Cognition, 28 (1988) 3–71

Connectionism and cognitive architecture: A critical analysis*

JERRY A. FODOR
CUNY Graduate Center

ZENON W. PYLYSHYN
University of Western Ontario

Abstract

This paper explores differences 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.

*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, and Edward Stabler. Reprints may be obtained by writing to either author: Jerry Fodor, CUNY Graduate Center, 33 West 42 Street, New York, NY 10036, U.S.A.; Zenon Pylyshyn, Centre for Cognitive Science, University of Western Ontario, London, Ontario, Canada N6A 5C2.

1. Introduction

Connectionist or *PDP* models are catching on. There are conferences and new books nearly every day, and the popular science press hails this new wave of theorizing as a breakthrough in understanding the mind (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 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 pseudoscientific 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 (e.g., Dreyfus & 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 & 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, 1986), and so on and on. It also appeals to many young cognitive scientists who view the approach as not only anti-establishment (and therefore desirable) but also rigorous and mathematical (see, however, footnote 2). 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 Fodor 1976, 1987; Newell, 1980, 1982; Pylyshyn, 1980, 1984a, b). In contrast, Connec-

tionists 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 Connectionist call "distributed representation".[1] The term 'Connectionist model' (like 'Turing Machine' or 'Von Neumann machine') is thus applied to a family of mechanisms that differ in details but share a galaxy of architectural commitments. We shall return to the characterization of these commitments below.

Connectionist networks have been analysed extensively—in some cases

[1]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 aggregate of units.

using advanced mathematical techniques.[2] They have also been simulated on computers and shown to exhibit interesting aggregate properties. For example, they can be "wired" to recognize patterns, to exhibit rule-like behavioral regularities, and to realize virtually any mapping from patterns of (input) parameters to patterns of (output) parameters—though in most cases multi-parameter, multi-valued mappings require very large numbers of units. Of even greater interest is the fact that such networks can be made to learn; this is achieved by modifying the weights on the connections as a function of certain kinds of feedback (the exact way in which this is done constitutes a preoccupation of Connectionist research and has lead to the development of such important techniques as "back propagation").

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

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

In the present paper we will proceed as follows. First, we discuss some methodological questions about levels of explanation that have become enmeshed in the substantive controversy over Connectionism. Second, we try to say what it is that makes Connectionist and Classical theories of mental

[2]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.

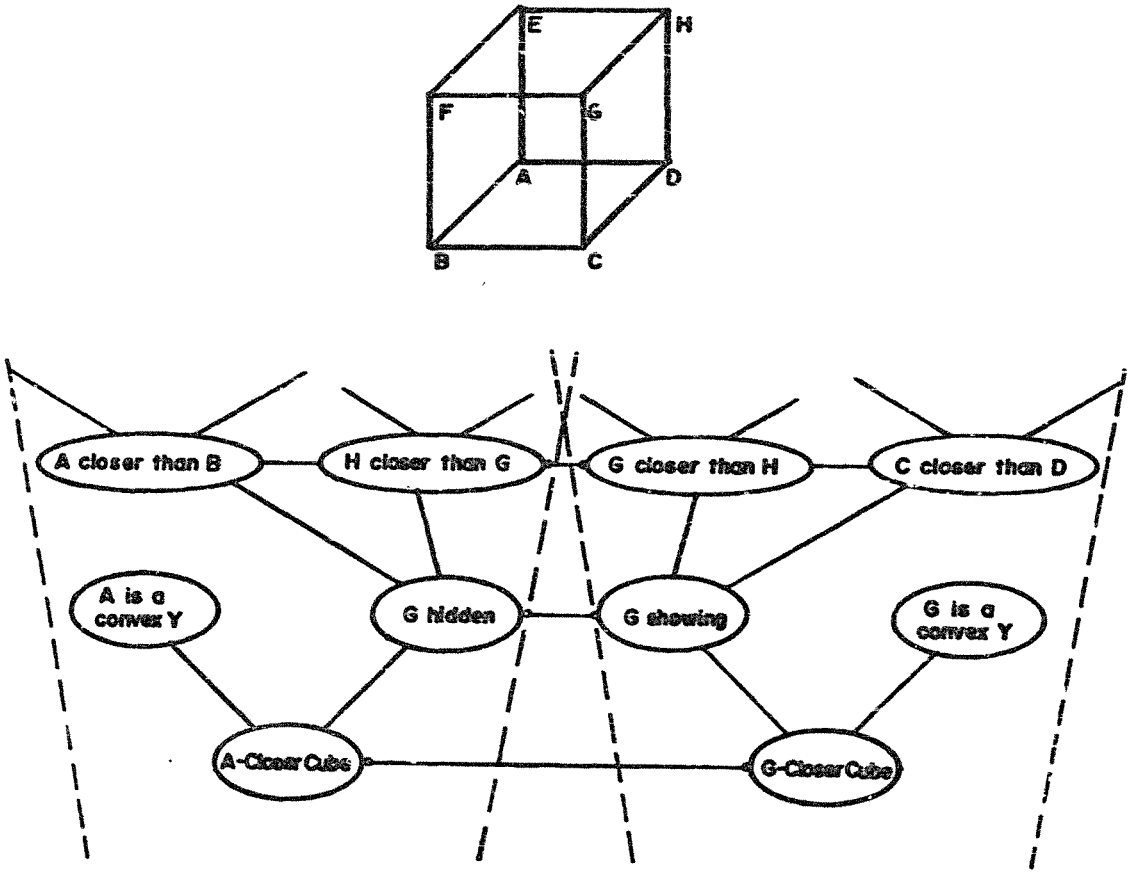
structure incompatible. Third, we review and extend some of the traditional arguments for the Classical architecture. Though these arguments have been somewhat recast, very little that we'll have to say here is entirely new. But we hope to make it clear how various aspects of the Classical doctrine cohere 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 and McClelland (1986a, p. 121) say, PDPs "are explicitly concerned with the problem of internal representation". 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 & 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.

Figure 1. *A Connectionist network model illustrating the two stable representations of the Necker cube. (Reproduced from Feldman and Ballard, 1982, p. 221, with permission of the publisher, Ablex Publishing Corporation.)*



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 "sub-symbolic"—and therefore presumably *not* representational. But this is misleading: Connectionist modeling is consistently Representationalist in practice, and Representationalism is generally endorsed by the very theorists who also like the idea of cognition 'emerging from the subsymbolic'. Thus, Rumelhart and McClelland (1986a, p. 121) insist that PDP models are "... strongly committed to the study of representation and process". Similarly, though Smolensky (1988, p. 2) takes Connectionism to articulate regularities at the "sub-symbolic level" of analysis, it turns out that sub-sym-

bolic 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".[3] 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 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.) Correspond-

[3]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.

ingly, 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.[4]

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.[5] In fact, the question whether Connectionist networks should be treated as models at some level of implementation is moot, and will be discussed at some length in Section 4.

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

[4]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.

[5]That Classical architectures can be implemented in networks is not disputed by Connectionists; see for example Rumelhart and McClelland (1986a, p. 118): "... 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.".

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.

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.[6] We will consider this idea in Section 4, where we will argue that it too is *not* what divides Classical from Connectionist architecture; Classical models *permit* a principled distinction between the etiologies of mental processes that are explicitly rule-governed and mental processes that aren't; but they don't demand one.

In short, the issue between Classical and Connectionist architecture is not about the explicitness of rules; as we'll presently see, Classical architecture is not, per se, committed to the idea that explicit rules mediate the etiology of behavior. And it is not about the reality of representational states; Classicists and Connectionists are all Representational Realists. And it is not about nonrepresentational architecture; a Connectionist neural network can perfectly well implement a Classical architecture at the cognitive level.

So, then, what *is* the disagreement between Classical and Connectionist architecture about?

[6]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 Section 4).

2. The nature of the dispute

Classicists and Connectionists all assign semantic content to *something*. Roughly, Connectionists assign semantic content to 'nodes' (that is, to units or aggregates of units; see footnote 1)—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.[7] 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 for 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 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

[7]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.

committed to "complex" mental representations or to "symbol structures".[8]

(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 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.[9] This is why Newell (1980) speaks of computational systems such as brains and Classical computers as "*physical* symbols systems".

[8]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 and 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.

[9]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

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

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 physical instantiation 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 syntax 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, $\mathbf{F}[P\&Q] = B(\mathbf{F}[P],\mathbf{F}[Q])$, where the function B specifies the physical relation that holds between physical states $\mathbf{F}[P]$ and $\mathbf{F}[Q]$. Here the property B serves to physically encode, (or 'instantiate') the relation that holds between the expressions P and Q, on the one hand, and the expressions $P\&Q$ 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 (1984a, pp. 54–69), where it is referred to as an 'instantiation function' and in Stabler (1985), where it is called a 'realization mapping'.

(1977) *Actors*—so long as the objects or the messages have a combinatorial structure which is causally implicated in the processing. Classical architecture is also neutral on the question whether the operations on the symbols are constrained to occur one at a time or whether many operations can occur at the same time.

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

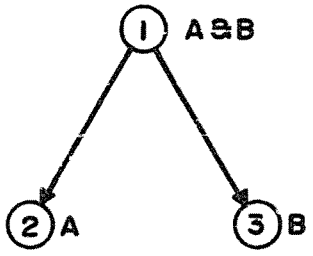
2.1. Complex mental representations

To begin with, consider a case of the most trivial sort: two machines, one Classical in spirit and one Connectionist.[10] Here is how the Connectionist machine might reason. There is a network of labelled nodes as in Figure 2. Paths between the nodes indicate the routes along which activation can spread (that is, they indicate the consequences that exciting one of the nodes has for determining the level of excitation of 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 7).

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:

[10]This illustration has not 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.

Figure 2. *A possible Connectionist network for drawing inferences from A&B to A or to B.*



'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.[11] By contrast, in the Connectionist machine none of this is 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., 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: (i) by failing to understand the difference between what arrays of symbols do in Classical machines and what node labels do in Con-

[11]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.

nectionist 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.

2.1.1. 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 diagram 2 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.[12]

2.1.2. Connectionist networks and graph structures

The *second* 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:

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 signifi-

[12]Labels aren't part of the *causal structure* of a Connectionist machine, but they may play an essential role in its *causal history* insofar as designers 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 is carried out. *This* sort of case of labels 'having effects' is thus quite different from the way that symbol tokens (e.g., tokened data structures) can affect the causal processes of a Classical machine.

cance *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.

2.1.3. Distributed representations

The *third* 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.[13] 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.

2.1.4. 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

[13]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.

of microfeatures. The model here is something like the relation between a defined expression and its defining feature analysis: thus, the concept BACHELOR might be thought to correspond to a vector in a space of features that includes ADULT, HUMAN, MALE, and MARRIED; i.e., as an assignment of the value + to the first two features and − 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 (p. 11), 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."[14] This treatment of the relation between

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

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.[15]

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

[15]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.

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 atomic 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 atomic (i.e., micro-) features, whose status is analogous to 'has-a-handle'. In particular, they have no internal syntactic analysis, and there is no structural relation (except the orthographic one) between the feature 'Mary-object' that occurs in the set {John-subject; loves; Mary-object } and the feature 'Mary-subject' that occurs in the set {Mary-subject; loves; John-object}. (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, McClelland, Rumelhart & Hinton, 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 complex symbols 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 microfeatures.

Consider a system in which the mental representation that is entertained when one believes that John loves Mary is the feature set {*+John-subject; +loves; +Mary-object*}. What representation corresponds to the belief that John loves Mary and Bill hates Sally? Suppose, pursuant to the present proposal, that it's the set {*+John-subject; +loves; +Mary-object; +Bill-subject; +hates; +Sally-object*}. We now have the problem of distinguishing that belief from the belief that John loves Sally and Bill hates Mary; and from the belief that John hates Mary and Bill loves Sally; and from the belief that 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-object* specifies whether it's the loving or the hating that she is the object of; similarly, mutatis mutandis, for the representation of John as *+John-subject.*

What has gone wrong isn't disastrous (yet). All that's required is to enrich the system of representations by recognizing features that correspond not to (for example) just being a subject, but rather to being the subject of a loving of Mary (the property that John has when John loves Mary) and being the subject 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-subject-hates-Mary-object.*

The disadvantage of this proposal is that it requires rather a lot of microfeatures.[16] 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 vocabulary of basic expressions 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 microfeatures 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 20-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 atomic symbols 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 subject) 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 atomic features. It is, of course, no accident that the Connectionist proposal for dealing with role relations runs into these sorts of problems. Subject, object and the rest are Classically defined *with respect to the geometry of constituent structure trees*. And Connectionist representations don't have constituents.

The idea that we should capture role relations by allowing features like *John-subject* 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 (and orthogonal) issue about whether the representation of commonsense concepts is 'distributed' (i.e., from questions like whether it's CUP or 'has-a-handle' or both that is semantically primitive in the language of thought).

It's worth adding that these problems about expressing the role relations are actually just a symptom of a more pervasive difficulty: A consequence of restricting the vehicles of mental representation to sets of atomic symbols is a notation that fails quite generally to express the way that concepts group

[16]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 microfeatures corresponding to these.

into propositions. To see this, let's continue to suppose that we have a network in which the nodes represent concepts rather than propositions (so that what corresponds to the thought that John loves Mary is a distribution of activation over the set of nodes {JOHN; LOVES; MARY} rather than the activation of a single node labelled JOHN LOVES MARY). Notice that it cannot plausibly be assumed that all the nodes that happen to be active at a given time will correspond to concepts that are constituents of the *same* proposition; least of all if the architecture is "massively parallel" so that many things are allowed to go on—many concepts are allowed to be entertained—simultaneously in a given mind. Imagine, then, the following situation: at time t, a man is looking at the sky (so the nodes corresponding to SKY and BLUE are active) and thinking that John loves Fido (so the nodes corresponding to JOHN, LOVES, and FIDO are active), and the node FIDO is connected to the node DOG (which is in turn connected to the node ANIMAL) in such fashion that DOG and ANIMAL are active too. We can, if you like, throw it in that the man has got an itch, so ITCH is also on.

According to the current theory of mental representation, this man's mind at t is specified by the vector {+JOHN, +LOVES, +FIDO, +DOG, +SKY, +BLUE, +ITCH, +ANIMAL}. And the question is: *which subvectors of this vector correspond to thoughts that the man is thinking?* Specifically, what is it about the man's representational state that determines that the simultaneous activation of the nodes, {JOHN, LOVES, FIDO} constitutes his thinking that John loves Fido, but the simultaneous activation of FIDO, ANIMAL and BLUE does *not* constitute his thinking that Fido is a blue animal? It seems that we made it too easy for ourselves when we identified the thought that John loves Mary with the vector {+JOHN, +LOVES, +MARY}; at best that works only on the assumption that JOHN, LOVES and MARY are the only nodes active when someone has that thought. And that's an assumption to which no theory of mental representation is entitled.

It's important to see that this problem arises precisely because the theory is trying to use sets of atomic representations to do a job that you really need complex representations for. Thus, the question we're wanting to answer is: Given the total set of nodes active at a time, what distinguishes the subvectors that correspond to propositions from the subvectors that don't? This question has a straightforward answer if, contrary to the present proposal, complex representations are assumed: When representations express concepts that belong to the same proposition, they are not merely simultaneously active, but also *in construction with each other*. By contrast, representations that express concepts that don't belong to the same proposition may be simultaneously active; but, they are ipso facto *not* in construction with each other.

In short, you need two degrees of freedom to specify the thoughts that an

intentional system is entertaining at a time: one parameter (active vs inactive) picks out the nodes that express concepts that the system has in mind; the other (in construction vs not) determines how the concepts that the system has in mind are distributed in the propositions that it entertains. For symbols to be "in construction" in this sense is just for them to be constituents of a complex symbol. Representations that are in construction form parts of a geometrical whole, *where the geometrical relations are themselves semantically significant*. Thus the representation that corresponds to the thought that John loves Fido is not a *set* of concepts but something like a *tree* of concepts, and it's the geometrical relations in this tree that mark (for example) the difference between the thought that John loves Fido and the thought that Fido loves John.

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

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

But, on second thought, this proposal merely begs the question that it set out to solve. For, if there's a problem about what justifies assigning the proposition *John loves Fido* as the content of the set {JOHN, LOVES, FIDO}, there is surely the same problem about what justifies assigning the proposition *Fido is the subject of bites* to the set {FIDO, SUBJECT-OF, BITES}. If this is not immediately clear, consider the case where the simultaneously active nodes are {FIDO, SUBJECT-OF, BITES, JOHN}. Is the propositional content that Fido bites or that John does?[17]

[17]It's especially important at this point not to make the mistake of confusing diagrams of Connectionist networks with constituent structure diagrams (see section 2.1.2 above). Connecting SUBJECT-OF with FIDO and BITES does not mean that when all three are active FIDO is the subject of BITES. A network diagram is not a specification of the internal structure of a complex mental representation. Rather, it's a specification of a pattern of causal dependencies among the states of activation of nodes. Connectivity in a network determines which sets of simultaneously active nodes are possible; but it has no *semantical* significance.

The difference between the paths between nodes that network diagrams exhibit and the paths between nodes that constituent structure diagrams exhibit is precisely that the latter but not the former specify parameters of mental representations. (In particular, they specify part/whole relations among the constituents of complex symbols.) Whereas network theories define semantic interpretations over sets of (causally intercon-

Strikingly enough, the point that we've been making in the past several paragraphs is very close to one that Kant made against the Associationists of his day. In "Transcendental Deduction (B)" of The First Critique, Kant remarks that:

> ... if I investigate ... the relation of the given modes of knowledge in any judgement, and distinguish it, as belonging to the understanding, from the relation according to laws of the reproductive imagination [e.g., according to the principles of association], which has only subjective validity, I find that a judgement is nothing but the manner in which given modes of knowledge are brought to the objective unity of apperception. This is what is intended by the copula "is". It is employed to distinguish the objective unity of given representations from the subjective .... Only in this way does there arise from the relation a *judgement*, that is a relation which is *objectively valid*, and so can be adequately distinguished from a relation of the same representations that would have only subjective validity—as when they are connected according to laws of association. In the latter case, all that I could say would be 'If I support a body, I feel an impression of weight'; I could not say, 'It, the body, is heavy'. Thus to say 'The body is heavy' is not merely to state that the two representations have always been conjoined in my perception, ... what we are asserting is that they are combined *in the object* ... (CPR, p. 159; emphasis Kant's)

A modern paraphrase might be: A theory of mental representation must distinguish the case when two concepts (e.g., THIS BODY, HEAVY) are merely *simultaneously entertained* from the case where, to put it roughly, the property that one of the concepts expresses is predicated of the thing that the other concept denotes (as in the thought: THIS BODY IS HEAVY). The relevant distinction is that while both concepts are "active" in both cases, in the latter case but *not* in the former the active concepts are in construction. Kant thinks that "this is what is intended by the copula 'is' ". But of course there are other notational devices that can serve to specify that concepts are in construction; notably the bracketing structure of constituency trees.

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

nected) representations of concepts, theories that acknowledge complex symbols define semantic interpretations over sets of representations of concepts *together with specifications of the constituency relations that hold among these representations.*

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

So much, then, for our long digression. We have now reviewed one of the 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.*

2.2. Structure sensitive operations

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

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

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

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

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

On the one hand, from the semantic point of view, (1) entails (2) (so, of

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

[18]And it doesn't work uniformly for English conjunction. Compare: *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 thought 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

nitive 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, commonsense 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.

2.2.1. Learning

If a Connectionist model is intended to learn, there will be processes that determine the weights of the connections among its units as a function of the character of its training. Typically in a Connectionist machine (such as a 'Boltzman Machine') the weights among connections are adjusted until the system's behavior comes to model the statistical properties of its inputs. In

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.

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 principle 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.

2.2.2. Reasoning

Association operates to alter the structure of a network *diachronically* as a function of its training. Connectionist models also contain a variety of types of 'relaxation' processes which determine the *synchronic* behavior of a network; specifically, they determine what output the device provides for a given pattern of inputs. In this respect, one can think of a Connectionist

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

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 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.