Connectionism and Cognitive Architecture: A Critical Analysis[1]

Jerry A. Fodor and Zenon W. Pylyshyn
Rutgers Center for Cognitive Science,
Rutgers University,
New Brunswick, NJ

ABSTRACT

This paper explores the difference between Connectionist proposals for cognitive architecture and the sorts of models that have traditionally been assumed in cognitive science. We claim that the major distinction is that, while both Connectionist and Classical architectures postulate representational mental states, the latter but not the former are committed to a symbol-level of representation, or to a 'language of thought': i.e., to representational states that have combinatorial syntactic and semantic structure. Several arguments for combinatorial structure in mental representations are then reviewed. These include arguments based on the 'systematicity' of mental representation: i.e., on the fact that cognitive capacities always exhibit certain symmetries, so that the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents. We claim that such arguments make a powerful case that mind/brain architecture is not Connectionist at the cognitive level. We then consider the possibility that Connectionism may provide an account of the neural (or 'abstract neurological') structures in which Classical cognitive architecture is implemented. We survey a number of the standard arguments that have been offered in favor of Connectionism, and conclude that they are coherent only on this interpretation.

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 (a typical example is the article in the May issue of *Science 86*, called "How we think: A new theory"). There are also, inevitably, descriptions of the emergence of

---

1. This paper is based on a chapter from a forthcoming book. Authors' names are listed alphabetically. We wish to thank the Alfred P. Sloan Foundation for their generous support of this research. The preparation of this paper was also aided by a Killam Research Fellowship and a Senior Fellowship from the Canadian Institute for Advanced Research to ZWP. We also gratefully acknowledge comments and criticisms of earlier drafts by: Professors Noam Chomsky, William Demopoulos, Lila Gleitman, Russ Greiner, Norbert Hornstein, Keith Humphrey, Sandy Pentland, Steven Pinker, David Rosenthal, Edward Stabler.

Reprints may be obtained by writing to either author at the Center for Cognitive Science, Psychology Annex/Busch Campus, Rutgers University, New Brunswick, NJ 08903

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. Connectionism gives solace both to philosophers who think that relying on the pseudo-scientific intentional or semantic notions of folk psychology (like goals and beliefs) mislead psychologists into taking the computational approach (e.g. P.M. Churchland, 1981; P.S. Churchland, 1986; Dennett, 1986); and to those with nearly the opposite perspective, who think that computational psychology is bankrupt because it doesn't address issues of intentionality or meaning (eg. Dreyfus and Dreyfus, in press). On the computer science side, Connectionism appeals to theorists who think that serial machines are too weak and must be replaced by radically new parallel machines (Fahlman and Hinton, 1986), while on the biological side it appeals to those who believe that cognition can only be understood if we study it as neuroscience (e.g., Arbib, 1975; Sejnowski, 1981). It is also attractive to psychologists who think that much of the mind (including the part involved in using imagery) is not discrete (e.g., Kosslyn & Hatfield, 1984), or who think that cognitive science has not paid enough attention to stochastic mechanisms or to "holistic" mechanisms (e.g. Lakoff, 1987), or to stochastic mechanisms, and so on and on. It also appeals to many young cognitive scientists who view the approach as not only ant-establishment (and therefore desirable) but also rigorous and mathematical (see, however, footnote 3). 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; or 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".[2] The term 'Connectionist model' (like 'Turing Machine' or 'Van 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 analysed extensively — in some cases using advanced mathematical techniques.[3] 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

---

2. The difference between Connectionist networks in which the state of a single unit encodes properties of the world (i.e. 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 (e.g. 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.

3. One of the attractions of Connectionism for many people is that it does employ some heavy mathematical machinery, as can be seen from a glance at many of the chapters of the two volume collection by Rumelhart, McClelland and the PDP Research Group (1986). But in contrast to many other mathematically sophisticated areas of cognitive science, such as automata theory or parts of Artificial Intelligence (particularly the study of search, or of reasoning and knowledge representation), the mathematics has not been used to map out the limits of what the proposed class of mechanisms can do. Like a great deal of Artificial Intelligence research, the Connectionist approach remains almost entirely experimental; mechanisms that look interesting are proposed and explored by implementing them on computers and subjecting them to empirical trials to see what they will do. As a consequence, although there is a great deal of mathematical work within the tradition, one has very little idea what various Connectionist networks and mechanisms are good for in general.

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 part four, 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.

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 P.S. 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 & McClelland (1986a) say, PDPs "are explicitly concerned with the problem of internal representation (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 bistability 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 1). Notice that, in this as in all other such 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.

==========
Insert FIGURE 1 about here
==========

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 sub-symbolic'. Thus, Rumelhart & McClelland (1986a) insist that PDP models are "…strongly committed to the study of representation and process (p. 121)" Similarly, though Smolensky (1988) takes Connectionism to articulate regularities at the "sub-symbolic level" of analysis, it turns out that sub-symbolic 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 sub-symbolic theories is just that "entities that are typically represented in the symbolic paradigm by [single] symbols are typically represented in the sub-symbolic paradigm by a large number of sub-symbols (p. 2)"[4]. Both the conceptual and the sub-symbolic levels thus postulate representational states, but sub-symbolic 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 & 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

4. Smolensky seems to think that the idea of postulating a level of representations with a semantics of subconceptual features is unique to network theories. This is an extraordinary view considering the extent to which *Classical* theorists have been concerned with feature analyses in every area of psychology from phonetics to visual perception to lexicography. In fact, the question whether there are 'sub-conceptual' features is *neutral* with respect to the question whether cognitive architecture is Classical or Connectionist.

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, etc (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.[5]

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 (e.g. 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.[6] 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 Part 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: "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 'sub-symbolic' processes." (Rumelhart, 1984, p 60). 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.

---

5. Sometimes, however, even Representationalists fail to appreciate that it is *representation* that distinguishes cognitive from noncognitive levels. Thus, for example, although Smolensky (1988) is clearly a Representationalist, his official answer to the question "What distinguishes those dynamical systems that are cognitive from those that are not?" makes the mistake of appealing to complexity rather than intentionality: "A river… fails to be a cognitive dynamical system only because it cannot satisfy a *large* range of goals under a *large* range of conditions." But, of course, that depends on how you individuate goals and conditions; the river that wants to get to the sea wants first to get half way to the sea, and then to get half way more, …., and so on; quite a lot of goals all told. The real point, of course, is that states that represent goals play a role in the etiology of the behaviors of people but not in the etiology of the 'behavior' of rivers.

6. That Classical architectures can be implemented in networks is not disputed by Connectionists; see for example Rumelhart and McClelland (1986a): "… one can make an arbitrary computational machine out of linear threshold units, including, for example, a machine that can carry out all the operations necessary for implementing a Turing machine; the one limitation is that real biological systems cannot be Turing machines because they have finite hardware (p. 118)".

But in fact it's not. For, *of course* there are 'sub-symbolic' 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: viz 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.[7] We will consider this idea in Part 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 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?

Part 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 2) — i.e. to the sorts of things that are typically labeled in Connectionist diagrams; whereas Classicists assign semantic content to *expressions* — i.e. to the sorts of things that get written on the tapes of Turing machines and stored at addresses in Von Neumann machines.[8]

---

7. There is a different idea, frequently encountered in the Connectionist literature, that this one is easily confused with: viz. that the distinction between regularities and exceptions is merely stochastic (what makes 'went' an irregular past tense is just that the *more frequent* construction is the one exhibited by 'walked'). It seems obvious that if this claim is correct it can be readily assimilated to Classical architecture. See Part 4.

8. This way of putting it will do for present purposes. But a subtler reading of Connectionist theories might take it to be total machine *states* that have content, e.g. the state of *having such and such a node excited*. Postulating connections among labelled nodes would then be equivalent to postulating causal relations among the corresponding content bearing machine states: To say that the excitation of the node labelled 'dog' is caused by the excitation of nodes labelled [d], [o], [g] is to say that the machine's representing its input as consisting of the phonetic sequence [dog] causes it to represent its input as consisting of the word 'dog'. And so forth. Most of the time the distinction between these two ways of talking does not matter for our purposes, so we shall adopt one or the other as convenient.

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".[9]

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 "(AvBvC)&(DvEvF)", from which it derives the expression "(AvBvC)".

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

---

9. Sometimes the difference between simply postulating representational states and postulating representations with a combinatorial syntax and semantics is marked by distinguishing theories that postulate *symbols* from theories that postulate *symbol systems*. The latter theories, but not the former, are committed to a "language of thought". For this usage, see Kosslyn & Hatfield (1984) who take the refusal to postulate symbol systems to be the characteristic respect in which Connectionist architectures differ from Classical architectures. We agree with this diagnosis.

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.[10] 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

---

10. Perhaps the notion that relations among physical properties of the brain instantiate (or encode) the *combinatorial structure* of an expression bears some elaboration. One way to understand what is involved is to consider the conditions that must hold on a mapping (which we refer to as the 'physical instantiation mapping') from expressions to brain states if the causal relations among brain states are to depend on the combinatorial structure of the encoded expressions. In defining this mapping it is not enough merely to specify a physical encoding for each symbol; in order for the *structures* of expressions to have causal roles, structural relations must be encoded by physical properties of brain states (or by sets of functionally equivalent physical properties of brain state).

    Because, in general, Classical models assume that the expressions that get physically instantiated in brains have a generative syntax, the definition of an appropriate mapping has to be built up in terms of (a) the definition of a primitive mapping from atomic symbols to relatively elementary physical states, and (b) a specification of how the structure of complex expressions maps onto the structure of relatively complex or composite physical states. Such a structure-preserving mapping is typically given recursively, making use of the combinatorial system by which complex expressions are built up out of simpler ones. For example, the physical instantiation mapping ***F*** for complex expressions would be defined by recursion, given the definition of ***F*** for *atomic* symbols and given the *structure* of the complex expression, the latter being specified in terms of the 'structure building' rules which constitute the generative syntax for complex expressions. Take, for example, the expression '(A&B)&C'. A suitable definition for a mapping in this case might contain the statement that for any expressions P and Q, ***F***[P&Q] = (***F***[P],***F***[Q]), where the function specifies the physical relation that holds between physical states ***F***[P] and ***F***[Q]. Here the property serves to physically encode, (or 'instantiate') the relation that holds between the expressions P and Q, on the one hand, and the expression P&Q on the other.

    In using this rule for the example above, P and Q would have the values 'A&B' and 'C' respectively, so that the mapping rule would have to be applied twice to pick the relevant physical structures. In defining the mapping recursively in this way we ensure that the relation between the expressions 'A' and 'B', and the composite expression 'A&B', is encoded in terms of a physical relation between constituent states that is identical (or functionally equivalent) to the physical relation used to encode the relation between expressions 'A&B' and 'C', and their composite expression '(A&B)&C'. This type of mapping is well known because of its use in Tarski's definition of an interpretation of a language in a model. The idea of a mapping from symbolic expressions to a structure of physical states is discussed in Pylyshyn (1984, p 54-69), where it is referred to as an 'instantiation function' and in Stabler (1983), where it is called a 'realization mapping'.

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, Part 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.

Complex mental representations

To begin with, consider a case of the most trivial sort; two machines, one Classical in spirit and one Connectionist[11] Here is how the Connectionist machine might reason. There is a network of labelled nodes as in Figure 2.

==========
Insert Figure 2 about here
==========

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 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; see footnote 8).

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

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 (viz. tokens of the expression 'A&B') literally contain, as proper parts, objects to which the content A is ascribed (viz. tokens of the expression 'A'.) Moreover, the semantics (e.g. the satisfaction conditions) of the expression 'A&B' is determined in a uniform way by the semantics of its constituents.[12] By

---

11. This illustration has no any particular Connectionist model in mind, though the caricature presented is, in fact, a simplified version of the Ballard (1987) Connectionist theorem proving system (which actually uses a more restricted proof procedure based on the *unification* of Horn clauses). To simplify the exposition, we assume a 'localist' approach, in which each semantically interpreted node corresponds to a single Connectionist unit; but nothing relevant to this discussion is changed if these nodes actually consist of patterns over a cluster of units.

12. This makes the "compositionality" of data structures a defining property of Classical architecture. But, of course, it leaves open the question of the degree to which *natural* languages (like English) are also compositional.

contrast, in the Connectionist machine none of this true; the object to which the content A&B is ascribed (viz. node 1) is causally connected to the object to which the content A is ascribed (viz. to node 2); but there is no structural (e.g. 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 (e.g. expressions like 'A') but indefinitely many of which have other symbols as syntactic and semantic parts (e.g. expressions like 'A&B').

It is easy to overlook this difference between Classical and Connectionist architectures when reading the Connectionist polemical literature or examining a Connectionist model. There are at least four ways in which one might be lead to do so: (1) by failing to understand the difference between what arrays of symbols do in Classical machines and what node labels do in Connectionist machines; (2) by confusing the question whether the nodes in Connectionist networks have *constituent* structure with the question whether they are *neurologically distributed*; (3) by failing to distinguish between a representation having semantic and syntactic constituents and a concept being encoded in terms of microfeatures, and (4) by assuming that since representations of Connectionist networks have a graph structure, it follows that the nodes in the networks have a corresponding constituent structure. We shall now need rather a long digression to clear up these misunderstandings.

(i) The role of labels in Connectionist theories.

In the course of setting out a Connectionist model, intentional content will be assigned to machine states, and the expressions of some language or other will, of course, be used to express this assignment; for example, nodes may be labelled to indicate their representational content. Such labels often have a combinatorial syntax and semantics; in this respect, they can look a lot like Classical mental representations. The point to emphasize, however, is that it doesn't follow (and it isn't true) that the nodes to which these labels are assigned have a combinatorial syntax and semantics. 'A&B', for example, can be tokened on the tape of the Classical machine *and can also appear as a label in a Connectionist machine* as it does in the diagram above. And, of course, the expression 'A&B' is syntactically and semantically complex: it has a token of 'A' as one of its syntactic constituents, and the semantics of the expression 'A&B' is a function of the semantics of the expression 'A'. But it isn't part of the intended reading of the diagram that node 1 itself has constituents; the node —unlike its label —has no semantically interpreted parts.

It is, in short, important to understand the difference between Connectionist labels and the symbols over which Classical computations are defined. The difference is this: Strictly speaking, the labels play *no role at all* in determining the operation of a Connectionist machine; in particular, the operation of the machine is unaffected by the syntactic and semantic relations that hold among the expressions that are used as labels. To put this another way, the node labels in a Connectionist machine are not part of the causal structure of the machine. Thus, the machine depicted in Figure 2 will continue to make the same state transitions regardless of what labels we assign to the nodes. Whereas, by contrast, the state transitions of Classical machines are causally determined *by the structure —including the constituent structure —of the symbol arrays that the machines transform*: change the symbols and the system behaves quite differently. (In fact, since the behavior of a Classical machine is sensitive to the syntax of the representations it computes

on, even interchanging *synonymous* — semantically equivalent — representations affects the course of computation). So, although the Connectionist's labels and the Classicist's data structures both constitute languages, only the latter language constitutes a medium of computation.[13]

(ii) Distributed representations

The *second* mistake that can lead to a failure to notice that the mental representations in Connectionist models lack combinatorial syntactic and semantic structure is the fact that many Connectionists view representations as being *neurologically distributed*; and, presumably, whatever is distributed must have parts. It doesn't follow, however, that whatever is distributed must have *constituents*; being neurologically distributed is very different from having semantic or syntactic constituent structure.

You have constituent structure when (and only when) the parts of semantically evaluable entities are themselves semantically evaluable. Constituency relations thus hold among objects all of which are at the representational level; they are, in that sense, *within* level relations.[14] By contrast, neural distributedness — the sort of relation that is assumed to hold between 'nodes' and the 'units' by which they are realized — is a *between* level relation: The nodes, but not the units, count as representations. To claim that a node is neurally distributed is presumably to claim that its states of activation correspond to patterns of neural activity — to aggregates of neural 'units' — rather than to activations of single neurons. The important point is that nodes that are distributed in this sense can perfectly well be syntactically and semantically atomic: Complex spatially-distributed implementation in no way implies constituent structure.

There is, however, a different sense in which the representational states in a network might be distributed, and this sort of distribution also raises questions relevant to the constituency issue.

(iii). Representations as 'distributed' over microfeatures

Many Connectionists hold that the mental representations that correspond to commonsense concepts (CHAIR, JOHN, CUP, etc.) are 'distributed' over galaxies of lower level units which themselves have representational content. To use common Connectionist terminology (see Smolensky, 1988), the higher or "conceptual level" units correspond to vectors in a "sub-conceptual" 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,

---

13. Labels aren't part of the casual structure of a Connectionist machine, but they may play an essential role in its causal history inasmuch as designers do, in fact, wire their machines to respect the semantical relations that the labels express. For example, in Ballard's (1987) Connectionist model of theorem proving, there is a mechanical procedure for wiring a network which will carry out proofs by unification. This procedure is a function from a set of node labels to a wired-up machine. There is thus an interesting and revealing respect in which node labels are relevant to the operations that get performed when the function is executed. But, of course, the machine on which the labels have the effect is not the machine whose states they are labels of; and the effect of the labels occurs at the time that the theorem-proving machine is constructed, not at the time its reasoning process are carried out. *This* sort of case of labels 'having effects' is thus quite different from the way that symbol tokens (eg. tokened data structures) can affect the causal processes of a Classical machine.

14. Any relation specified as holding among representational states is, by definition, within the 'cognitive level'. It goes without saying that relations that are 'within-level' by this criterion can count as 'between-level' when we use criteria of finer grain. There is, for example, nothing to prevent hierarchies of levels of representational states.

and MARRIED; i.e. as an assignment of the value + to the first two 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.

Since microfeatures are frequently assumed to be derived automatically (i.e. via learning procedures) from the statistical properties of samples of stimuli, we can think of them as expressing the sorts of properties that are revealed by multivariate analysis of sets of stimuli (e.g. by multidimensional scaling of similarity judgments). In particular, they need not correspond to English words; they can be finer-grained than, or otherwise atypical of, the terms for which a non-specialist needs to have a word. Other than that, however, they are perfectly ordinary semantic features, much like those that lexicographers have traditionally used to represent the meanings of words.

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. (p. 11)"[15] This treatment of the relation between commonsense 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.[16]

---

15. Smolensky (1988) remarks that "unlike symbolic tokens, these vectors lie in a topological space, in which some are close together and others are far apart. (p 14)" 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 Part 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.

16. The primary use that Connectionists make of microfeatures is in their accounts of generalization and abstraction (see, for example, Hinton, McClelland & Rumelhart, 1986). Roughly, you get generalization by using overlap of microfeatures to define a similarity space, and you get abstraction by making the vectors that correspond to *types* be subvectors of the ones that correspond to their *tokens*. Similar proposals have quite a long history in traditional Empiricist analysis; and have been roundly criticized over the centuries. (For a discussion of abstractionism see Geach, 1957; that similarity is a primitive relation —hence not reducible to partial identity of feature sets —was, of course, a main tenet of Gestalt psychology, as well as more recent approaches based on "prototypes"). The treatment of microfeatures in the Connectionist literature would appear to be very close to early proposals by Katz and Fodor (1963) and Katz and Postal (1964), where both the idea of a feature analysis of concepts and the idea that relations of semantical containment among concepts should be identified with set-theoretic relations among feature arrays are explicitly endorsed.

Now, it is among the major misfortunes of the Connectionist literature that the issue about whether commonsense 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 commonsense 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'.

Real constituency does have to do with parts and wholes; the symbol 'Mary' is literally a part of the symbol 'John loves Mary'. It is because their symbols enter into real-constituency relations that natural languages have both atomic symbols and complex ones. By contrast, the definition relation can hold in a language where *all* the symbols are syntactically atomic; e.g. a language which contains both 'cup' and 'has-a-handle' as primitive predicates. This point is worth stressing. The question whether a representational system has real-constituency is independent of the question of microfeature analysis; it arises both for systems in which you have CUP as semantically primitive, and for systems in which the semantic primitives are things like '+ has-a-handle' and CUP and the like are defined in terms of these primitives. It really is very important not to confuse the semantic distinction between primitive expressions and defined expressions with the syntactic distinction between atomic symbols and complex symbols.

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: viz, 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 (i.e. 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-agent; loves; Mary-object'} and the feature 'Mary-subject' that occurs in the set {Mary-subject; loves; John-object}. (See, for example, the discussion in Hinton (1987) of "role-specific descriptors that represent the conjunction of an identity and a role [by the use of

which] we can implement part-whole hierarchies using set intersection as the composition rule." See also, Hinton, McClelland and Rumelhart (1986 p. 82-85) where what appears to be the same treatment is proposed in somewhat different terms).

Since, as we remarked, these sorts of ideas aren't elaborated in the Connectionist literature, detailed discussion is probably not warranted here. But it's worth a word to make clear what sort of trouble you would get into if you were to take them seriously.

As we understand it, the proposal really has two parts: On the one hand, it's 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's 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 sub-sets, 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-actor; +loves; +Mary-patient}). 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-agent; +loves; +Mary-patient; +Bill-agent; +hates; +Sally-patient}*. 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; etc. 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-patient* specifies whether it's the loving or the hating that she is the patient of; similarly, mutatis mutandis, with the representation of John as *+John-actor*.

What has gone wrong isn't disasterous (yet). All that's required is to enrich the system of representations by recognizing features that correspond not to (for example) just being an agent, but rather to being the agent of a loving of Mary (the property that John has when John loves Mary) and being the agent of a hating of Sally (the property that Bill has when Bill hates Sally.) So, the representation of John that's entertained when one believes that John loves Mary and Bill hates Sally might be something like *+John-agent-hates-Mary-object*.

The disadvantage of this proposal is that it requires rather a lot of primitive features.[17] How many? Well, a number of the order of magnitude of the *sentences* of a natural language (whereas one might have hoped to get by with a primitive vocabulary that is not vastly larger than the *lexicon* of a natural language; after all, natural languages do.) We leave it to the reader to estimate the number of primitive features you would need, assuming that there is a distinct belief corresponding to every grammatical sentence of English of up to, say, fifteen words of length, and assuming that there is an average of, say, five roles associated with each belief. (Hint: George Miller once estimated that the number of well-formed twenty word sentences of English is of the order of magnitude of the number of seconds in the history of the universe.)

The alternative to this grotesque explosion of primitives would be to have *a combinatorial syntax and semantics for the features*. But, of course, this is just to give up the game since the syntactic and semantic relations that hold among the parts of the complex feature *+((John agent) loves (Mary object))* are the very same ones that Classically hold among the constituents of the complex symbol "John loves Mary"; these include the role relations which Connectionists had proposed to reconstruct using just sets of primitive features. The idea that we should capture role relations by allowing features like *John-agent* thus turns out to be bankrupt; and there doesn't seem to be any other way to get the force of structured symbols in a Connectionist architecture. Or, if there is, nobody has given any indication of how to do it. This becomes clear once the crucial issue about structure in mental representations is disentangled from the relatively secondary issue about whether the representation of commonsense concepts is 'distributed' (i.e. from questions like whether it's CUP or 'has-a-handle' that is primitive in the language of thought.)

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.*

(iv) Connectionist networks and graph structures

The *fourth* reason that the lack of syntactic and semantic structure in Connectionist representations has largely been ignored may be that Connectionist networks look like general graphs; and it is, of course, perfectly possible to use graphs to describe the internal structure of a complex symbol. That's precisely what linguists do when they use 'trees' to exhibit the constituent structure of sentences. Correspondingly, one could imagine a graph notation that expresses the internal structure of mental representations by using arcs and labelled nodes. So, for example, you might express the syntax of the mental representation that corresponds to the thought that John loves the girl like this:

---

17. Another disadvantage is that, strictly speaking it doesn't work; although it allows us to distinguish the belief that John loves Mary and Bill hates Sally from the belief that John loves Sally and Bill hates Mary, we don't yet have a way to distinguish believing that (John loves Mary because Bill hates Sally) from believing that (Bill hates Sally because John loves Mary). Presumably nobody would want to have primitive features corresponding to these.

John ⟶ loves ⟶ the girl

Under the intended interpretation, this would be the structural description of a mental representation whose content is that John loves the girl, and whose constituents are: a mental representation that refers to *John*, a mental representation that refers to *the girl*, and a mental representation that expresses the two-place relation represented by '⟶ loves ⟶'.

But although graphs can sustain an interpretation as specifying the logical syntax of a complex mental representation, this interpretation is inappropriate for graphs of Connectionist networks. Connectionist graphs are not structural descriptions of mental representations; they're specifications of causal relations. All that a Connectionist can mean by a graph of the form 'X ⟶ Y' is: *states of node X causally affect states of node Y.* In particular, the graph can't mean *'X is a constituent of Y'* or *'X is grammatically related to Y'* etc., since these sorts of relations are, in general, not defined for the kinds of mental representations that Connectionists recognize.

Another way to put this is that the links in Connectionist diagrams are not generalized pointers that can be made to take on different functional significance *by an independent interpreter*, but are confined to meaning something like "sends activation to". The intended interpretation of the links as causal connections is intrinsic to the theory. If you ignore this point, you are likely to take Connectionism to offer a much richer notion of mental representation than it actually does.

So much, then, for our long digression. We have now reviewed one of major respects in which Connectionist and Classical theories differ; viz. their accounts of mental *representations*. We turn to the second major difference, which concerns their accounts of mental *processes*.

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 '$(S1 \text{ and } S2)_S$' 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.[18]

Here is another example. We can reconstruct such truth preserving inferences as *if Rover bites then something bites* on the assumption that (a) the sentence 'Rover bites' is of the syntactic type **F**a, (b) the sentence 'something bites' is of the syntactic type **E**x(**F**x) and (c) every formula of the first type entails a corresponding formula of the second type (where the notion 'corresponding formula' is cashed syntactically; roughly the two formulas must differ only in that the one has an existentially bound variable at the syntactic position that is occupied by a constant in the other.) Once again the point to notice is the blending of syntactical and semantical notions: The rule of existential generalization applies to formulas in virtue of their syntactic form. But the salient property that's preserved under applications of the rule is semantical: What's claimed for the transformation that the rule performs is that it is *truth* preserving.[19]

There are, as it turns out, examples that are quite a lot more complicated than these. The whole of the branch of logic known as proof theory is devoted to exploring them.[20] It would not

---

18. And it doesn't work uniformly for English conjunction. Cf: *John and Mary are friends* ⇏ **John are friends*; or *The flag is red, white and blue* ⇏ *The flag is blue*. Such cases show either that English is not the language of thought, or that, if it is, the relation between syntax and semantics is a good deal subtler for the language of thought than it is for the standard logical languages.

19. It needn't, however, be strict truth-preservation that makes the syntactic approach relevant to cognition. Other semantic properties might be preserved under syntactic transformation in the course of mental processing —e.g., warrant, plausibility, heuristic value, or simply *semantic non-arbitrariness*. The point of Classical modeling isn't to characterize human thought as supremely logical; rather, it's to show how a family of types of semantically coherent (or knowledge-dependent) reasoning are mechanically possible. Valid inference is the paradigm only in that it is the best understood member of this family; the one for which syntactical analogues for semantical relations have been most systematically elaborated.

20. It is not uncommon for Connectionists to make disparaging remarks about the relevance of logic to psychology, even though they accept the idea that inference is involved in reasoning. Sometimes the suggestion seems to be that it's all right if Connectionism can't reconstruct the theory of inference that formal deductive logic provides since it has something even better on offer. For example, in their report to the U.S. National Science Foundation, McClelland, Feldman, Adelson, Bower & McDermott (1986) state that "…connectionist models realize an evidential logic *in contrast to* the symbolic logic of conventional computing (p 6; our emphasis)" and that "evidential logics are becoming increasingly important in cognitive science and have a natural map to connectionist modeling. (p 7)" It is, however, hard to understand the implied contrast since, on the one hand, evidential logic must surely be a fairly conservative extension of "the symbolic logic of conventional computing" (i.e. most of the theorems of the latter have to come out true in the former) and, on the other, there is not the slightest reason to doubt that an evidential logic would 'run' on a Classical machine. Prima facie, the problem about evidential logic isn't that we've got one that we don't know how to implement; it's that we haven't got one.

be unreasonable to describe Classical Cognitive Science as an extended attempt to apply the methods of proof theory to the modeling of thought (and similarly, of whatever other mental processes are plausibly viewed as involving inferences; preeminently learning and perception.) Classical theory construction rests on the hope that syntactic analogues can be constructed for nondemonstrative inferences (or informal, common-sense reasoning) in something like the way that proof theory has provided syntactic analogues for validity.

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 that they operate upon. 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.

(i) 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 'Boltzman 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 recapitulates 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.[21]

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: viz, 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 (i.e., 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: E.g. by correlation between firing of the 'A&B' node and firing of the 'A' node, or by correlations of the firing of both with some feedback signal. Like traditional Associationism, Connectionism treats learning as basically a sort of statistical modeling.

(ii) 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 (e.g. intrinsic thresholds; time since last firing, etc.) (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, etc). And the weights, in turn, are a function of the statistical properties of events in the environment (or of relations between patterns of events in the environment and

21. Compare the "little s's" and "little r's" of neo-Hullean "mediational" Associationists like Charles Osgood.

implicit 'predictions' made by the network, etc.) 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.

Part III: The need for Symbol Systems: Productivity, Systematicity, Compositionality 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 compositionality, 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.

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.[22] However, this unbounded expressive power must presumably be achieved by finite

22. This way of putting the productivity argument is most closely identified with Chomsky (e.g. Chomsky, 1965; 1968). However, one does not have to rest the argument upon a basic assumption of infinite generative capacity. Infinite generative capacity can be viewed, instead, as a consequence or a corollary, of theories formulated so as to capture the greatest number of generalizations with the fewest independent

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 non-atomic. So linguistic (and mental) representations must constitute *symbol systems* (in the sense of footnote 9). 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. Classicists have traditionally offered a mixture of methodological and empirical considerations in favor of the latter view.

From a methodological perspective, the least that can be said for assuming productivity is that it precludes solutions that rest on inappropriate tricks (such as storing all the pairs that define a function); tricks that would be unreasonable in practical terms even for solving finite tasks that place sufficiently large demands on memory. The idealization to unbounded productive capacity forces the theorist to separate the finite specification of a method for solving a computational problem from such factors as the resources that the system (or person) brings to bear on the problem at any given moment.

The empirical arguments for productivity have been made most frequently in connection with linguistic competence. They are familiar from the work of Chomsky (1968) who has claimed (convincingly, in our view) that the knowledge underlying linguistic competence is generative — i.e. that it allows us *in principle* to generate (/understand) an unbounded number of sentences. It goes without saying that no one does, or could, *in fact* utter or understand tokens of more than a finite number of sentence types; this is a trivial consequence of the fact that nobody can utter or understand more than a finite number of sentence tokens. But there are a number of considerations which suggest that, despite de facto constraints on performance, ones knowledge of ones language supports an unbounded productive capacity in much the same way that ones knowledge of addition supports an unbounded number of sums. Among these considerations are, for example, the fact that a speaker/hearer's performance can often be improved by relaxing time constraints, increasing motivation, or supplying pencil and paper. It seems very natural to treat such manipulations as affecting the transient state of the speaker's memory and attention rather than what he knows about —or how he represents —his language. But this treatment is available only on the assumption that the character of the subject's performance is determined by interactions between the available knowledge base and the available computational resources.

Classical theories are able to accommodate these sorts of considerations because they assume architectures in which there is a functional distinction between memory and program. In a system such as a Turing machine, where the length of the tape is not fixed in advance, changes

principles. This more neutral approach is, in fact, very much in the spirit of what we shall propose below. We are putting it in the present form for expository and historical reasons.

in the amount of available memory *can be affected without changing the computational structure of the machine*; viz by making more tape available. By contrast, in a finite state automaton or a Connectionist machine, adding to the memory (e.g. by adding units to a network) alters the connectivity relations among nodes and thus does affect the machine's computational structure. Connectionist cognitive architectures cannot, by their very nature, support an expandable memory, so they cannot support productive cognitive capacities. The long and short is that if productivity arguments are sound, then they show that the architecture of the mind can't be Connectionist. Connectionists have, by and large, acknowledged this; so they are forced to reject productivity arguments.

The test of a good scientific idealization is simply and solely whether it produces successful science in the long term. It seems to us that the productivity idealization has more than earned its keep, especially in linguistics and in theories of reasoning. Connectionists, however, have not been persuaded. For example, Rumelhart and McClelland (1986a) say that they "… do not agree that [productive] capabilities are of the essence of human computation. As anyone who has ever attempted to process sentences like 'The man the boy the girl hit kissed moved' can attest, our ability to process even moderate degrees of center-embedded structure is grossly impaired relative to an ATN [Augmented Transition Network] parser…. What is needed, then, is not a mechanism for flawless and effortless processing of embedded constructions… The challenge is to explain how those processes that others have chosen to explain in terms of recursive mechanisms can be better explained by the kinds of processes natural for PDP networks. (p 119)"

These remarks suggest that Rumelhart and McClelland think that the fact that center-embedding sentences are hard is somehow an *embarrassment* for theories that view linguistic capacities as productive. But of course it's not since, according to such theories, performance is an effect of interactions between a productive competence and restricted resources. There are, in fact, quite plausible Classical accounts of why center-embeddings ought to impose especially heavy demands on resources, and there is a reasonable amount of experimental support for these models. (See, for example, Wanner and Maratsos, 1978.)

In any event, it should be obvious that the difficulty of parsing center-embeddings can't be a consequence of their recursiveness per se since there are many recursive structures that are strikingly easy to understand. Consider: 'this is the dog that chased the cat that ate the rat that lived in the house that Jack built.' The Classicist's case for productive capacities in parsing rests on the transparency of sentences like these.[23] In short, the fact that center embedded sentences are hard perhaps shows that there are some recursive structures that we can't parse. But what Rumelhart and McClelland need if they are to deny the productivity of linguistic capacities is the

23. McClelland & Kawamoto (1986) discuss this sort of recursion briefly. Their suggestion seems to be that parsing such sentences doesn't really require recovering their recursive structure: "…the job of the parser [with respect to right-recursive sentences] is to spit out phrases in a way that captures their *local* context. Such a representation may prove sufficient to allow us to reconstruct the correct bindings of noun phrases to verbs and prepositional phrases to *nearby* nouns and verbs" (p 324; emphasis ours). It is, however, by no means the case that all of the semantically relevant grammatical relations in readily intelligible embedded sentences are local in surface structure. Consider: '*Where* did the man who owns the cat that chased the rat that frightened the girl say that he was going to move to (X)?' or '*What* did the girl that the children loved to listen to promise your friends that she would read (X) to them?' Notice that, in such examples, a binding element (italicized) can be arbitrarily displaced from the position whose interpretation it controls (marked 'X') without making the sentence particularly difficult to understand. Notice too that the 'semantics' doesn't determine the binding relations in either example.

much stronger claim that there are no recursive structures that we can parse; and this stronger claim would appear to be simply false.

Rumelhart and McClelland's discussion of recursion (p. 119-120) nevertheless repays close attention. They are apparently prepared to concede that PDPs can model recursive capacities only indirectly — viz., by implementing Classical architectures like ATNs; so that *if* human cognition exhibited recursive capacities, that would suffice to show that minds have Classical rather than Connectionist architecture at the psychological level. "We have not dwelt on PDP implementations of Turing machines and recursive processing engines *because we do not agree with those who would argue that such capacities are of the essence of human computation*" (p. 119, our emphasis). Their argument that recursive capacities *aren't* "of the essence of human computation" is, however, just the unconvincing stuff about center-embedding quoted above.

So the Rumelhart & McClelland view is apparently that if you take it to be independently obvious that some cognitive capacities are productive, then you should take the existence of such capacities to argue for Classical cognitive architecture and hence for treating Connectionism as at best an implementation theory. We think that this is quite a plausible understanding of the bearing that the issues about productivity and recursion have on the issues about cognitive architecture; in Part 4 we will return to the suggestion that Connectionist models can plausibly be construed as models of the implementation of a Classical architecture.

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.

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'.[24]

---

24. See Pinker, 1984, Chapter 4, for evidence that children never go through a stage in which they distinguish between the internal structures of NPs depending on whether they are in subject or object position; i.e. the dialects that children speak are always systematic with respect to the syntactic structures that can appear in these positions.

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.

Notice that you can make this argument for constituent structure in sentences without idealizing to astronomical computational capacities. There are productivity arguments for constituent structure, but they're concerned with our ability — in principle — to understand sentences that are arbitrarily long. Systematicity, by contrast, appeals to premises that are much nearer home; such considerations as the ones mentioned above, that no speaker understands the form of words 'John loves the girl' except as he also understands the form of words 'the girl loves John'. The assumption that linguistic capacities are productive "in principle" is one that a Connectionist might refuse to grant. But that they are systematic *in fact* no one can plausibly deny.

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 the sentences about John and the girl unless he were able to think both 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 (i.e. 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;[25] 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

25. 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 (i.e. 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.

structure and there is a language of thought. So the architecture of the mind is not a Connectionist network.[26]

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.

But perhaps not being punctate is a property only of the minds of language users; perhaps the representational capacities of infraverbal organisms do have just the kind of gaps that Connectionist models permit? A Connectionist might then claim that he can do everything "up to language" on the assumption that mental representations lack combinatorial syntactic and semantic structure. Everything up to language may not be everything, but it's a lot. (On the other hand, a lot may be a lot, but it isn't everything. Infraverbal cognitive architecture mustn't be so represented as to make the eventual acquisition of language in philogeny and in ontogeny require a miracle.)

It is not, however, plausible that only the minds of verbal organisms are systematic. Think what it would mean for this to be the case. It would have to be quite usual to find, for example, animals capable of representing the state of affairs *a**R**b*, but incapable of representing the state of affairs *b**R**a*. Such animals would be, as it were, *a**R**b* sighted but *b**R**a* blind since, presumably, the representational capacities of its mind affect not just what an organism can think, but also what it can perceive. In consequence, such animals would be able to learn to respond selectively to *a**R**b* situations but quite *un*able to learn to respond selectively to *b**R**a* situations. (So that, though you could teach the creature to choose the picture with the square larger than the triangle, you couldn't for the life of you teach it to choose the picture with the triangle larger than the square.)

It is, to be sure, an empirical question whether the cognitive capacities of infraverbal organisms are often structured that way, but we're prepared to bet that they are not. Ethological cases are the exceptions that prove the rule. There *are* examples where salient environmental configurations act as 'gestalten'; and in such cases it's reasonable to doubt that the mental representation of the stimulus is complex. But the point is precisely that these cases are *exceptional;* they're exactly the ones where you expect that there will be some special story to tell about the ecological significance of the stimulus: that it's the shape of a predator, or the song

---

26. These considerations throw further light on a proposal we discussed in Part II. Suppose that the mental representation corresponding to the thought that John loves the girl is the feature vector {*+John-agent; +loves; +the-girl-patient*} where '*John-agent*' and '*the-girl-patient*' are primitive, atomic features; as such, they bear no more relation to '*John-patient*' and '*the-girl-agent*' than they do to one another or to, say, '*has-a-handle*'. Since this theory recognizes no structural relation between '*John-agent*' and '*John-patient*', it offers no reason why a representational system that provides the means to express one of these concepts should also provide the means to express the other. This treatment of role relations thus makes a mystery of the (presumed) fact that anybody who can entertain the thought that John loves the girl can also entertain the thought that the girl loves John (and, mutatis mutandis, that any natural language that can express the proposition that John loves the girl can also express the proposition that the girl loves John). This consequence of the proposal that role relations be handled by "role specific descriptors that represent the conjunction of an identity and a role" (Hinton, 1987) offers a particularly clear example of how failure to postulate internal structure in representations leads to failuree to capture the systematicity of representational systems.

of a conspecific… etc. Conversely, when there is no such story to tell you expect structurally similar stimuli to elicit correspondingly similar cognitive capacities. That, surely, is the least that a respectable principle of stimulus generalization has got to require.

That infraverbal cognition is pretty generally systematic seems, in short, to be about as secure as any empirical premise in this area can be. And, as we've just seen, it's a premise from which the inadequacy of Connectionist models as cognitive theories follows quite straightforwardly; as straightforwardly, in any event, as it would from the assumption that such capacities are generally productive.

Compositionality of representations

Compositionality is closely related to systematicity; perhaps they're best viewed as aspects of a single phenomenon. We will therefore follow much the same course here as in the preceding discussion: first we introduce the concept by recalling the standard arguments for the compositionality of natural languages. We then suggest that parallel arguments secure the compositionality of mental representations. Since compositionality requires combinatorial syntactic and semantic structure, the compositionality of thought is evidence that the mind is not a Connectionist network.

We said that the systematicity of linguistic competence consists in the fact that "the ability to produce/understand some of the sentences is intrinsically connected to the ability to produce/understand certain of the others". We now add that which sentences are systematically related is not arbitrary from a semantic point of view. For example, being able to understand 'John loves the girl' goes along with being able to understand 'the girl loves John', and there are correspondingly close semantic relations between these sentences: in order for the first to be true, John must bear to the girl the very same relation that the truth of the second requires the girl to bear to John. By contrast, there is no intrinsic connection between understanding either of the John/girl sentences and understanding semantically unrelated formulas like 'quarks are made of gluons' or 'the cat is on the mat' or '2+2=4'; it looks as though semantical relatedness and systematicity keep quite close company.

You might suppose that this covariance is covered by the same explanation that accounts for systematicity per se; roughly, that sentences that are systematically related are composed from the same syntactic constituents. But, in fact, you need a further assumption, which we'll call the 'principle of compositionality': insofar as a language is systematic, a lexical item must make approximately the same semantic contribution to each expression in which it occurs. It is, for example, only insofar as 'the' 'girl', 'loves' and 'John' make the same semantic contribution to 'John loves the girl' that they make to 'the girl loves John' that understanding the one sentence implies understanding the other. Similarity of constituent structure accounts for the semantic relatedness between systematically related sentences only to the extent that the semantical properties of the shared constituents are context-independent.

Here it's idioms that prove the rule: being able to understand 'the', 'man', 'kicked' and 'bucket' isn't much help with understanding 'the man kicked the bucket', since 'kicked' and 'bucket' don't bear their standard meanings in this context. And, just as you'd expect, 'the man kicked the bucket' is *not* systematic even with respect to syntactically closely related sentences like 'the man kicked over the bucket' (for that matter, it's not systematic with respect to the 'the man kicked the bucket' read literally.)

It's uncertain exactly how compositional natural languages actually are (just as it's uncertain exactly how systematic they are). We suspect that the amount of context induced variation of lexical meaning is often overestimated because other sorts of context sensitivity are misconstrued as violations of compositionality. For example, the difference between 'feed the chicken' and 'chicken to eat' must involve an *animal/food* ambiguity in 'chicken' rather than a violation of compositionality since if the context 'feed the…' could *induce* (rather than select) the meaning *animal*, you would expect 'feed the veal', 'feed the pork' and the like.[27] Similarly, the difference between 'good book', 'good rest' and 'good fight' is probably not meaning shift but syncategorematicity. 'Good *NP*' means something like *NP that answers to the relevant interest in NPs*: a good book is one that answers to our interest in books (viz. it's good to read); a good rest is one that answers to our interest in rests (viz. it leaves one refreshed); a good fight is one that answers to our interest in fights (viz. it's fun to watch or to be in, or it clears the air); and so on. It's because the meaning of 'good' is syncategorematic and has a variable in it for relevant interests, that you can know that a good flurg is a flurg that answers to the relevant interest in flurgs without knowing what flurgs are or what the relevant interest in flurgs is. (See Ziff, 1960).

In any event, the main argument stands: systematicity depends on compositionality, so to the extent that a natural language is systematic it must be compositional too. This illustrates another respect in which systematicity arguments can do the work for which productivity arguments have previously been employed. The standard argument for compositionality is that it is required to explain how a finitely representable language can contain infinitely many nonsynonymous expressions.

Considerations about systematicity offer one argument for compositionality; considerations about entailment offer another. Consider predicates like '…is a brown cow'. This expression bears a straightforward semantical relation to the predicates '…is a cow' and '…is brown'; viz that the first predicate is true of a thing if and only if both of the others are. I.e. '… is a brown cow' severally entails '…is brown' and '…is a cow' and is entailed by their conjunction. Moreover — and this is important — this semantical pattern is not peculiar to the cases cited. On the contrary, it holds for a very large range of predicates (see '…is a red square,' '…is a funny old German soldier,' '…is a child prodigy;' and so forth.)

How are we to account for these sorts of regularities? The answer seems clear enough; '… is a brown cow' entails '…is brown' because (a) the second expression is a constituent of the first; (b) the syntactical form '(adjective noun)$_N$' has (in many cases) the semantic force of a conjunction, and (c) 'brown' retains its semantical value under simplification of conjunction. Notice that you need (c) to rule out the possibility that 'brown' means *brown* when in it modifies

27. We are indebted to Steve Pinker for this point.

a noun but (as it might be) *dead* when it's a predicate adjective; in which case '… is a brown cow' wouldn't entail '…is brown' after all. Notice too that (c) is just an application of the principle of composition.

So, here's the argument so far: you need to assume some degree of compositionality of English sentences to account for the fact that systematically related sentences are always semantically related; and to account for certain regular parallelisms between the syntactical structure of sentences and their entailments. So, beyond any serious doubt, the sentences of English must be compositional to some serious extent. But the principle of compositionality governs the semantic relations between words *and the expressions of which they are constituents.* So compositionality implies that (some) expressions *have* constituents. So compositionality argues for (specifically, presupposes) syntactic/semantic structure in sentences.

Now what about the compositionality of mental representations? There is, as you'd expect, a bridging argument based on the usual psycholinguistic premise that one uses language to express ones thoughts: Sentences are used to express thoughts; so if the ability to use some sentences is connected with the ability to use certain other, semantically related sentences, then the ability to think some thoughts must be correspondingly connected with the ability to think certain other, semantically related thoughts. But you can only think the thoughts that your mental representations can express. So, if the ability to think certain thoughts is interconnected, then the corresponding representational capacities must be interconnected too; specifically, the ability to be in some representational states must imply the ability to be in certain other, semantically related representational states.

But then the question arises: *how could* the mind be so arranged that the ability to be in one representational state is connected with the ability to be in others that are semantically nearby? What account of mental representation would have this consequence? The answer is just what you'd expect from the discussion of the linguistic material. Mental representations must have internal structure, just the way that sentences do. In particular, it must be that the mental representation that corresponds to the thought that John loves the girl contains, as its parts, the same constituents as the mental representation that corresponds to the thought that the girl loves John. That would explain why these thoughts are *systematically* related; *and, to the extent that the semantic value of these parts is context-independent, that would explain why these systematically related thoughts are also semantically related*. So, by this chain of argument, evidence for the compositionality of sentences is evidence for the compositionality of the representational states of speaker/hearers.

Finally, what about the compositionality of infraverbal thought? The argument isn't much different from the one that we've just run through. We assume that animal thought is largely systematic: the organism that can perceive (hence learn) that *a**R**b* can generally perceive (/learn) that *b**R**a*. But, systematically related thoughts (just like systematically related sentences) are generally semantically related too. It's no surprise that being able to learn that the triangle is above the square implies being able to learn that the square is above the triangle; whereas it would be *very* surprising if being able to learn the square/triangle facts implied being able to learn that quarks are made of gluons or that Washington was the first President of America.

So, then, what explains the correlation between systematic relations and semantic relations in infraverbal thought? Clearly, Connectionist models don't address this question; the fact that a network contains a node labelled X has, so far as the constraints imposed by Connectionist architecture are concerned, *no implications at all* for the labels of the other nodes in the network; in particular, it doesn't imply that there will be nodes that represent thoughts that are semantically close to X. This is just the semantical side of the fact that network architectures permit arbitrarily punctate mental lives.

But if, on the other hand, we make the usual Classicist assumptions (viz. that systematically related thoughts share constituents and that the semantic values of these shared constituents are context independent) the correlation between systematicity and semantic relatedness follows immediately. For a Classicist, this correlation is an 'architectural' property of minds; it couldn't but hold if mental representations have the general properties that Classical models suppose them to.

What have Connectionists to say about these matters? There is some textual evidence that they are tempted to deny the facts of compositionality wholesale. For example, Smolensky (1988) claims that: "Surely… we would get quite a different representation of 'coffee' if we examined the difference between 'can with coffee' and 'can without coffee' or 'tree with coffee' and 'tree without coffee'; or 'man with coffee' and 'man without coffee'… context insensitivity is not something we expect to be reflected in Connectionist representations….".

It's certainly true that compositionality is not generally a feature of Connectionist representations. Connectionists can't acknowledge the facts of compositionality because they are committed to mental representations that don't have combinatorial structure. But to give up on compositionality is to take 'kick the bucket' as a model for the relation between syntax and semantics; and the consequence is, as we've seen, that you make the systematicity of language (and of thought) a mystery. On the other hand, to say that 'kick the bucket' is aberrant, and that the right model for the syntax/semantics relation is (e.g.) 'brown cow', is to start down a trail which leads, pretty inevitably, to acknowledging combinatorial structure in mental representation, hence to the rejection of Connectionist networks as cognitive models.

We don't think there's any way out of the need to acknowledge the compositionality of natural languages and of mental representations. However, it's been suggested (see Smolensky, op cit) that while the principle of compositionality is false (because content isn't context invariant) there is nevertheless a "family resemblance" between the various meanings that a symbol has in the various contexts in which it occurs. Since such proposals generally aren't elaborated, it's unclear how they're supposed to handle the salient facts about systematicity and inference. But surely there are going to be serious problems. Consider, for example, such inferences as

(i) Turtles are slower than rabbits.
(ii) Rabbits are slower than Ferraris.
……
(iii) Turtles are slower than Ferraris.

The soundness of this inference appears to depend upon (a) the fact that the same relation (viz, *slower than*) holds between turtles and rabbits on the one hand, and rabbits and Ferraris on the other; and (b) the fact that relation is transitive. If, however, it's assumed (contrary to the principle of compositionality) that 'slower than' means something different in premises (i) and (ii) (and presumably in iii as well) — so that, strictly speaking, the relation that holds between turtles and rabbits is *not* the same one that holds between rabbits and Ferraris — then it's hard to see why the inference should be valid.

Talk about the relations being 'similar' only papers over the difficulty since the problem is then to provide a notion of similarity that will guaranty that if (i) and (ii) are true, so too is (iii). And, so far at least, no such notion of similarity has been forthcoming. Notice that it won't do to require just that the relations all be similar in respect of their *transitivity*, i.e. that they all be transitive. On that account, the argument from 'turtles are slower than rabbits' and 'rabbits are furrier than Ferraris' to 'turtles are slower than Ferraris' would be valid since 'furrier than' is transitive too.

Until these sorts of issues are attended to, the proposal to replace the compositional principle of context invariance with a notion of "approximate equivalence … across contexts" (Smolensky,1988) doesn't seem to be much more than hand waiving.

The systematicity of inference

In Part 2 we saw that, according to Classical theories, the syntax of mental representations mediates between their semantic properties and their causal role 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,[28] 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 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 (viz $S_1\&S_2\&S_3\&\ldots S_n$); and the process of drawing the

28. 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 (AvBv CvDvE) and (-B&-C&-D&-E) to A is of the same logical type as the inference from AvB and –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.

inference corresponds, in both cases, to the same formal operation of detaching the constituent that expresses the conclusion.

The idea that organisms should exhibit similar cognitive capacities in respect of logically similar inferences is so natural that it may seem unavoidable. But, on the contrary: there's nothing in principle to preclude a kind of cognitive model in which inferences that are quite similar from the logician's point of view are nevertheless computed by quite different mechanisms; or in which some inferences of a given logical type are computed and other inferences of the same logical type are not. Consider, in particular, the Connectionist account. 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 3) would do:

==========

Insert Figure 3 about here

==========

But notice that *a Connectionist can equally 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. Analogously, there's no reason why you shouldn't get minds that simplify the premise *John loves Mary and Bill hates Mary* but no others; or minds that simplify premises with 1, 3, or 5 conjuncts, but don't simplify premises with 2, 4, or 6 conjuncts; or, for that matter, minds that simplify only premises that were acquired on Tuesdays… etc.

In fact, the Connectionist architecture is *utterly indifferent* as among these possibilities. That's because it recognizes no notion of syntax according to which thoughts that are alike in inferential role (e.g. thoughts that are all subject to simplification of conjunction) are expressed by mental representations of correspondingly similar syntactic form (e.g. by mental representations that are all syntactically conjunctive). 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.

Summary

It is perhaps obvious by now that all the arguments that we've been reviewing — the argument from systematicity, the argument from compositionality, and the argument from semantic coherence — are really much the same: If you hold the kind of theory that acknowledges structured representations, it must perforce acknowledge representations with *similar* or *identical* structures. In the linguistic cases, constituent analysis implies a taxonomy of sentences by their syntactic form, and in the inferential cases, it implies a taxonomy of arguments by their logical form. So, if your theory also acknowledges mental processes that are structure sensitive, then it will predict that similarly structured representations will generally play similar roles in thought. A theory that says that the sentence 'John loves the girl' is made out of the same parts as the sentence 'the girl loves John', and made by applications of the same rules of composition, will have to go out of its way to explain linguistic competence which embrace one sentence but not the other. And similarly, if a theory says that the mental representation that corresponds to the thought that *P&Q&R* has the same (conjunctive) syntax as the mental representation that corresponds to the thought that *P&Q*, and that mental processes of drawing inferences subsume mental representations in virtue of their syntax, it will have to go out of its way to explain inferential capacities which embrace the one thought but not the other. Such a competence would be, at best, an embarrassment for the theory, and at worst a refutation.

By contrast, since the Connectionist architecture recognizes no combinatorial structure in mental representations, gaps in cognitive competence should proliferate arbitrarily. It's not just that you'd expect to get them from time to time; it's that, on the 'no-structure' story, *gaps are the unmarked case*. It's the *systematic* competence that the theory is required to treat as an embarrassment. But, as a matter of fact, inferential competences are *blatantly* systematic. So there must be something deeply wrong with Connectionist architecture.

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.[29]

---

29. 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 representations 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

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… etc. There is therefore nothing to stop a Connectionist from stipulating — as an independent postulate of his theory of mind — that all biologically instantiated networks are, de facto, systematic.

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.

Part 4: The lure of connectionism

The widespread 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 in relation to neural speeds: the "hundred step" constraint.*** It has been observed (e.g., Feldman & Ballard, 1982) that the time required to execute computer instructions is in the order of nanoseconds, whereas 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

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.

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 & 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 (e.g., 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 & 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.

- ***Conventional computer models are committed to a different etiology for "rule governed" behavior and "exceptional" behavior.*** Classical psychological theories, which are based on conventional computer ideas, typically distinguish between mechanisms that cause regular and divergent behavior by postulating systems of explicit unconscious rules to explain the former, and then attributing departures from these rules to secondary (performance) factors. Since the divergent behaviors occur very frequently, a better strategy would be to try to account for both types of behavior in terms of the same mechanism.

- ***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 & Hinton, 1987; Smolensky, 1988). We appear to know many things that we would have great difficulty in describing verbally, including how to ride a bicycle, what our close friends look like, and how to recall the name of the President, etc. 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 suggests 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.

- ***Storage in conventional architectures is passive.*** Conventional computers have a passive memory store which is accessed in what has been called a "fetch and execute cycle." This appears to be quite unlike human memory. For example, according to Kosslyn and Hatfield (1984),

> "In computers the memory is static: once an entry is put in a given location, it just sits there until it is operated upon by the CPU…. But consider a very simple experiment: Imagine a letter *A* over and over again… then switch to the letter *B*. In a model employing a Von Neumann architecture the 'fatigue' that inhibited imaging the *A* would be due to some quirk in the way the CPU executes a given instruction…. Such fatigue should generalize to all objects imaged because the routine responsible for imaging was less effective. But experiments have demonstrated that this is not true: specific objects become more difficult to image, not all objects. This finding is more easily explained by an analogy to the way invisible ink fades of its own accord…: with invisible ink, the representation itself is doing something — there is no separate processor working over it… (p1022 & p1029)".

- ***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 & 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 know 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, etc., 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.

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 (e.g. 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, resistance to damage and noise, and the passivity of memory.)

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'.

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.

If you ever have doubts about whether a proposal pertains to the implementation level or the symbolic level, a useful heuristic is to ask yourself whether what is being claimed is true of a conventional computer — such as the DEC VAX — at *its* implementation level. Thus although most algorithms that run on the VAX are serial,[30] at the implementation level such computers are 'massively parallel'; they quite literally involve simultaneous electrical activity throughout almost the entire device. For example, every memory access cycle involves pulsing every bit in a significant fraction of the system's memory registers — since memory access is essentially a destructive read and rewrite process, the system clock regularly pulses and activates most of the central processing unit, and so on.

The moral is that 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[31] of neurons is invariably the relevant implementation property (at least for higher level cognitive processes like learning and memory) the 100 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

---

30. Even in the case of a conventional computer, whether it should be viewed as executing a serial or a parallel algorithm depends on what 'virtual machine' is being considered in the case in question. After all, a VAX *can* be used to simulate (i.e., to implement) a virtual machine with a parallel architecture. In that case the relevant algorithm would be a parallel one.

31. There are, in fact, a number of different mechanisms of neural interaction (e.g., the "local interactions" described by Rakic, 1975). Moreover, a large number of chemical processes take place at the dendrites, covering a wide range of time scales, so even if dendritic transmission were the only relevant mechanism, we still wouldn't know what time scale to use as our estimate of neural action in general (see, for example, Black, 1986).

are going on in parallel in cognition, and that these processes interact with one another (e.g., 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' (1985) "Connection Machine", as well as any of a number of 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.

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 PDP's 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.[32] 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. Come to think of it, we already had it in the old style "ferrite core" memories!

The physical requirements of a Classical symbol-processing system are easily misunderstood. (Confounding of physical and functional properties is widespread in psychological theorizing in general; for a discussion of this confusion in relation to metrical properties in models of mental imagery, see Pylyshyn 1981). 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 (see the discussion in footnote 10). However, since the relations to be physically realized is *functional* adjacency, there is no necessity that physical instantiations of adjacent

---

32. Unless the 'units' in a Connectionist network really are assumed to have different spatially-focused loci in the brain, talk about distributed representation is likely to be extremely misleading. In particular, if units are merely *functionally* individuated, any amount of distribution of functional entities is compatible with any amount of spatial compactness of their neural representations. But it is not clear that units do, in fact, correspond to any anatomically identifiable locations in the brain. In the light of the way Connectionist mechanisms are designed, it may be appropriate to view units and links as functional/mathematical entities (what psychologists would call "hypothetical constructs") whose neurological interpretation remains entirely open. (This is, in fact, the view that some Connectionists take; see Smolensky, 1988). The point is that distribution over mathematical constructs does not buy you damage resistance; only *neural* distribution does!

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. In a VAX (to use our heuristic again) pairs of symbols may certainly be functionally adjacent, but the symbol tokens are nonetheless spatially spread through many locations in physical memory.

In short, the fact that a property (like the position of a symbol within an expression) is functionally local has no implications one way or the other for damage-resistance or noise tolerance unless the functional-neighborhood metric corresponds to some appropriate *physical* dimension. When that is the case, we may be able to predict adverse consequences that varying the physical property has on objects localized in functional space (e.g., varying the voltage or line frequency might damage the left part of an expression). But, of course, the situation is exactly the same for Connectionist systems: even when they are resistant to spatially-local damage, they may not be resistant to damage that is local along some other physical dimensions. Since spatially-local damage is particularly frequent in real world traumas, this may have important practical consequences. But so long as our knowledge of how cognitive processes might be mapped onto brain tissue remains very nearly nonexistent, its message for cognitive science remains moot.

"Soft" constraints, continuous magnitudes, stochastic mechanisms, and active symbols

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, 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, etc., 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: There is, to our knowledge, no reason to believe that something like Newell's (1969) "hierarchy of weak methods" or Laird, Rosenberg and Newell's (1986) "universal subgoaling," is in principle incapable of dealing with the problem of graceful degradation. (Nor, to our knowledge, has any argument yet been offered that Connectionist architectures are in principle capable of dealing with it. In fact current Connectionist models are every bit as graceless in their modes of failure as ones based on Classical architectures. For example, contrary to some claims, models such as that of McClelland and Kawamoto (1986) fail quite unnaturally when given incomplete information.)

In short, the Classical theorist can view stochastic properties of behavior as emerging from interactions between the model and the intrinsic properties of the physical medium in which it is realized. It is essential to remember that, from the Classical point of view, overt behavior is par excellence an interaction effect, and symbol manipulations are supposed to be only one of the interacting causes.

These same considerations apply to Kosslyn and Hatfield's remarks (quoted earlier) about the commitment of Classical models to 'passive' versus 'active' representations. It is true, as Kosslyn and Hatfield say, that the representations that Von Neumann machines manipulate 'don't *do* anything' until a CPU operates upon them (they don't decay, for example). But, even on the absurd assumption that the mind has *exactly* the architecture of some contemporary (Von Neumann) computer, it is obvious that its behavior, and hence the behavior of an organism, is determined not just by the logical machine that the mind instantiates, but also by the protoplasmic machine in which the logic is realized. Instantiated representations *are* therefore bound to be active, even according to Classical models; the question is whether the kind of activity they exhibit should be accounted for by the cognitive model or by the theory of its implementation. This question is empirical and must not be begged on behalf of the Connectionist view. (As it is, for example, in such passages as "The brain itself does not manipulate symbols; the brain is the medium in which the symbols are floating and in which they trigger each other. There is no central manipulator, no central program. There is simply a vast collection of 'teams'- patterns of neural firings that, like teams of ants, trigger other patterns of neural firings… We feel those symbols churning within ourselves in somewhat the same way we feel our stomach churning." (Hofstadter, 1983, p. 279). This appears to be a serious case of *Formicidae in machina*: ants in the stomach of the ghost in the machine.)

Explicitness of rules

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

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.[33]

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 (i.e. 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

---

33. An especially flagrant example of how issues about architecture get confused with issues about the explicitness of rules in the Connectionist literature occurs in PDP Chapter 4., where Rumelhart and McClelland argue that PDP models provide "… a rather plausible account of how we can come to have innate 'knowledge'. To the extent that stored knowledge is assumed to be in the form of explicit, inaccessible rules … it is hard to see how it could 'get into the head' of the newborn. It seems to us implausible that the newborn possesses elaborate symbol systems and the systems for interpreting them required to put these explicit, inaccessible rules to use in guiding behavior. On our account, we do not need to attribute such complex machinery. If the innate knowledge is simply the prewired connections, it is encoded from the start in just the right way to be of use by the processing mechanisms.(142)" A priorizing about what it does and doesn't seem likely that newborns possess strikes us as a bad way to do developmental cognitive psychology. But Rumelhart and McClelland's argument is doubly beside the point since a Classicist who shares their prejudices can perfectly well avail himself of the same solution that they endorse. Classical architecture does *not* require "complex machinery" for "interpreting" explicit rules since classical machines do not *require* explicit rules at all. Classical architecture is therefore *neutral* on the Empiricism/Nativism issue (and so is Connectionism, as Rumelhart and Mcclelland elsewhere correctly remark.)

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.

One of the important inventions in the history of computers —the stored-program computer — makes it *possible* for programs to take on the role of data structures. But nothing in the architecture *requires* that they always do so. Similarly, Turing demonstrated that there exists an abstract machine (the so-called Universal Turing Machine) which can simulate the behavior of any target (Turing) machine. A Universal machine is "rule-explicit" about the machine it is simulating (in the sense that it has an explicit representation of that machine which is sufficient to specify its behavior uniquely). Yet the target machine can perfectly well be "rule-implicit" with respect to the rules that govern *its* behavior.

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.[34]

If, therefore, there should prove to be persuasive arguments for rule explicit cognitive processes, that would be very embarrassing for Connectionists. A natural place to look for such arguments would be in the theory of the acquisition of cognitive competences. For example, much traditional work in linguistics (see Prince and Pinker, 1988) and all recent work in mathematical learning theory (see Osherson, Stov & Weinstein, 1984), assumes that the characteristic output of a cognitive acquisition device is a recursive rule system (a grammar, in the linguistic case). Suppose such theories prove to be well-founded; then that would be incompatible with the assumption that the cognitive architecture of the capacities acquired is Connectionist.

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 the Ballard (1986) paper). On the other hand, Smolensky (1988) views Connectionist units as mathematical objects which can be given an interpretation in either neural

34. Of course, it *is* possible to simulate a "rule explicit process" in a Connectionist network by first implementing a Classical architecture in the network. The slippage between networks as architectures and as implementations is ubiquitous in Connectionist writings, as we remarked above.

or psychological terms. Most Connectionists find themselves somewhere in between, frequently referring to their approach as "brain style" theorizing.[35]

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 multidimension "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.

Consider, for example, one of the most salient properties of neural systems: they are networks which transmit activation culminating in state changes of some quasi-threshold elements. Surely it is not warranted to conclude that reasoning consists of the spread of excitation among representations, or even among semantic components of representations. After all, a VAX is also correctly characterized as consisting of a network over which excitation is transmitted culminating in state changes of quasi-threshold elements. Yet at the level at which it processes representations, a VAX is *literally* organized as a Von Neuman architecture.

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.

---

35. 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 & McClelland, 1986a, V1, 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.

There are, no doubt, cases where special empirical considerations suggest detailed structure/function correspondences or other analogies between different levels of a system's organization. For example, the input to the most peripheral stages of vision and motor control *must* be specified in terms of anatomically projected patterns (of light, in one case, and of muscular activity in the other); and independence of structure and function is perhaps less likely in a system whose input or output must be specified somatotopically. Thus, at these stages it is reasonable to expect an anatomically distributed structure to be reflected by a distributed functional architecture. When, however, the cognitive process under investigation is as abstract as reasoning, there is simply no reason to expect isomorphisms between structure and function; as, indeed, the computer case proves.

Perhaps this is all too obvious to be worth saying. Yet it seems that the commitment to "brain style" modeling leads to many of the characteristic Connectionist claims about psychology, and that it does so via the implicit — and unwarranted — assumption that there ought to be similarity of structure among the different levels of organization of a computational system. This is distressing since much of the psychology that this search for structural analogies has produced is strikingly recidivist. Thus the idea that the brain is a neural network motivates the revival of a largely discredited Associationist psychology. Similarly, the idea that brain activity is anatomically distributed leads to functionally distributed representations for concepts which in turn leads to the postulation of micro-features; yet the inadequacies of feature-based theories of concepts are well-known and, to our knowledge, micro-feature theory has done nothing to address them.(See Bolinger, 1965; J.D. Fodor, 1977). Or again, the idea that the strength of a connection between neurons is affected is by the frequency of their co-activation gets projected onto the cognitive level. The consequence is a resurgence of statistical models of learning that had been widely acknowledged (both in Psychology and in AI) to be extremely limited in their applicability (e.g., Minsky & Papert, 1972, Chomsky, 1957).

So although, *in principle*, knowledge of how the brain works could direct cognitive modeling in a beneficial manner, *in fact* a research strategy has to be judged by its fruits. The main fruit of "brain style modeling" has been to revive psychological theories whose limitations had previously been pretty widely appreciated. It has done so largely because assumptions about the structure of the brain have been adopted in an all-too-direct manner as hypotheses about cognitive architecture; it's an instructive paradox that the current attempt to be thoroughly modern and 'take the brain seriously' should lead to a psychology not readily distinguishable from the worst of Hume and Berkeley. 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.

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.[36] 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 PDP's 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 pushdown stacks and other LISP facilities.[37] Rumelhart & McClelland (1986a, p 117), who are convinced that

---

36. Rumelhardt & McClelland maintain that PDP models are more than *just* theories of implementation because (1) they add to our understanding of the problem (p 116), (2) studying PDPs can lead to the postulation of different macrolevel processes (p 126). Both these points deal with the heuristic value of "brain style" theorizing. Hence, though correct in principle, they are irrelevant to the crucial question whether Connectionism is best understood as an attempt to model neural implementation, or whether it really does promise a "*new* theory of the mind" incompatible with Classical information-processing approaches. It is an empirical question whether the heuristic value of this approach will turn out to be positive or negative. We have already commented on our view of the recent history of this attempt.

37. Even in this case, where the model is specifically designed to implement Lisp-like features, some of the rhetoric fails to keep the implementation-algorithm levels distinct. This leads to talk about "emergent properties" and to the claim that even when they implement Lisp-like mechanisms, Connectionist systems "can compute things in ways in which Turing machines and von Neumann computers can't" (Touretzky, 1986). Such a claim suggests that Touretzky distinguishes different "ways of computing" not in terms of different algorithms, but in terms of different ways of implementing the same algorithm. While nobody has proprietary rights to terms like "ways of computing", this is a misleading way of putting it; it means that a DEC machine has a "different way of computing" from an IBM machine even when executing the identical program.

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 & 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 Connnectionist 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*.

Part 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 29), 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: viz 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, 1986a) "… a reasonable basis for modeling cognitive processes in general" (p. 110). It could still be held that networks sustain *some* cognitive processes. A good bet might be that they sustain such processes as can be analyzed as the drawing of statistical inferences; 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.

This is, for example, one way of understanding what's going on in the argument between Rumelhart and McClelland (1986b) and Prince and Pinker (1988), though neither paper puts it in quite these terms. In effect, Rumelhart and McClelland postulate a mechanism which, given a corpus of pairings that a 'teacher' provides as data, computes the statistical correlation between the phonological form of the ending of a verb and the phonological form of its past tense inflection. (The magnitude of the correlations so computed is analogically represented by the weights that the network exhibits at asymptote.) Given the problem of inflecting a new verb stem ending in a specified phonological sequence, the machine chooses the form of the past tense that was most highly correlated with that sequence in the training set. By contrast, Prince and Pinker argue (in effect) that more must be going in learning past tense morphology than merely estimating correlations since the statistical hypothesis provides neither a close fit to the ontogenetic data nor a plausible account of the adult competence on which the ontogenetic processes converge. It seems to us that Pinker and Prince have, by quite a lot, the best of this argument.

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.

References

Arbib, M. (1975). Artificial intelligence and brain theory: Unities and diversities. *Biomedical Engineering,3* 238-274.

Ballard, D.H. (1986). Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences, 9*, 67-120.

Ballard, D.H. (1987). Parallel Logical Inference and Energy Minimization. Report TR142, Computer Science Department, University of Rochester.

Black, I.B. (1986) Molecular memory mechanisms. In G. Lynch, *Synapses, Circuits, and the Beginnings of Memory*, Cambridge, Mass., M.I.T. Press, A Bradford Book.

Bolinger, D. (1965). The atomization of meaning, *Language, 41*, 555-573.

Broadbent, D. (1985). A question of levels: Comments on McClelland and Rumelhart. *Journal of Experimental Psychology: General, 114*, 189-192.

Carroll, L. (1956). What the tortoise said to achilles and other riddles. In Newman, J.R. (ed.) *The World of Mathematics: Volume Four*. New York, N.Y., Simon and Schuster.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, Cambridge, Mass., M.I.T. Press.

Chomsky, N. (1968). *Language and Mind*. New York: Harcourt, Brace and World.

Churchland, P.M. (1981). Eliminative materialism and the propositional attitudes, *Journal of Philosophy, 78*, 67-90.

Churchland, P.S. (1986). *Neurophilosophy*, Cambridge, Mass. M.I.T. Press.

Cummins, R. (1983). *The Nature of Psychological Explanation*, Cambridge, Mass., M.I.T. Press.

Dennett, D. (1986). The logical goegraphy of computational approaches: A view from the east pole. In Brand,M. & Harnish, M. (eds.) *The Representation of Knowledge*, Tuscon, AZ, The University of Arizona Press.

Dreyfus, H. & Dreyfus, S. (in press) Making a mind vs modelling the brain: A.I. back at a branch point. *Daedalus*.

Fahlman, S.E. & Hinton, G.E. (1987). Connectionist architectures for artificial intelligence. *Computer,20*, 100-109.

Feldman, J.A. (1986). Neural representation of conceptual knowledge. Report TR189. Department of Computer Science, University of Rochester.

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

Fodor, J. (1976). *The Language of Thought*, Harvester Press, Sussex. (Harvard University Press paperback).

Fodor, J. (1987). *Psychosemantics*, M.I.T. Press, Cambridge, Ma.

Fodor, J.D. (1977) *Semantics: Theories of Meaning in Generative Grammar*, New York: Thomas Y. Crowell.

Frohn, H., Geiger, H. & Singer, W. (1987). A self-organizing neural network sharing features of the mammalian visual system. *Biological Cybernetics, 55*, 333-343.

Geach, P. (1957) *Mental Acts*, London, Routledge and Kegan Paul.

Hewitt, C. (1977). Viewing control structures as patterns of passing messages. *The Artificial Intelligence Journal, 8*, 232-364.

Hillis, D. (1985). *The Connection Machine*, Cambridge, Mass., M.I.T. Press.

Hinton, G. (1987) . Representing part-whole Hierarchies in connectionist networks. *Proceedings AAAI-87*. Palo Alto: Kaufman.

Hofstadter, D.R. (1983) Artificial intelligence: Sub-cognition as computation. In F. Machlup & U. Mansfield, *The Study of Information: Interdisciplinary Messages*, New York, N.Y., John Wiley & Sons.

Katz, J.J. (1972). *Semantic Theory*, New York: Harper & Row.

Katz, J.J. & Fodor, J.A. (1963). The structure of a semantic theory, *Language, 39*, 170-210.

Katz, J. & Postal, P. (1964). *An Integrated Theory of Linguistic Descriptions*, Cambridge, Mass. M.I.T.Press.

Kosslyn, S.M. & Hatfield, G. (1984). Representation without symbol systems. *Social Research, 51*, 1019-1054.

Laird, J., Rosenbloom, P. & Newell, A. *Universal Subgoaling and Chunking: The Automatic Generation and Learning of Goal Hierarchies*, Boston, Mass: Kluwer Academic Publishers.

Lakoff, G. (1986). Connectionism and congitive linguistics. Seminar delivered at Princeton University, December 8, 1986.

Mackworth, A. (1987). Constraint propagation. In Shapiro, S.C. (ed.), *The Encyclopedia of Artificial Intelligence, Volume 1*, New York, N.Y., John Wiley & Sons.

McClelland, J.L., Feldman, J., Adelson, B., Bower, G. & McDermott, D. (1986). Connectionist models and cognitive science: Goals, directions and implications. *Report to the National Science Foundation, June 1986*.

McClelland, J.L. & Kawamoto, A.H. (1986) Mechanisms of sentence processing: Assigning roles to constituents. In McClelland, Rumelhart and the PDP Research Group, (eds.) *Parallel Distributed Processing: Volume 2*, Cambridge, Mass. M.I.T. Press, A Bradford Book.

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986) The appeal of parallel distributed processing. In Rumelhart, McClelland and the PDP Research Group, (eds.) *Parallel Distributed Processing: Volume 1*, Cambridge, M.I.T. Press, A Bradford Book.

Minsky, M. & Papert, F. (1972). *Artificial Intelligence Progress Report*, AI Memo 252, Massichusetts Institute of Technology.

Newell, A. (1969). Heuristic programming: Ill-structured problems. In Aronofsky, J. (ed.) *Progress in Operations Research, III*, New York, John Wiley & Sons.

Newell, A.(1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87-127.

Osherson, D., Stov, M. & Weinstein, S. (1984). Learning theory and natural language, *Cognition, 17*, 1-28.

Pinker, S. (1984). *Language, Learnability and Language Development*. Cambridge: Harvard University Press.

Prince, A. & Pinker, S. (1987). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, In press.

Pylyshyn, Z.W. (1984a). *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, M.I.T. Press, A Bradford Book.

Pylyshyn, Z.W. (1984b). Why computation requires symbols. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society, Bolder, Colorado, August, 1984*. Hillsdale, N.J: Earlbaum.

Pylyshyn, Z.W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review, 88*, 16-45.

Pylyshyn, Z.W. (1980). Cognition and computation: Issues in the foundations of cognitive science, *Behavioral and Brain Sciences, 3:1*, 154-169.

Rakic, P. (1975). Local circuit neurons. *Neurosciences Research Program Bulletin, 13*, 299-313.

Rumelhart, D.E. (1984). The emergence of cognitive phenomena from sub-symbolic proesses. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society, Bolder, Colorado, August, 1984*. Hillsdale, N.J: Earlbaum.

Rumelhart, D.E. & McClelland, J.L. (1985). Level's indeed! A response to Broadbent. *Journal of Experimental Psychology: General, 114*, 193-197

Rumelhart, D.E. & McClelland, J.L. (1986a). PDP Models and general issues in cognitive science. In Rumelhart, McClelland and the PDP Research Group (eds.) *Parallel Distributed Processing, Volume 1*, Cambridge, M.I.T. Press, A Bradford Book.

Rumelhart, D.E. & McClelland, J.L. (1986b). On learning the past tenses of english verbs. In Rumelhart, McClelland and the PDP Research Group (eds.) *Parallel Distributed Processing, Volume 1*, Cambridge, M.I.T. Press, A Bradford Book.

Schneider, W. (1987). Connectionism: Is it a paradigm shift for psychology? *Behavior Research Methods, Instruments, & Computers, 19*, 73-83.

Sejnowski, T.J. (1981). Skeleton filters in the brain. In G. E. Hinton and J.A. Anderson (eds.) *Parallel Models of Associative Memory*. Hillsdale, N.J.: Earlbaum.

Simon, H.A. & Chase, W.G. (1973). Skill in chess. *American Scientist, 621*, 394-403.

Smolensky, P. (1988). On the proper treatment of connectionism. *The Behavioral and Brain Sciences, 11*, (in press).

Stabler, E. (1983). How are grammers represented? *Behavioral and Brain Sciences, 6*, 391-420

Stich, S. (1983). *From Folk Psychology to Cognitive Science*, M.I.T. Press, Cambridge, MA.

Touretzky, D.S. (1986). BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees, *Procedings of the Eighth Annual Conference of the Cognitive Science Society*, Amherst, Mass. August, 1986], Hillsdale, N.J: Earlbaum.

Wanner, E. & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan & G. A. Miller (eds), *Linguistic Theory and Psychological Reality*. Cambridge, Ma: MIT Press.

Watson, J. (1930). *Behaviorism*, U. of Chicago Press, Chicago.

Ziff, P (1960). *Semantic Analysis*. Ithica: Cornell University Press.

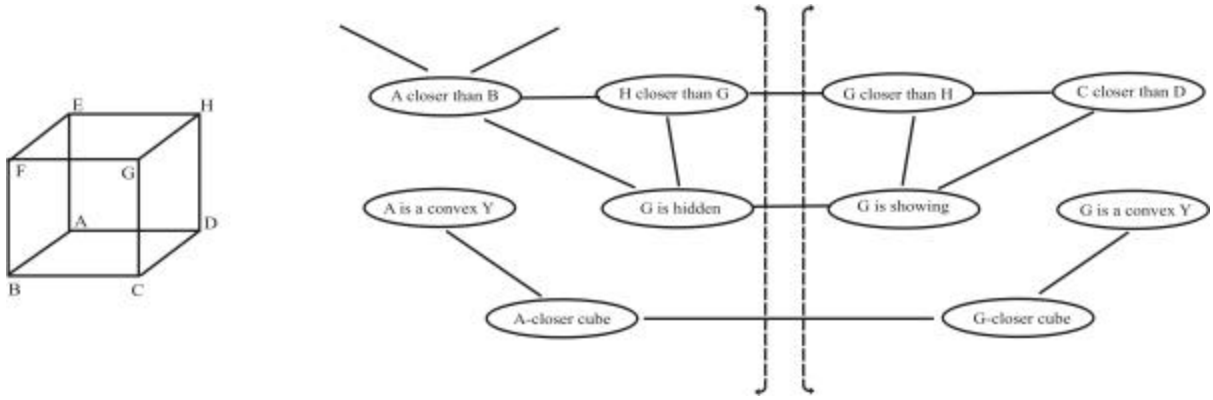

Figure 1

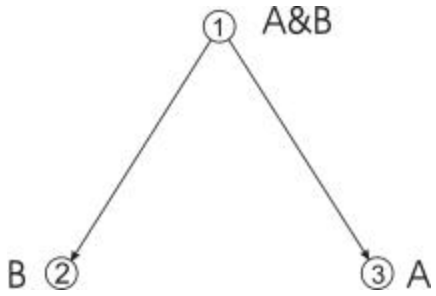

Figure 2