CONNECTIONISM AND INTENTIONALITY

WILLIAM BECHTEL

DEPARTMENT OF PHILOSOPHY
GEORGIA STATE UNIVERSITY

ABSTRACT

Connectionism offers greater promise than symbolic approaches to cognitive science for explaining the *intentionality* of mental states, that is, their ability to be *about* other phenomena. In symbolic cognitive science symbols are essentially arbitrary so that there is nothing that intrinsically relates them to their referents. The causal process of transduction is inadequate to explain how mental states acquire intentionality, in part because it is incapable of taking into account the contextual character of mental states. In contrast, representations employed in connectionist models can be much more closely connected to the things they represent. The ability to produce these representations in response to external stimuli is controlled by weights which the system acquires through a learning process. In multi-layer systems the particular representations that are formed are also determined by processes internal to the system as it learns to produce the overall desired output. Finally, the representations produced are sensitive both to contextual variations in the objects being represented and in the system doing the representing. These features suggest that connectionism offers significant resources for explaining how representations are *about* other phenomena and so possess *intentionality*.

THE PROBLEM OF INTENTIONALITY

Explaining the intentionality of mental states, the fact that they are *about* phenomena that are generally situated outside of the cognitive agent, has been a central concern in the philosophy of cognitive science. The challenge is to explain what it is in virtue of which a mental state is about a particular phenomenon and so has a particular content. One of the factors that makes this challenge difficult was identified by Brentano (1874/1973). He noted that a mental state such as a belief seems to involve a relation between the believer and external phenomena, but that this relation is unlike ordinary relations. If Sam believes that Sarah is a neurologist, Sam's state of mind seems to stand in a relation to Sarah. Normally, for a relation to exist both relata must exist. Yet, Sam could well have this belief and Sarah not exist. His mental state is still *about* Sarah, and not anyone or anything else. Thus, intentionality cannot be handled simply in terms of relations.

The problem of explaining intentionality is a serious one for symbolic cognitive science since it takes seriously an aspect of the ordinary logic of sentences about mental states. These sentences typically have the form of propositional attitudes, in which the verb (e.g., *believes*) represents a relation between a person and a *proposition*. The proposition

then becomes the *bearer* of the intentionality since it is what represents the possible or actual condition in the world to which the person's belief is directed. To explain such states, many practitioners of symbolic cognitive have assumed that there are symbols in the mind corresponding to propositions and that the mind manipulates these symbols via procedures much like those posited in formal logic. The use of symbols as bearers of intentionality helps solve the problem of how mental states can be about non-existing entities, since there are a variety of procedures through which we can imagine formal symbols being introduced which do not correspond to actual entities. The challenge, however, is to explain how symbols, whether or not they do refer to real things, have the specific representational content they have. I will briefly examine why this problem is a difficult one for symbolic cognitive science. My main endeavor will then be to explore how connectionist approaches offer promise in explaining this aspect of intentionality.

Intentionality and Symbolic Cognitive Science

This is not the place to review in detail the difficulties that arise in explaining the intentionality of formal symbols (see Bechtel, 1988). Rather, I will try to capture some of the problems informally so as to set up the contrast with connectionist approaches. The main problem the symbolic approach faces in explaining intentionality is that primitive symbols are treated as *atomic* and *arbitrary*. As a result, there is nothing about the symbol itself that determines its referent. The question then arises as to what it is that determines the referent of a symbol. What makes the mental symbol for Sarah refer to Sarah? The most plausible approach is to treat the symbol for Sarah as having a particular referent because of the way it is employed by the cognitive system. The set of symbols is manipulated by formal rules in a manner that is appropriate to the referential function of those symbols. As a formal system, the cognitive system is construed as a *syntactic engine*. The model here is the manner in which formal proof procedures in logic are *truth* preserving because they mirror the relations between objects in the world. *Truth* is a semantic property, relating a proposition to the situation in the world that satisfies it. Proof procedures do not utilize semantic information but provide a formal means of manipulating symbols that respects the semantic property of truth. Similarly, the formal operations of the cognitive system's syntactic engine do not rely on the referents of these symbols but provide a means for properly manipulating symbols so as to facilitate the system's negotiation with these objects. Since reference is a semantic relation, the syntactic engine can be seen as simulating a *semantic engine* (Dennett, 1981).

According to the view just characterized, there is nothing about the formal symbols that determines their semantic content. This can be appreciated by the simple thought experiment in which a formal system, a computer program, that is satisfactorily performing one task is employed to perform another task and does it equally well. There is nothing about the formal symbols in the program that makes them more about the objects involved in the first task than those encountered in the second task. We, the users of the program, must supply the interpretation. This point is closely related to one Searle (1980) derives from his famous Chinese room argument in which he pictures himself manipulating symbols in a purely formal manner using a set of rules. He does

this in such a manner as to carry on a conversation in Chinese without understanding a single word of Chinese or knowing that he is conversing in Chinese. The Chinese characters could have quite different semantics and that would not alter Searle's behavior. Since humans do know what their mental states are about, Searle objects that the formal symbol approach totally fails to capture the *intrinsic intentionality* of mental states.

While some theorists have been satisfied with the view that all there is to intentionality is accounted for whenever a syntactic engine simulates a semantic engine, many others have agreed with Searle that we need to explain how humans, at least, are real semantic engines. We must explain, they maintain, how our mental state have determinant contents and should not be subject to whatever reinterpretation an external party chooses to employ. However, few have been satisfied with Searle's own explanation of intrinsic intentionality, which appeals to the biological character of mental states. An alternative perspective, suggested by Dreyfus and Dreyfus (1986), is to focus the difficulty on the *context-free* character of formal symbols. In characterizing formal symbols as context-free we are noting that how symbols are processed depends only on what is formal represented and no other aspects of the environment. Dreyfus and Dreyfus attributed the reliance on context-free formal symbols to traditional philosophy, which has provided much of the theoretical framework for cognitive science:

> According to Heidegger, traditional philosophy is defined from the start by its focusing on facts in the world while "passing over" the world as such. *This means that philosophy has from the start systematically ignored or distorted the everyday context of human activity.* The branch of the philosophical tradition that descends from Socrates through Plato, Descartes, Leibniz, and Kant to conventional AI takes it for granted, in addition, that understanding a domain consists in having a *theory* of that domain. *A theory formulates the* relationships among objective, *context-free elements* (simples, primitives, features, attributes, factors, data points, cues, etc.) in terms of abstract principles (covering laws, rules, programs, etc.) (pp. 24-25).

For a formal symbolic system, the system's total knowledge about context must be provided in terms of formal symbols, that is, in other explicit representations in the system. The hope has been that we could build in enough explicit representations to enable the system to deal adequately with all contexts that arise in the real world, but this is precisely what Hubert Dreyfus has long been questioning (see Dreyfus, 1979). The problem for a formal symbol system is that there does not seem to be any other way to bring context into play.

The main alternative to trying to account for the intentionality of symbols in terms of formal relations between symbols has been to analyze their meaning or intentionality in terms of their relations to the objects that they represent. One possibility that has been pursued has been to treat the causal mechanisms that produce the symbols in us as the source of intentionality. Dretske (1981), for example, characterizes the causal relation between the object in the world and the symbol in the head in terms of the *information* that is transmitted and then tries to explain intentionality in terms of how the symbol bears information *about* the object. When a symbol is activated without being caused by its referent, it is still about the object which would normally cause its activation. This proposal has been challenged from a number of perspectives. In particular, it has been

argued that such causal relations are inadequate to account for the possibility of error or misrepresentation (e.g., the possibility of representing non-existent objects), which, as we have already noted, is an important characteristic of intentional states (see Churchland & Churchland, 1983 and Fodor, 1984).

An additional objection to treating the causal relation between referent and symbol as the basis for intentionality is that such an approach is not able to accommodate the role of contextual factors such as those the Dreyfuses have emphasized. When we use representations intentionally, the particular referent that is intended for the system may vary with the context. This problem is readily seen when we consider the representational function of words in a natural language. Barsalou (1987) has shown surprising variability in people's prototypicality ratings of exemplars of concepts over time, suggesting that the representational function of words as well as their internal representations change with context. The problem for capturing this in a symbolic account is that symbols are fixed entities. Moreover, the relata in the causal link between an object and a symbol will have to involve something like the typical entity that generates the symbol in the cognizer. The causal theory cannot explain how on the different occasions when a symbol is used, there may be significant variation among intended referents. This variability in intended referents is an aspect of intentionality that cannot be accounted for either in terms of formal relations between symbols or in terms of the typical causal ancestor of the symbol. Contextual sensitivity is, however, something connectionist systems are more adept at dealing with. Hence, there is motivation to explore the potential of connectionism in accounting for intentionality.

A Connectionist Perspective on Intentionality

Part of the problem with the symbolic approach is that it limits contact with the world to a process of transduction through which a sensory input is transformed into a symbol. Some of the potential of connectionism in accounting for intentionality stems from the alternative perspective it provides on the transduction process. Sensory input will be provided to the network by activating certain nodes in the network. These nodes will then cause other nodes to activate. The initial activation process culminates in the activation of the units constituting the representation. This processing within the system is of a piece with the causal transmission of signals in the external world and so provides the potential for direct contact of the representational states of the system with their referents. Despite the fact that there is a direct continuity in the sort of processing involved, this process may still seem to be very like the kind of transduction envisaged in a symbolic model: the sensory input causes a representation to be activated in the system. But there are several crucial differences between this connectionist process and the type of transduction required in symbolic systems that render the connectionist approach better suited for explaining intentionality.

One of the ways in which connectionist models have an advantage over symbolic accounts is that in at least one respect connectionist representations will not be arbitrary in the way that symbolic representations are. This is a result of the fact that connectionist systems have the capacity to learn how to generate their representations and also

what representations to employ. At the level of basic representations, only the first of these capacities is generally employed in current connectionist systems. In simple, two-level, feedforward networks trained through procedures such as the least mean squares learning algorithm, for example, the weights required to produce the output representation are learned. Through the learning process the network selects how to attend to features of the input. Only some input units are relevant for determining the weights of particular output units, and the weights from these input units adjust accordingly. Thus, the process of generating the representation involves an adaptation of the representational system to the external referent. Here is one initial respect in which the representations developed in connectionist systems are closely tied to that which they are supposed to represent. This linkage makes the relation between representation and represented somewhat less arbitrary than in symbolic systems.

The representations that simple networks learn, however, are chosen by the researcher. Since any input can be paired with any output, there is still a strong sense in which the representations are arbitrarily related to what they represent. To reduce this sense of arbitrariness we need to consider systems which possess the ability to create their own internal representations. Something like this capacity is found in multi-level feedforward networks trained through processes like backpropagation. The hidden units in such systems develop specific response characteristics in the course of training the output units to produce the desired patterns of activation. Sometimes it is possible to determine, through detailed analysis of when the hidden units become active, what representational function is performed by each of these units. For example, Hinton (1987) designed a network to learn information about relations in two family trees. The input units specified one person and a kin relationship, and the output units were to identify the person standing in that relationship. Between the input and output units were three layers of hidden units. The input and output units were coded in a localist fashion, one person or relationship per unit. However, because the number of hidden units was much smaller than the number of input units, the network was forced to find distributed representations for the input and output. For example, the twenty four input units encoding the possible individuals fed into a set of six hidden units. Through the course of learning via back propagation, the network had to find a way to represent all the information about these individuals that it required in order to determine the correct output person. The network developed a representational system that identified persons in terms of their tree (British or Italian), their generation, and the branch of the tree from which they came. The important point for our purposes is that in Hinton's simulation the network developed its own distributed representation in the course of adjusting connection strengths so as to minimize its error in solving the task for which it required the information. Since the network is determining these representations on the hidden units, they are far less arbitrary than do symbols in a symbolic system.

In existing networks the internal representations that are constructed are grounded in an already existing representation chosen by the researcher. Thus, in Hinton's simulations, one set of input units encodes the name of the person whose relative is being sought, and the other set of input units encodes the relations. Hence, the representations that are learned (i.e., the activation pattern over the hidden units) are comparable to higher

order concepts or complex symbols that might be acquired in symbolic systems. (They are not fully comparable to these since they are far more sensitive to small variations in input information than are symbolic representations. For example, while one unit encodes whether or not the input person is English, it generates higher activation levels for some English persons than others.) This limitation, however, results from the fact that these systems do not use sensory stimuli directly as inputs. If one were to train a system that took as inputs the outputs of sensory receptors that directly picked up information from the environment, then the responses of hidden units could be thought of as defining the system's most basic categorization of inputs and hence as providing the system with its most basic representations.

The crucial point to be emphasized is that representations on hidden units result from the system's attempt to accommodate to its environment. They cease to be states which could have been causally connected to any sensory input and, hence, arbitrary as far as the operation of the system was concerned. Since these representations constitute a learned response of the system to a given set of inputs that the system then uses in order to respond in the desired way to those inputs, these representations are naturally seen as being *about* the entities supplying the input. (The tightness of this connection is evident in the fact that in order for researchers to analyze the operation of hidden units, they must try to identify what input patterns will in fact generate the response of particular hidden units.) A connectionist system such as I have described is thus able to develop representations in much the way Dreyfus and Dreyfus portray human systems as learning:

> By playing with all sorts of liquids and solids every day for years, a child may simply learn to discriminate prototypical cases of solids, liquids, and so on and learn typical skilled responses to their typical behavior in typical circumstances (Dreyfus and Dreyfus, 1987, p. 33).

The contrast between this process and the way interpretations are generally assigned to symbols in symbolic systems is clear. There are not separate processes of learning to use a symbolic representation and learning how to assign an interpretation to it. The connectionist representation is developed as part of the system's adaptation to its environment.

One of the failures Dreyfus and Dreyfus claimed befell symbolic representational systems was their lack of context sensitivity. The responses that connectionist systems make to their environments are quite sensitive to the particular stimuli they receive as well as to other processing that is occurring in the networks. Particularly when units can take on continuous activations, there is enormous variability in the responsiveness of individual units. As a result, the system does not need to have discrete symbols by which it can represent each variation in context. For example, consider a case in which a representation is produced not from an input, but from activity elsewhere in the network that causes it to activate a pattern much like it would for a particular type of input such as a ball. On one occasion this activity may result in a pattern more like that typically generated by a baseball, while on another occasion it might result in a pattern more like that typically produced by a basketball. This variation is then available to enable the system to adjust its response in light of differences in input circumstances and

internal conditions. This ability to vary representations in appropriate manners may not be a unmitigated benefit since it will be necessary to ensure that the ultimate response of the system is appropriate to the context and is not a bizarre one. Thus, the responsiveness of the system to the representations must itself be tuned to the variability in the representation itself. But at least it is possible for such a connectionist system to represent objects differently depending upon context and so these systems are not restricted, as are symbolic systems, to representing context in yet other arbitrary symbols.

The connectionist approach to modeling cognition thus offers promise in explaining the *aboutness* or *intentionality* of mental states. Representational states, especially those of hidden units, constitute the system's own learned response to inputs. Since they constitute the system's adaptation to the input, there is a clear respect in which they are *about* those inputs. They are about the situations to which they are responses in much the way biological adaptations are adapted to situations like those which figured in the process of their selection. The fact that these representations are also sensitive to context, both external and internal to the system, enhances the plausibility of this claim that the representations are representations of particular states. The connectionist approach thus makes a start on explaining the *aboutness* of representations. Unfortunately, there is more to be done to explain intentionality. We must also explain how mental states can represent things that do not exist. This seemed relatively easy to do in symbolic systems, since we could simply incorporate a symbol to stand in for the non-existent object. Yet we could not explain why the arbitrary symbol had the referent it did. A detailed explanation of how connectionist systems could make reference to non-existing objects is beyond this paper. But the outlines of how this is possible can be sketched. In interactive networks, activations can be brought about by activity in the network itself, and not just from external inputs. It is conceivable that activation patterns could be induced that do not correspond to anything normally caused by input patterns. These would be representations of non-existent objects. We know they are *about* these objects, and not others, because they are the representations that would be produced if the system ever did confront such an object. Thus, if a representational pattern was created by internal processes in the system which would be produced by the system encountering a unicorn, then it would be a representation of a unicorn, not of Santa Claus. The network's response to the production of these states can be viewed as its further thinking about the non-existent objects.

The proposals advanced here are simply intended to show the promise of connectionism in helping us understand the intentionality of mental states. They do not show that connectionist accounts will be successful or that symbolic analyses cannot invoke similar strategies in order to explain intentionality themselves. (The causal analysis of the intentionality of symbolic states most nearly parallels the account proposed here and could conceivably employ some of the strategies outlined here to flesh out that account. What distinguishes the two accounts is that the causal account does not treat the representation as an adaptation on the part of the cognitive system, and so does not as clearly overcome the problem that the symbolic representation remains rather arbitrary and so not intrinsically linked to its referent. It is simply the symbolic state that happened to be caused by the sensory input.) There are challenges to be faced in

devising connectionist networks that will have the right semantics to model cognition. For example, just designing a system that has a context sensitive representation of an external referent does not ensure that it can use this representation appropriately in solving other problems. But perhaps there is even a virtue here in that this constitutes an empirical research problem about the intentional representations in a network, and not simply a problem to be solved by *a priori* philosophical speculation. Since so little has been achieved in the attempt to explain the *aboutness* or intentionality of mental state, the fact that connectionism offers a plausible promissory note is one reason to take it seriously.

REFERENCES

Barsalou, L. (1987). The instability of graded structures: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101-140). Cambridge, England: Cambridge University Press.

Bechtel, W. (1988). *Philosophy of mind. An overview for cognitive science.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Brentano, F. (1874/1973). *Psychology from an empirical standpoint* (A.C. Pancurello, D. B. Terrell, & L. L. McAlister, Trans.). New York: Humanities.

Churchland, P. S. & Churchland, P. M. (1983). Stalking the wild epistemic engine. *Nous, 17,* 44-52.

Dennett, D. C. (1981). Three kinds of intentional psychology. In R. Healey (Ed.), *Reduction, time and reality* (pp. 37-61). Cambridge: Cambridge University Press.

Dretske, F. I. (1983). *Knowledge and the flow of information.* Cambridge, MA: MIT Press/Bradford Books.

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence.* (2nd edition). New York: Harper & Row.

Dreyfus, H. L. and Dreyfus, S. E. (1987). *Mind over machine. The power of human intuition and expertise in the era of the computer.* New York: The Free Press.

Fodor, J. A. (1984). Semantics, Wisconsin Style. *Synthese, 59,* 231-250.

Hinton, G. (1986). Learning distributed representations of concepts. *Proceedings of the eighth annual conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences, 3,* 417-424.