

1

Thunder in the Gap

John Dinsmore
Washington University

This is the unfolding story of *symbolicism* and *connectionism*, the argumentative paradigms who inhabit the modern study of cognition, and of the gap that stands between them. The purpose of this initial chapter is to provide a general overview of the debate between the paradigms, and to show where each of the remaining chapters slices through this debate. It also provides the introductory background, both technical and philosophical, necessary for the comprehension of the remaining chapters. I begin by looking at symbolism and connectionism independently, what each is and what each has accomplished or failed to accomplish. Then I look at ways in which the gap between them might be handled.

1. THE SYMBOLIC PARADIGM

Symbolic theories are *representation-oriented* in the sense that each begins by positing basic syntactic structures assumed to possess a transparent compositional semantics. Momentarily, we see that connectionist models, in contrast, are process-oriented. The greatest problems for symbolic theories begin in looking for the kinds of *processes* needed to produce the properties observed in human cognition.

In this section I try to characterize symbolism a little more accurately, and give a general assessment of the many strengths but especially of the apparent limitations found in current work within the symbolic paradigm.

1.1. Symbolic Mechanisms

The central principles that characterize traditional work in the symbolic paradigm can be summarized as follows:

- there are such things as *symbols*, which can be combined into larger *symbolic structures* (or *expressions*),
- these symbolic structures have a *combinatorial semantics* whereby what a symbolic structure represents is a function of what the parts represent, and
- at the same time all cognitive *processes* (reasoning) are manipulations of these symbolic structures.

This position is represented by Fodor and Pylyshyn (1988)—one of the most cited references in this book—Pylyshyn (1984), and Newell and Simon (1976; Newell, 1982). The latter incorporate this position as the central idea of what they called the *physical symbol system hypothesis*.

Other properties have often been attributed to symbolic systems, but either do not adhere very firmly on closer inspection, or do not turn out to be of critical importance in the debate between the paradigms. Most of these involve the thesis that the symbolic paradigm is *computationalist*, that is, that it takes the *computer metaphor* very seriously in talking about human cognition. Newell and Simon in fact borrowed specific assumptions related to modern digital computers, such as *universality* (the ability to stimulate any other system), *discreteness*, and the ability to execute *programs*. But as Derthick (1990) pointed out, these assumptions seem to go by the wayside in Fodor and Pylyshyn's well-known arguments against the viability of a connectionist cognitive science. Another feature sometimes attributed to the symbolic paradigm is the assumption of *seriality*, but there are simply too many exceptions in modern artificial intelligence, such as *marker-passing systems* discussed later, to take this attribute seriously, and the advent of massive parallelism in computer architecture no longer supports seriality as characteristic of computers in any case. Nevertheless, I return later to the issue of whether the abstract concept of *computation* offers a principled distinction between the paradigms.

1.2. Examples of Symbolic Models

The principles listed earlier are central in the sense that symbolic theories tend to conform to these principles, but, as becomes evident, systems often deviate from them in restricted ways. I look at a couple of kinds of symbolic models that come up most often in discussing contrasts between the paradigms.

1.2.1. Logic and Rule-based Systems. Perhaps the prototype of symbolic models is the *logic-based system*, especially common in artificial intelligence, whereby a proposition is expressed discretely by a predicate (a symbol for a relation) and a linear sequence of arguments (each a symbol for an object). Inference procedures define manipulations of those structures, generally in a manner that preserves meanings or truth conditions. For example, such a procedure would allow the presence of the following expressions,

bird(Tweety)
 bird(X) \rightarrow feathered (X)

to result in the derivation of the following expression.

feathered (Tweety)

Alongside logic-based systems are *rule-based* or *production* systems, common in psychology and linguistics as well as artificial intelligence. These systems interpret *rules* (each of which has a *condition* that potentially matches some symbolic structure and an *action* that specifies some symbolic manipulations) by repeatedly selecting a rule (generally one of many) whose condition is satisfied and then executing the action. For instance, the following rule might cause the inference discussed earlier to be derived.

condition: bird(X)
 action: add feathered (X)

1.2.2. Associative Networks. Also known as *semantic networks*, these were introduced in psychology (Quillian, 1968) to model *associative memory*. We observe associative memory when the sound of a voice causes us to think about the person belonging to the voice. An associative network represents the concepts in an encyclopedic way, as a set of nodes interconnected with links labeled for the relations holding between concepts. The original semantic network models attributed to particular symbolic nodes *activations* that increased when a node was processed, but automatically *spread* to connected nodes. For instance, the kind of *semantic priming* that occurs in a sentence like *The astronomer married a star*, whereby the interpretation of *star* as a celestial object, strongly suggested by *astronomer*, confuses the interpretation, can be modeled in an associative network by activation spreading from the node for astronomer.

In many associative network models the passing of discrete symbolic *markers* in a much more controlled manner replaces spreading activation (see Lange, chap. 10 in this volume). Marker passing, however, is capable of achieving many specific inferences through graph-traversal instead of by rule application. The links serve as an indexing mechanism to facilitate matching one structure against another and finding intersections of sets or concepts.

Associative networks are of interest in the symbolism/connectionism debate because they are perhaps the most connectionist of the symbolic models. In fact, although work in associative networks is generally assumed to belong firmly in the symbolic tradition, we doubtlessly find some stretching of some of the central premises of the symbolic paradigm in this framework: In associative networks manipulations are sensitive to current activation levels (or the presence of certain markers) yet these do not contribute to semantics. Thus processing in associative networks is not defined in terms of strictly symbolic structures. Lange (chap. 10 in this volume) and some others like to classify these as connectionist in a very

broad sense to underscore these similarities while at the same time recognizing that they are at the same time fundamentally symbolic.

1.3. Representations in Symbolic Systems

Philosophers, linguists, psychologists, researchers in artificial intelligence, and many others have been comfortable with the languagelike expressive abilities provided in the symbolic paradigm, the clarity with which the symbolic paradigm treats reasoning in science or other conscious domains and its natural affinity for dealing with concepts at a level subject to conscious introspection. Almost all of the practical successes in artificial intelligence in implementing higher-level abilities in such domains as expert systems, language understanding, machine translation, goal-oriented planning, and mathematical reasoning, have relied on symbolicism.

The symbolic paradigm is in effect designed around its support for representation, and it is therefore hardly surprising that this is where its strengths lie. Fodor and Pylyshyn (1988) saw this strength in *systematicity* and *productivity* of cognitive representations, which arise from the way the compositionality of symbolic expressions allows for the decomposition and recomposition of representations. Barnden (chap. 7 in this volume) contrasts these representational advantages with those of connectionist representations.

1.4. Learning

What has been simple for symbolic systems is simply acquiring information already encoded in some propositional language, that is, *learning by being told*. However symbolic models are very poor at adapting or organizing themselves dynamically on the basis of experience. Although a lot of attention has been given within artificial intelligence to symbolic learning algorithms (Michalski, Carbonell, & Mitchell, 1983, 1986), each of these generally works in some specialized domain and does not reorganize the system in any profound way. The lack of generality in learning means that computational symbolic models are invariably *hand-coded* with a particular algorithm in mind. This contrasts very markedly with connectionist systems for which very general and powerful learning techniques exist, which make it unnecessary to explicitly program the networks for specific behaviors.

1.5. Mysterious Processes

There appear to be many cognitive processes that do involve symbolic manipulations, in the sense of mapping one symbolic structure, A, into another, B, but nevertheless resist symbolic analysis, in the sense of an explicit specification of what the relation of A to B must be strictly in boolean terms of the symbolic

structures involved. Dinsmore (1991) called symbolic processes with this property *mysterious*. Let's look at a few processes that appear to be mysterious, given symbolic experience, at this time.

1.5.1. Holistic Processes. As an example, face recognition might be seen abstractly as a mapping that takes a set of symbolic structures representing facial features, and returns a symbol representing some individual. Now, humans can recognize the same face in a variety of contexts, from different angles, under different illuminations, with parts obscured by shadows or other objects, and so on. No known boolean combination of features determines this mapping, yet a given face is recognized fairly reliably. Dreyfus and Dreyfus (1986) described such processes as *holistic*, resisting the explicit decomposition of objects into component features.

1.5.2. Noise and Unexpected Input. Symbolic models are notoriously brittle when presented with unusual, unplanned, faulty, or noisy input. Even where the symbolic mapping is analyzable under ideal noise-free conditions, noise introduces a degree of complexity into the mapping that humans, but not symbolic models, are typically capable of overcoming.

Kwasny and Faisal (chap. 9 in this volume) discuss some examples in the context of parsing natural language. Symbolic parsers generally cannot successfully parse ungrammatical sentences like *John did hitting Jack*, even though they are close to grammatical sentences and can be processed by humans. The symbolic parsers that can handle such sentences do so by explicitly anticipating such anomalies, in effect making the grammar accepted by the parser more general to accommodate them. There is a problem because there are always new instances that will not be anticipated, so that the process involved must either become enormously complex or remain inadequate. Kwansy and Faisal in fact handle such cases by building in a connectionist component to account for the mysterious mapping.

1.5.3. Associative Memory. Associations between concepts have proven particularly resistant to symbolic analysis. Such associations should allow the retrieval of a complete memory given a partial description of the content of the memory, such as *high-ranking government official who is not too bright*. Examples from natural language processing are semantic priming (discussed earlier), *metonymy* (the reference to one object or concept by means of an expression that stands for an associated object or concept) and the resolution of anaphoric, for example, pronominal, reference.

1.5.4. The Problem of Mysterious Processes for Symbolicism. In summary, mysterious processes can be seen to map symbolic representations to others as required in the symbolic paradigm, but the symbolic vocabulary often does not

seem to allow us to fully analyze the mapping, not to the degree that we can predict which representations will get mapped onto which, or such that we could develop a purely symbolic artificially intelligent program that achieves behavior that depends on mysterious processes. It follows that there can never be a complete, strictly symbolic theory of cognition. Work in connectionism, to which we now turn, suggests that dropping down to a lower, *subsymbolic*, level allows mysterious processes to be analyzed more successfully in terms of a nonsymbolic vocabulary.

2. THE CONNECTIONIST PARADIGM

The connectionist paradigm is, in marked contrast to the symbolic paradigm, *process-oriented*. That is, it starts with the assumption that the basic objects posited in a theory of cognition will be processed in a simple and usually uniform way. Unfortunately, this orientation then leaves connectionist models with the rather daunting task of discovering how representations emerge from the simple processing components.

This section provides a brief description of connectionism and lists some of its accomplishments and problems. Keep in mind that not as much work has been done to date in connectionist theories as in symbolic theories. It is hardly surprising, therefore, that (a) connectionism has logged fewer substantial achievements than symbolism, (b) fewer clear weaknesses in the connectionist paradigm have been demonstrated, and (c) researchers in connectionism tend to have very high expectations of future success.

2.1. Connectionism in a Nutshell

A connectionist system is a network of very simple processing *units* that stores knowledge in the *weights* of the *connections* between units and realizes computation in the dynamic global behavior that results from the local interactions among units. The following few pages give a basic introduction to the fundamentals of connectionism equivalent to a one-semester university course, and in particular introduces some of the terminology used in the other chapters of this book. The uninitiated will have to hold onto their hats, but in principle should come out with enough background to appreciate the rest of the book.

2.1.1. Units and Connections. A *connectionist* model consists of a set of *units* linked together in a network by a set of *connections*. Units and connections are neurologically inspired by neurons and synapses respectively. At a given time each unit has an *activation* level and produces an *output*, which affects other units through connections to those units.

Typically certain units are assumed to communicate with the outside world. *Input units* are often *clamped* onto certain activation values (as if due to some environmental influence, much like receptors in the retina of the eye react to impinging light). *Output units* are monitored for their output values (as if producing effects in the outside world). Figure 1.1 illustrates the functioning of an intermediate unit.

The specific behavior of a unit varies from one connectionist model to another. Activations and outputs are either continuous or discrete scalar values. The output of one unit can affect the activation level of another unit if there is a connection between the two units. A connection has a *weight*, which determines the extent to which the output of one unit can affect the activation of the other unit. A weight typically assumes a continuous value in a scale between +1.0 and -1.0. A high positive weight and a high output from the source unit will tend to raise the activation level of the destination unit. A negative weight and a high output from the source unit will tend to diminish the activation level of the destination unit. Positively weighted connections are said to be *excitatory*, whereas negatively weighted connections are said to be *inhibitory*. The activation of a unit typically changes very rapidly, but weights on connections typically change very slowly over time as part of *learning*, described later.

A unit adjusts its own activation level in response to the input it receives from other units through incoming connections. How a unit does this is described by an *activation function*. Almost always the input to a unit is determined by a *sigma-pi* computation, that is, by summing up individual incoming signals, each of which is the output produced by the unit on the other side of a connection scaled down by the weight of the incoming connection. Activation functions vary from one connectionist model to another. Very often a *sigmoid* function is used to squish the activation down to fit in a range from 0 to 1. Often the activation function will take the previous activation of the unit into account such that, for instance, if the input signal drops the activation of the unit will only gradually *decay*. Often some kind of *bias* is added in to influence the value of the activation function.

A unit produces an *output level*, which is what other units see, as a function of the activation level. Again output functions vary from one connectionist model to another. Very often the output is simply identified with the activation level. Very often the output will be a discrete value, 0 or 1, while the activation will have a continuous value. Often the output is 0 until the activation level crosses some *threshold* value.

Often the output is a nondeterministic *stochastic* function of the activation value, such that even with a fixed activation level the output will toggle between 0 and 1 with a higher activation level making it more likely that at a given time the output will be 1 rather than 0. The analogy is drawn between these last models and the statistical behavior of *thermodynamic* systems. One such connec-

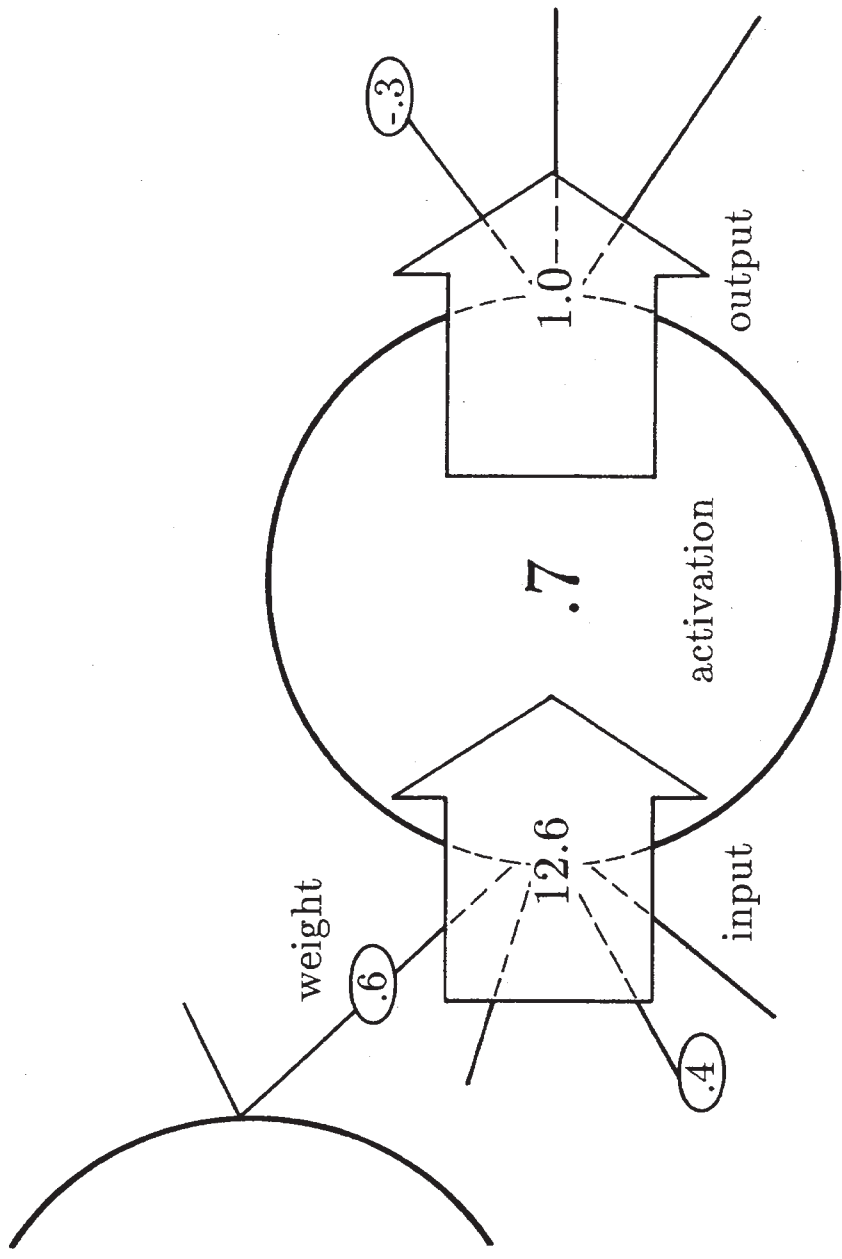


FIG. 1.1. A unit and its connections illustrating summation of input, adjustment of activation, and production of output.

tionist system is called the *Bolzman machine*. Often the stochastic function is varied so that it becomes increasingly deterministic as the global computation proceeds, much as a drop in temperature makes a thermodynamic system more predictable. The use of stochastic output functions with the element of decreasing temperature built in is referred to as *simulated annealing*.

2.1.2. *Configurations of Units and Connections.* The computational properties of a connectionist network result not only from the particular activation and output functions used, but also—not surprisingly—of the way the units and connections are interconnected. Figure 1.2 illustrates a very common connectionist architecture, a *feed-forward network*, in which units are grouped into *layers*, with a connection running from each unit in one layer to each unit in the layer above. If the units in the input layer of Fig. 1.2 are clamped onto certain activation levels, activation changes will propagate upward through the hidden layers and the units in the output layer will converge on certain activation values. Effectively, some pattern of activation on the input will be mapped onto some pattern of activation on the output, with the mapping dependent on the connection weights and the particular activation and output functions used in the system. This configuration is thus being used as a *pattern associator*. A pattern

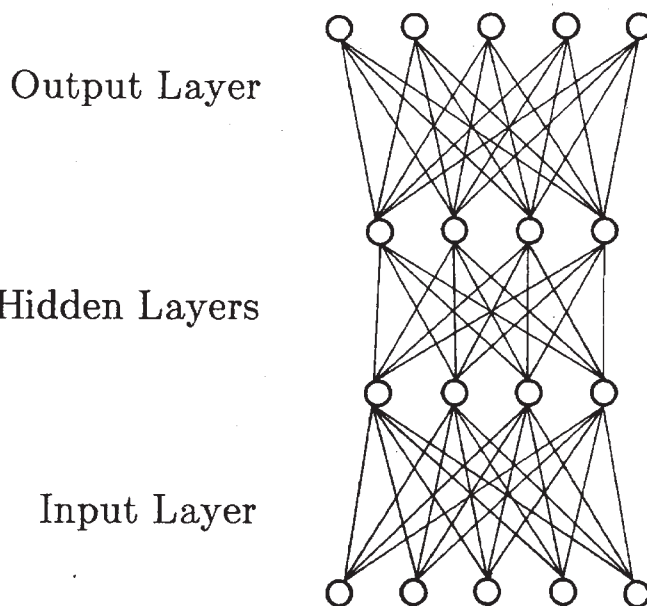


FIG. 1.2. A simple feedforward network in which activation flows from the input layer and through hidden layers to produce a pattern of activation on the output layer.

associator that maps a pattern of activation on the input layer onto the same pattern on the output layer is an *autoassociator*. Autoassociators are interesting for reasons discussed later. The mappings can also be very complex. This can be appreciated intuitively if one thinks of each unit on the first hidden layer as serving to measure how closely some possible pattern on the input layer (which pattern depends on the weights of the incoming connections) is realized, each unit on the second hidden layer as serving to measure how well some possible pattern on the first hidden layer is realized and so on.

Many alternative configurations can be regarded as variations of that of Fig. 1.2. A common kind of variation forces a layer or some subset of units in a layer (sometimes called a *clique*) to assume a pattern of activation in which a *single* unit is *on* or highly activated and the remaining units are *off* or minimally activated. This effect can be produced by interconnecting the units within the clique by *inhibitory* connections, producing competition among units as the more highly activated units try to turn off the less activated, as well as each other. Such a subnetwork is a *winner-take-all* network. This is used in the McClelland and Rumelhart model of letter perception discussed by Adams, Aizawa, and Fuller (chap. 3 in this volume) and by Aizawa (chap. 4 in this volume). Another variation is the *simple recurrent network* discussed by Lee and Gasser (chap. 8 in this volume) and Blank, Meeden, and Marshall (chap. 6 in this volume), which makes use of connections from a hidden layer down to a clique in a lower layer to make the behavior of the network effectively dependent on the previous states of the network.

2.1.3. Learning. A real strength of connectionist networks is their ability to adaptively organize themselves to acquire appropriate behavior. A connectionist network can learn from experience through procedures that adjust weights gradually in the direction of more appropriate responses to environmental input. The most common learning method in layered networks like that illustrated in Fig. 1.2 is *backpropagation*. Backpropagation is a *supervised* technique that requires the presentation of a series of training exemplars, each of which represents an input pattern and a target output pattern. The backpropagation algorithm begins by comparing the output pattern actually produced with the target output pattern and propagates a series of error-correcting weight changes backward level by level from the output units to the input units. The connection weight changes are gradual so that the network can adapt simultaneously to a large number of training exemplars to produce the desired mapping with a low error rate.

2.1.4. Semantics. Connectionist networks are useful for low-level pattern recognition tasks, for instance those that might well fall within a housefly's capabilities as well as within a human's. Nevertheless, as Fodor and Pylyshyn (1988) pointed out, virtually all connectionists agree with symbolicists that pos-

tulating complex representational states is essential to a theory of cognition. Some connectionist models, called *localist*, assume that individual units stand for individual objects or concepts. For instance, one unit might represent your dog Fido, another your grandmother's rocking chair.

Other models, called *distributed*, make the more intriguing assumption that objects and concepts are represented as patterns of activity over many units, with each unit participating in many patterns. Distributed models exhibit many distinctive properties, to be described presently, whereas localist models tend to be more akin to symbolic models. Lange (chap. 10 in this volume) contrasts the differences between localist and distributed connectionist models in more detail.

For the most part the chapters in this book focus on distributed connectionist models. Processing in all connectionist models is realized only in direct, interactions among individual units, so under the assumption of distributed representations the level at which processing is described (individual units) is lower than the level at which representations are described (patterns of activation over many units). Distributed connectionist models are therefore said to make use of *sub-symbolic* processing.

A very interesting aspect of distributed representations is that the acquisition of new representations seems to follow from general learning techniques like backpropagation. However, with localist models, or for that matter symbolic models, there is no obvious account of how the decision is made that some given node is now going to stand for your grandmother's rocking chair. Experiments with backpropagation in autoassociative networks (a specific kind of pattern associator defined earlier) reveal the ability of a network to acquire new distributed representations of locally encoded concepts. The idea is to train the network on the identity mapping, but set up the network so that the hidden levels have fewer units than the input and output layers. Each pattern of activation that appears on a hidden level can be viewed as representing the same thing as the autoassociated pattern—but more compactly—but is in effect invented by the network itself. This technique is exploited by Blank, Meeden, and Marshall (chap. 6 in this volume).

2.2. Does Connectionism Take the Brain Seriously?

Much of the appeal of connectionism is doubtlessly its basis in brain architecture, which potentially gives it more of a hard science flavor. On the other hand, this basis tends to be quite tenuous. As Aizawa (chap. 4 in this volume) discusses, many connectionist mechanisms have no known neural counterparts. For instance, backpropagation, as its name suggests, requires the propagation of a signal backward along connections in a manner that is not generally observed physically. Also, less frequently observed, many neural mechanisms have no counterparts in connectionist models, for the most part because they are poorly

understood. For instance, little in existing connectionist models correspond to hormonal and other chemical mechanisms, or to large-scale anatomical structures observed in the brain.

2.3. Limitations of Connectionist Learning

Although the availability of generalized learning techniques is one of the great achievements of connectionist research, these techniques have critical limitations. First, learning is generally achieved by providing highly reliable training instances in a very systematic way, normally iterating over all desirable input-output correspondences a number of times. The quality and systematicity of training by a more realistic, random sampling of the environment would fall far short of that achieved artificially. Aizawa (chap. 4 in this volume) discusses the problem of the absence of external sources of supervision.

Second, learning is generally achieved only after hundreds of thousands of training instances. The slow speeds of human neural architecture, in contrast to that of the computers on which simulations are run, make so many iterations infeasible.

Third, these techniques may have difficulty in scaling up tractably from the relatively small networks and relatively simple problems for which they have been tested. This is not only because the time required to train a network could well explode exponentially with an increase in the size of the network, but also because the learning algorithms for connectionist networks are in essence *hill-climbing* techniques, which have been shown in artificial intelligence to be of limited usefulness for problems of significant complexity (Minsky & Papert, 1988, epilogue).

But aside from the slow rate and lack of reliability of connectionism's standard learning techniques, the most pervasive and successful kind of learning in symbolic models is virtually unknown in connectionist models, that is, learning *by being told*. No mechanisms have been proposed for connectionist systems to account for the extremely rapid acquisition by humans of very complex information, found, for instance, in the interpretation of conscious rules, like *put the green ones over here and the red ones over there*, that begin to affect behavior immediately after acquisition.

2.4. Representing Things in Connectionist Models

Connectionism is often criticized for not being able to deal naturally with compositional representation structures (Fodor & Pylyshyn, 1988), or for simply providing *implementations* of structures properly described at the symbolic level without adding anything to symbolic descriptions. Representing things adequately is one of the fundamental challenges to connectionism.

2.4.1. *Simple Compositional Structures.* Over the last decade quite a bit of progress has been made in representing simple relations like *Bif sees Matilda* in distributed connectionist networks. In symbolic theories this is easy: simply concatenate symbols. In connectionism it was not initially clear how this would be done: *Bif* might be represented by a pattern of activation, as might *Matilda* and the concept of seeing, but what would represent the whole thing?

Blank, Meeden, and Marshall (chap. 6 in this volume) present one very general approach to this that allows for the representation of arbitrary sequences of representations. This approach makes use of a technique of compressing a concatenative structure into a smaller hidden layer by training for autoassociation. The representation on the hidden layer can be used as one of the concatenated elements on the input layer in producing a compressed representation for a yet larger structure. The compressed representation is created by the system itself, through backpropagation, and does not have any transparent internal structure, yet can be easily translated into a concatenative structure, thus demonstrating that the original content is faithfully preserved. Blank, Meeden, and Marshall show additionally how many of the processing advantages of general connectionist models over symbolic models are preserved in their model.

Similarly, Lee and Gasser (chap. 8 in this volume) make use of the framework of a *simple recurrent network* in representing the sequences of phonemes that represent words in their system for learning the past tense. Simple recurrent networks represent concatenative structures as temporally ordered activation patterns.

2.4.2. *Rules.* We have seen that many symbolic systems are based on rules that match symbolic structures and typically produce new symbolic structures. By training a pattern associator like that in Fig. 1.2 the network can do a fairly good job, by simple cumulative connection weight adjustments, at picking up the generalizations expressed by a set of symbolic rules. Although it has behavioral properties like rules, the representation itself is implicit because in the connectionist model individual rules cannot be recognized (Rumelhart & McClelland, 1986). Lee and Gasser (chap. 8 in this volume) give a very good discussion of the acquisition of such rulelike behavior in natural language phonology.

Pinker and Prince (1988) pointed out that symbolic linguistic models not only depend on rules, but on ordering of rules and on the existence of hidden *underlying representations* on which these rules operate. For instance, the English past tense is expressed in a number of ways, each of them derived by the application of some rule to a common underlying form. Lee and Gasser (chap. 8 in this volume) show how an appropriately trained network might even create its own underlying representations as patterns on hidden units as it learns the implicit rules that manipulate those representations.

Barnden (chap. 7 in this volume), on the other hand, shows why explicit rules like those in symbolic models are needed in addition to the implicit rules realized

by pattern association. For instance, he points out that a network might represent the sentence, *Most communists in town are atheists*, implicitly as a pattern associator that when presented with a pattern representing a *communist (in town)* on its input layer would produce a pattern representing *atheist* on its output layer. However, it is difficult to see how such an associator would be used in building a higher-level representation for *If most communists in town are atheists then the town is eligible for a grant from the Fundie Fund*.

2.4.3. *Complex Relations/Propositional Attitudes*. Then there are higher-order representations. For instance, given a complex distributed representation for *John loves Mary*, how does the connectionist network represent that the agent *knows* that John loves Mary? Barnden (chap. 7 in this volume) discusses the multitudinous problems that arise in connectionist attempts to represent individuals' beliefs. These problems, as the reader will see, are far greater than that of representing simple relations, which connectionism is just beginning to get a handle on.

2.5. Observations About Connectionist Semantics

One of the remarkable capabilities of distributed connectionist systems is the automatic creation of new representations. New representations in a symbolic system are just assumed when needed (for instance, by a call to *gensym*), leaving the question open of how the system fixes the referent of the symbol. The creation of representations is illustrated by the technique of inducing the system to create a more compact representation on the hidden layer through autoassociation previously described. Such representations tend to develop what Blank, Meeden, and Marshall (chap. 6 in this volume) call a *microsemantics* whereby representations that are learned in the same kinds of contexts tend to be similar activation patterns. Chalmers (chap. 2 in this volume) argues that this does a far better job at fixing referents than is found in symbolic models.

Also, because connectionism typically makes a clear break between processing and semantics, connectionist models are in principle immune to many of the criticisms already raised to symbolic models. First, although certain patterns of activation may denote things out there in the world, there is no requirement that all patterns of activation do this. Thus, processes without semantic counterparts, as found, for instance, in making aesthetic judgments, are entirely consistent with the connectionist paradigm (this is not to say that anyone has any specific idea at this time of how these would occur in a connectionist network).

2.6. Mysterious Processes from a Connectionist Point of View

We have seen how connectionist systems are able to organize themselves, developing their own distributed representations. This results in some remarkable

properties, which are especially apparent when we look at processes that are mysterious from a symbolic point of view. These processes seem to emerge naturally as consequences of the connectionist architecture.

Connectionist models are very good at *generalizing* information. For instance, Lee and Gasser's (chap. 8 in this volume) system is trained to produce past-tense forms of a set of verbs. When presented with verb roots it has not been trained on, it will generally produce the correct form of the past tense; it has made generalizations equivalent, roughly, to linguistic rules of phonology. Representations created in similar contexts will also tend to be realized as similar patterns of activation. As a result, connectionist systems seem to naturally *classify* concepts. Blank, Meeden, and Marshall show how their system develops classifications like *aggressive animal* based on the sentences in which such concepts are mentioned.

Connectionist models have been especially successful in finding a *best match* when a number of competing interpretations are available. Kwasny and Faisal (chap. 9 in this volume) describe a hybrid (with symbolic and connectionist components) system that overcomes the brittleness normally associated with rule selection in symbolic systems by handling this decision in a connectionist component. In the presence of ungrammatical natural language input the system naturally selects the rule that matches best. In general, this illustrates the ability of connectionist systems to reach interpretations fairly reliably in spite of *noise* in the input.

Likewise, the most common mechanisms in symbolic models to access memories rely on knowing where to look for specific information. How to find a complete memory given a partial description of its content has been a problem. However, such retrieval occurs naturally in connectionist models. For instance Rumelhart, Smolensky, McClelland, and Hinton (1986) developed a model that learns generalizations about the objects that occur in different kinds of rooms. Given an input representing, for instance, a toaster and a sink, the network naturally completes the representation by filling in a refrigerator, an oven, and so on.

Another kind of task that connectionism seems to handle effortlessly is that of satisfaction of multiple simultaneous *soft constraints* as an interactive process that finds a compromise (Adams, Aizawa, & Fuller, chap. 3 in this volume). Rumelhart, Smolensky, McClelland and Hinton (1986) demonstrated how the system described in the previous paragraph naturally makes compromises among conflicting schemas for types of rooms when given an input representation, for instance, for a room with a bathtub and a toaster.

Connectionist networks tend to be relatively good at handling what are arguably holistic processes, such as recognizing faces. In fact, Blank, Meeden, and Marshall's (chap. 6 in this volume) show how a connectionist system actually develops holistic compositional representations. Fodor and Pylyshyn (1988) argued that compositional representations must be *concatenative*, that is, decomposable into parts such that structure-sensitive processes can manipulate the representation. Blank, Meeden, and Marshall's model for learning short sen-

tences, on the other hand, actually translates a concatenative representation into a holistic encoding, in which subrepresentations cannot be individually isolated. This result is particularly interesting because they show that what in symbolic terms would be structure-sensitive processes can operate on the holistic representations themselves. First, the holistic representation can be translated back into a concatenative representation. Second, pattern associators can be trained that can determine from the holistic whether, for instance, an aggressive animal is mentioned in the original sentence. Third, syntactic transformations such as passivization can be performed directly over the holistic encodings.

2.7. High-Level Processing

There are a lot of high-level symbolic processes, and some not so high-level, that are not matched in existing connectionist systems. For instance, it is hard to get multiple instantiations of a generic representation in connectionist models and we don't know how to set up or forget complex associations between generalizations and specific instances quickly as is done through role and variable binding in symbolic models.

Especially focusing on propositional attitudes, Barnden (chap. 7 in this volume) offers an extended discussion of the kinds of problems that occur in high-level reasoning tasks that are particularly challenging to connectionism. Among these are the need to match different short-term information structures, the need to deal with anomalous combinations of concepts, and the need to perform embedded reasoning, for instance to reason about a particular person's beliefs.

More generally, full cognitive systems are complex. It has been very difficult for connectionist systems to move beyond modeling one component of a cognitive system. McCarthy (1988) called this deficient property *elaboration tolerance*. This is certainly largely a result of the fact that a pattern of activity created through weight adjustments by one connectionist network will generally not be interpretable by another.

Aizawa (chap. 4 in this volume) discusses the reliance of current connectionist models on outside intervention, and investigates the viability of adding additional connectionist modules to systems to make them more complete. For instance, the Boltzman machine's knowledge of the world is reflected in the relative frequency in which patterns of activation occur. However there is no immediate way to use knowledge in other processes, for instance in reporting beliefs. He proposes a special *recall module* that is able to summarize relative frequencies in order to proceed with such processes. But Aizawa worries that such a module would be entirely ad hoc, designed to acquire behaviors that do not emerge naturally from the underlying connectionist architecture.

2.8. Summary

There are some very basic differences between symbolic and connectionist systems, in the level at which processing is described, in the nature of representa-

tion, and so on. The capabilities of the systems seem to be quite different, in pattern matching, in holistic processing, in high-level compositional representation. What should we make of these differences?

3. THE GAP

Pinker and Prince (1988) discussed three different philosophies about the relationship of connectionism to symbolicism: *Implementational connectionism* describes the philosophy that connectionist networks simply provide an alternative means of implementing understood symbolic structures and processes. *Eliminative connectionism* describes the opposing, more radical, position that the symbolic level is not necessary at all to a complete description of cognition. Finally, *revisionist connectionism* describes a more subtle, intermediate position whereby connectionism will lead to different sets of symbols and operations and to new discoveries about the nature of symbolic processing, even while the levels remain distinct. I organize the discussion approximately around the first two possible positions and a generalization of the third.

3.1. Eliminativism

Eliminativism sees connectionism as the wave of the future, destined to sweep away symbolicism. Presumably no one seriously wants to eliminate connectionism itself (however much one may criticize the current direction of connectionist research, one must accept that the brain is a connectionist system of some sort and worth studying as such). The alternative to eliminativism would be a view in which the symbolic and the connectionist coexist, but at different *levels*. Let's look briefly at the case for and against eliminationism.

3.1.1. Profound Computational Differences? A potential argument for eliminativism is that the symbolic paradigm is inextricably linked to computationalism and computationalism is capable in principle of only an inadequate account of cognition, that is, that the computer metaphor must fail. Chalmers (chap. 2 in this volume) answers this by pointing out that connectionist networks can be simulated by computational devices and computational devices can be simulated by connectionist networks and therefore cannot be distinguished as computational versus noncomputational. Adams, Aizawa, and Fuller (chap. 3 in this volume) attack this question without recourse to the simulation argument by comparing connectionist architectures with the most general computational architecture, the Turing machine, and reasonable extensions thereof, with respect to a number of alleged differences such as explicitness, discreteness, distributed representations, and structure-sensitivity. They discover that these are not criterial in distinguishing symbolic models from connectionist models.

One of the most outspoken critics of computationalism is Searle, whose

Chinese Room argument hinges on showing that "syntax is not sufficient for semantics." Chalmers (chap. 2 in this volume) argues that his criticisms actually only apply to symbolic theories, in which semantics is defined for the same level at which syntactic manipulations take place. While not endorsing Searle's arguments in their entirety, Chalmers points out that they fail in distributed connectionist models in which representations have a rich internal (subsymbolic) structure. The fundamental differences between symbolism and connectionism is for Chalmers in the level at which semantics attaches.

3.1.2. Connectionism Needs Symbolicism. Clearly, connectionist networks with significant cognitive abilities will have to be far more complex than the simple computational models that now exist. But as models become more complex, they will become more opaque (recall also that connectionist networks are for the most part trained, not programmed). For this reason connectionist models cannot dispense with higher-level functional abstractions that summarize large parts of the network structure or behavior. Dyer (1988) imagined running a large connectionist network for 20 years and sending it to college and pointed out that there would be no way to understand what had been produced, even though the network might exhibit intelligent behavior.

Lachter and Bever (1988) suggested that symbolic principles are actually often snuck into the design of a connectionist network to achieve higher-level behavior. For instance, a certain model of speech production crucially arrays nodes at different levels of lexical and phonological representation. However, these levels are those originally posited in symbolic models as representations over which rules can operate. Pinker and Prince (1988) similarly argued that rulelike behavior is often achieved by putting a system into an unrealistic teaching environment tailored to result in behavior they wanted to see in order to make up for lack of explicit macrororganization. In short, the contribution of symbolic theories is not always properly acknowledged.

The recognition of higher-level structures is in no way inconsistent with the thesis that cognition emerges from connectionist principles. But like many kinds of physical systems, the connectionist architecture can be assumed to organize itself into higher-level structures and processes with their own properties that call for a higher level of description. We should always look for higher-level organizational principles.

3.2. Implementationalism

Implementationalism sees the symbolic system as a virtual machine that in humans happens to be implemented on top of connectionist hardware, but could just as well have been implemented to run on some other machine. Symbolicism and connectionism for the implementationalist belong to strictly distinct levels of description.

3.2.1. *The Justification for Separate Levels of Description.* Implementationalism has traditionally been an ally of the symbolic paradigm (e.g., Marr, 1982). Putnam (1973) in developing the philosophy of *functionalism*, drew the analogy to computer systems. A high-level language like LISP can be implemented in any of a variety of hardware machines, and therefore has properties that are independent of any particular machine. A single programming language is therefore *multiply realizable*. Additionally, the same hardware machine can be used to implement any high-level programming language, for instance, PROLOG and SMALLTALK in addition to LISP. Another example of multiple realizability is the implementation of the same mathematics by the abacus, by an electronic calculator and by the brain (Schwartz, chap. 5 in this volume). In fact all of science—physics, chemistry, biology, zoology, sociology, and so on—seems to naturally fall similarly into relatively independent levels of description. Dinsmore (1991, chap. 1) discussed levels of description in more detail.

3.2.2. *Symbolicism Needs Connectionism.* Against implementationalism Smolensky's (1988) *subsymbolic hypothesis* states that the cognitive system cannot be adequately described at the symbolic level (Chalmers, chap. 2 in this volume). However, if implementationalism were correct there wouldn't be so much interest in connectionism among cognitive scientists brought up in the symbolic tradition. More precisely, we observe that connectionist models that perform low-level tasks such as pattern recognition exhibit a number of naturally emergent properties like graceful degradation, noise tolerance, ability to generalize, and so on. However, such properties are also observed in processing at the symbolic level, but are unexplained in the symbolic paradigm. Dyer (1988) pointed out that if traditional symbolic operations were simply implemented in connectionist models many of the significant and useful cognitive properties of connectionist models would simply be lost at the symbolic level, where they in fact fulfill a great need. This would include the natural account of *mysterious processes* into which symbolic models have not provided much insight. The fact is that connectionist models seem to provide natural solutions to many of the most serious problems that arise in symbolic models.

3.3. Interactions Between Levels

The alternative to eliminativism and implementationalism is an intermediate position that accepts symbolism and connectionism as legitimate levels of description, but nevertheless does not regard them as strictly separate.

3.3.1. *Functionalism Without Multiple Realizability.* Schwartz (chap. 5 in this volume) argues that the separation of levels is much more complex than the philosophy of functionalism would suggest. In particular, he looks at Mendelian genetics (the functional level) and how it is implemented in terms of the structure

of the DNA molecule. Schwartz asks, for instance, what if the DNA molecule had three instead of two strands. This would entail a corresponding difference in Mendelian genetics; humans would seem to require three parents. Schwartz argues at the same time for the nonautonomy of levels but also for the need for high-level functionalist descriptions.

Actually, many of the same points can be made immediately and directly with respect to cognition. For instance, a symbolic theory that proposes a process involving a long sequence of serial steps, as in search with backtracking, can certainly be dismissed if it is known that humans carry out the process very quickly. This is because of the demonstrated slowness of neural architecture. An alternative proposal that divides the original process into many parallel subparts is more likely to be consistent with connectionist evidence, since massive parallelism is supported by the underlying connectionist architecture.

3.3.2. Revisionism. Revisionism is the position described by Pinker and Prince (1988) whereby symbolicism will remain its status as a functional level of description, but will nevertheless be influenced by connectionism. Thus, as they suggest, some complex symbol manipulations might be eliminated because of unanticipated computational powers of connectionism. There are probably many examples of mysterious processes for which this is likely. For instance, what is typically described as a rule application at the symbolic level is in distributed connectionist systems a process that exhibits properties of automatic generalization, best-match, noise tolerance, and so on. Knowing that these properties are accounted for in the connectionist implementation, the symbolicist might feel comfortable as accepting these as a primitive feature of rule application requiring no further analysis within the symbolic theory.

3.3.3. Limitivism. I interpret Pinker and Prince's revisionism as a very controlled view of the interrelation between the different levels of description. Smolensky (1988) advocated a position, which he called *limitivism*, that also accords symbolicism legitimacy as a level of description, but a somewhat more tenuous one. A good symbolic account is an approximation of a more accurate connectionist account that is of predictive value within certain limits. An analogy can be drawn to statistical thermodynamics and statements about temperature and flow of heat.

3.4. Hybridism

The successful cognitive scientist of the future may well be one who can shift comfortably from one level of description to the next in accordance with which provides a better conceptual grasp of the problem at hand, recognizing that two complementary perspectives are more powerful than a single, uniform one.

Kwasny and Faisal (chap. 9 in this volume) describe a natural language

parsing system that is based on symbolic work, but makes use of a connectionist component to make mysterious high-level decision about rule application. Basically, the symbolic component captures the systematicity that is captured in rules of grammar. The connectionist component, on the other hand, smooths the boundaries of what the parser can successfully process to include ungrammatical sentences or sentences with ambiguous or unknown words. It makes use of the generalization properties of the network to find a best match among the candidate symbolic rules.

Lange (chap. 10 in this volume) provides a general overview of a variety of hybrid systems and discusses a general framework for developing hybrid systems that combine distributed connectionist, localist connectionist, and symbolic marker-passing components. He considers the justification of hybrid systems to be in the development of temporary prototypes that can eventually be replaced with a single level. This replacement is desirable, he argues, because the difficulty of interfacing different levels in a natural way.

I feel that hybrid systems are a necessity and cannot in principle be replaced by single-level descriptions. Cognitive systems are complex, consisting of many component structures and processes. Looking over the range of these components we will often overextend the limits of the descriptive powers of symbolic models in two ways. First, connectionist processes may not have organized themselves into higher-level behaviors or structures that are subject to symbolic description. This may be the case, for instance, with emotional responses. Second, symbolic descriptions may simply be incomplete. This would be the case in mysterious processes like face recognition and associations between concepts.

Nevertheless, to the extent that connectionist networks have organized themselves into functionally motivated, systematic, high-level behaviors, symbolic descriptions can, and should, take over. Although parallel descriptions at the symbolic and connectionist levels may be possible, the symbolic description in this case is bound to be much simpler. Even when a full account in symbolic terms does not present itself, partial higher-level or functional descriptions of certain aspects of processes may be appropriate supplements to the connectionist description.

As the researcher moves through the various aspects of a cognitive problem moving up or down to the most useful level of description, a model will emerge that will look like a patchwork of symbolic and connectionist components. This will be a hybrid model.

4. CONCLUSION

Symbolicism and connectionism are not threats to one another. The success of one does not undermine the other. In fact, we should be happy when a problem makes sense from both perspectives. But failing this we should be free to exploit

their complementary strengths. When we realize this, the gap between the paradigms will be closed.

ACKNOWLEDGMENTS

I would like to thank Fred Adams, Ken Aizawa, John Barnden, Doug Blank, Dave Chalmers, and Stan Kwasny for reading and commenting on an earlier draft. Jackie Muehler drew the figures.

REFERENCES

- Derthick, M. (1990). [Review of *Connections and symbols*]. *Artificial Intelligence*, 43, 251–265.
- Dinsmore, J. (1991). *Partitioned representations: A study in mental representation, language understanding and linguistic structure*. Dordrecht: Kluwer.
- Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine*. New York: Free Press.
- Dyer, M. (1988). The promise and problems of connectionism. *Behavioral and Brain Sciences*, 11(1), 32–33.
- Fodor, J., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 3–72). Cambridge, MA: Bradford/MIT Press.
- Lachter, J., & Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. In S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 195–247). Cambridge, MA: Bradford/MIT Press.
- Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.
- McCarthy, J. (1988). Epistemological challenges for connectionism. *Behavioral and Brain Sciences*, 11(1), 44.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (1983). *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (1986). *Machine learning*. (Vol. II). Palo Alto, CA: Tioga.
- Minsky, M., & Papert, S. (1988). *Perceptions, (expanded ed.)*. Cambridge, MA: MIT Press.
- Newell, A. (1982). The knowledge level. *Artificial Intelligence*, 18(1), 87–127.
- Newell, A., & Simon, H. (1976). Computer science as an empirical inquiry: Symbols and search. *Communications of the ACM*, 19, 113–126.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 73–194). Cambridge, MA: Bradford/MIT Press.
- Putnam, H. (1973). Reductionism and the nature of psychology. *Cognition*, 2, 131–146.
- Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, MA: Bradford/MIT Press.
- Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 227–270). Cambridge, MA: MIT Press.
- Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 2, pp. 216–217). Cambridge, MA: MIT Press.

- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. Rumelhart, J. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 1, pp. 7-57). Cambridge, MA: MIT Press.
- Smolensky, P. (1988). On the proper treatment of connectionism. *The Behavioral and Brain Sciences*, 11(1), 1-74.