouldn't do cognitive science by studying "the interaction of lower levels" as opposed to studying processes at the cognitive level, since we surely have to do *both*. Some scientists study geological principles, others study "the interaction of lower level units" like molecules. But since the fact that there are genuine, autonomously-stateable principles of geology is never in dispute, people who build molecular-level models do not claim to have invented a "new theory of geology" that will dispense with all that old fashioned "folk-geological" talk about rocks, rivers, and mountains!

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

5 Conclusion

What, in light of all of this, are the options for the further development of connectionist theories? As far as we can see, there are four routes that they could follow:

(1) Hold out for unstructured mental representations as against the classical view that mental representations have a combinatorial syntax and semantics. Productivity and systematicity arguments make this option appear not attractive.

(2) Abandon network architecture to the extent of opting for structured mental *representations* but continue to insist upon an associationistic account of the nature of mental *processes*. This is, in effect, a retreat to Hume's picture of the mind (see footnote 5), and it has a problem that we don't believe can be solved; although mental representations are, on the present assumption, structured objects, *association is not a structure sensitive relation*. The problem is thus how to reconstruct the semantical coherence of thought without postulating psychological processes that are sensitive to the structure of mental representations. (Equivalently, in more modern terms, it's how to get the causal relations among mental representations to mirror their semantical relations without assuming a proof-theoretic treatment of inference and—more generally—a treatment of semantic coherence that is syntactically expressed, in the spirit of proof theory). This is the problem on which traditional associationism foundered, and the prospects for solving it now strike us as not appreciably better then they were a couple of hundred years ago. To put it a little differently: if you need structure in mental representations anyway to account for the productivity and systematicity of minds, why not postulate mental processes that are structure sensitive to account for the coherence of mental processes? Why not be a classicist, in short.

In any event, notice that the present option gives the classical picture a lot of what it wants: namely, the identification of semantic states with relations to structured arrays of symbols and the identification of mental processes with transformations of such arrays. Notice too that, as things now stand, this proposal is Utopian, since there are no serious proposals for incorporating constituent structure in connectionist architectures.

(3) Treat connectionism as an implementation theory. We have no principled objection to this view (though there are, as connectionists are discovering, technical reasons why networks are often an awkward way to implement classical machines). This option would entail rewriting quite a lot of the polemical material in the connectionist literature, as well as redescribing what the networks are doing as operating on symbol structures, rather than spreading of activation among semantically interpreted nodes.

Moreover, this revision of policy is sure to lose the movement a lot of fans. As we have pointed out, many people have been attracted to the connectionist approach because of its promise to (a) do away with the symbol level of analysis, and

(b) elevate neuroscience to the position of providing evidence that bears directly on issues of cognition. If connectionism is considered simply as a theory of how cognition is neurally implemented, it may constrain cognitive models no more than theories in biophysics, biochemistry, or, for that matter, quantum mechanics do. All of these theories are also concerned with processes that *implement* cognition, and all of them are likely to postulate structures that are quite different from cognitive architecture. The point is that 'implements' is transitive, and it goes all the way down.

(4) Give up on the idea that networks offer (to quote Rumelhart and McClelland 1986, p. 110) "a reasonable basis for modeling cognitive processes in general". It could still be held that they sustain *some* cognitive processes. A good bet might be that they sustain such processes as can be analyzed as the drawing of statistical inferences; as far as we can tell, what network models really are is just analog machines for computing such inferences. Since we doubt that much of cognitive processing does consist of analyzing statistical relations, this would be quite a modest estimate of the prospects for network theory compared to what the connectionists themselves have been offering.

There is an alternative to the empiricist idea that all learning consists of a kind of statistical inference, realized by adjusting parameters; it's the rationalist idea that some learning is a kind of theory construction, effected by framing hypotheses and evaluating them against evidence. We seem to remember having been through this argument before. We find ourselves with a gnawing sense of *deja vu.*

Notes

1. The difference between connectionist networks in which the state of a single unit encodes properties of the world (the so-called "localist" networks) and ones in which the pattern of states of an entire population of units does the encoding (the so-called "distributed-representation" networks) is considered to be important by many people working on connectionist models. Although connectionists debate the relative merits of localist (or "compact") versus distributed representations (for instance, Feldman 1986), the distinction will usually be of little consequence for our purposes, for reasons that we give later. For simplicity, when we wish to refer indifferently to either single-unit codes or aggregate distributed codes, we shall refer to the *nodes* in a network. When the distinction is relevant to our discussion, however, we shall explicitly mark the difference by referring either to units or to aggregates of units.

2. Smolensky (1988, p. 14) remarks that "unlike symbolic tokens, these vectors lie in a topological space, in which some are close together and others are far apart". However, this seems to radically conflate claims about the connectionist model and claims about its implementation (a conflation that is not unusual in the connectionist literature, as we'll see in section 4). If the space at issue is *physical*, then Smolensky is committed to extremely strong claims about adjacency relations in the brain—claims which there is, in fact, no reason at all to believe. But if, as seems more plausible, the space at issue is *semantical*, then what Smolensky says isn't true. Practically any cognitive theory will imply distance measures between mental representations. In classical theories, for example, the distance between two representations is plausibly related to the number of computational steps it takes to derive one representation from the other. In connectionist theories, it is plausibly related to the number of intervening nodes (or to the degree of overlap between vectors, depending on the version of connectionism one has in mind). The interesting claim is not that an architecture offers *a* distance measure but that it offers the *right* distance measure—one that is empirically certifiable.

3. It may be worth emphasizing that the structural complexity of a mental representation is not the same thing as, and does *not* follow from, the structural complexity of its content (that is, of what we're calling "the thought that one has"). Thus, connectionists and classicists can agree to agree that *the thought that P & Q* is complex (and has the thought that *P* among its parts) while agreeing to disagree about whether mental representations have internal syntactic structure.

4. The hedge is meant to exclude cases where inferences of the same logical type nevertheless differ in complexity in virtue of, for example, the length of their premises. The inference from $(A \vee B \vee C \vee D \vee E)$ and $(\neg B \,\&\, \neg C \,\&\, \neg D \,\&\, \neg E)$ to A is of the same logical type as the inference from $A \vee B$ and $\neg B$ to A. But it wouldn't be very surprising, or very interesting, if there were minds that could handle the second inference but not the first.

5. Historical footnote: connectionists are associationists, but not every associationist holds that mental representations must be unstructured. Hume didn't, for example. Hume thought that mental repre-

sentations are rather like pictures, and pictures typically have a compositional semantics. The parts of a picture of a horse are generally pictures of horse parts.

On the other hand, allowing a compositional semantics for mental representations doesn't do an associationist much good, so long as he is true to this spirit of his associationism. The virtue of having mental representations with structure is that it allows for structure-sensitive operations to be defined over them; specifically, it allows for the sort of operations that eventuate in productivity and systematicity. Association is not, however, such an operation; all *it* can do is build an internal model of redundancies in experience by altering the probabilities of transitions among mental states. So far as the problems of productivity and systematicity are concerned, an associationist who acknowledges structured representations is in the position of having the can but not the opener.

Hume, in fact, cheated. He allowed himself not just association but also "imagination", which he takes to be an "active" faculty that can produce new concepts out of old parts by a process of analysis and recombination. (The idea of a unicorn is pieced together out of the idea of a horse and the idea of a horn, for example.) Qua associationist, Hume had, of course, no right to active mental faculties. But allowing imagination in gave Hume precisely what modern connectionists don't have: an answer to the question how mental processes can be productive. The moral is that, if you've got structured representations, the temptation to postulate structure-sensitive operations and an executive to apply them is practically irresistible.

6. The PDP Research Group views its goal as being "to replace the 'computer metaphor' as a model of the mind with the 'brain metaphor'" (Rumelhart, Hinton, and McClelland 1986, p. 75). But the issue is not at all which *metaphor* we should adopt; metaphors (whether "computer" or "brain") tend to be a license to take one's claims as something less than serious hypotheses. As Pylyshyn (1984) points out, the claim that the mind has the architecture of a classical computer is *not* a metaphor but a *literal* empirical hypothesis.

Connectionism, Eliminativism, and the Future of Folk Psychology

13

William Ramsey
Stephen Stich
Joseph Garon

1990

1 Introduction

In the years since the publication of Thomas Kuhn's *Structure of Scientific Revolutions*, the term 'scientific revolution' has been used with increasing frequency in discussions of scientific change, and the magnitude required of an innovation before someone or other is tempted to call it a revolution has diminished alarmingly. Our thesis in this paper is that, if a certain family of connectionist hypotheses turn out to be right, they will surely count as revolutionary, even on stringent pre-Kuhnian standards. There is no question that connectionism has already brought about major changes in the way many cognitive scientists conceive of cognition. However, as we see it, what makes certain kinds of connectionist models genuinely revolutionary is the support they lend to a thorough-going eliminativism about some of the central posits of common-sense (or "folk") psychology. Our focus in this paper will be on beliefs or propositional memories, though the argument generalizes straightforwardly to all the other propositional attitudes. If we are right, the consequences of this kind of connectionism extend well beyond the confines of cognitive science, since these models, if successful, will require a radical reorientation in the way we think about ourselves.

Here is a quick preview of what is to come. Section 2 gives a brief account of what eliminativism claims, and sketches a pair of premises that eliminativist arguments typically require. Section 3 says a bit about how we conceive of common-sense psychology, and the propositional attitudes that it posits. It also illustrates one sort of psychological model that exploits and builds upon the posits of folk psychology. Section 4 is devoted to connectionism. Models that have been called