August 31, 1994

# Consciousness: Perspectives from Symbolic and Connectionist AI

William Bechtel
Program in Philosophy, Neuroscience, and Psychology
Department of Philosophy
Washington University in St. Louis

## 1. Computational Models of Consciousness

For many people, consciousness is one of the defining characteristics of mental states. Thus, it is quite surprising that consciousness has, until quite recently, had very little role to play in the cognitive sciences. Three very popular multi-authored overviews of cognitive science, Stillings et al. [33], Posner [26], and Osherson et al. [25], do not have a single reference to consciousness in their indexes. One reason this seems surprising is that the cognitive revolution was, in large part, a repudiation of behaviorism's proscription against appealing to inner mental events. When researchers turned to consider inner mental events, one might have expected them to turn to conscious states of mind. But in fact the appeals were to postulated inner events of information processing. The model for many researchers of such information processing is the kind of transformation of symbolic structures that occurs in a digital computer. By positing procedures for performing such transformation of incoming information, cognitive scientists could hope to account for the performance of cognitive agents. Artificial intelligence, as a central discipline of cognitive science, has seemed to impose some of the toughest tests on the ability to develop information processing accounts of cognition: it required its researchers to develop running programs whose performance one could compare with that of our usual standard for cognitive agents, human beings. As a result of this focus, for AI researchers to succeed, at least in their primary task, they did not need to attend to consciousness; they simply had to design programs that behaved appropriately (no small task in itself!).

This is not to say that conscious was totally ignored by artificial intelligence researchers. Some aspect of our conscious experience seemed critical to the success of any information processing model. For example, conscious agents exhibit selective attention. Some information received through their senses is attended to; much else is ignored. What is attended to varies with the task being performed: when one is busy with a conceptual problem, one may not hear the sounds of one's air conditioner going on and off, but if one is engaged in repairing the controls on the air conditioning system, one may be very attentive to these sounds. In order for AI systems to function in the real world (especially if they are embodied in robots) it is necessary to control attention, and a fair amount of AI research has been devoted to this topic.

Moreover, conscious thought seems to be linear: we are not aware of having multiple thoughts at once, but rather of one thought succeeding another. Typically other processing is occurring at the same time, but it is not brought into the linear sequence of consciousness. This is suggestive of the role of an executive in some AI systems, which is responsible for coordinating and directing the flow of information needed for action. Johnson-Laird [19], who argues for the need for much of the computation underlying behavior to proceed in parallel, proposes that consciousness arises with a high level processing system that coordinates lower level processes. Finally, conscious states are states people have access to, can report on to others, and can rely on in conducting their own actions. Capturing an aspect of this has turned out to be particularly important in a variety of computer programs, especially those we rely on to make or advise about decisions. We want to query them about their decisions or recommendations so as to evaluate whether they were reasonable. But Johnson-Laird points out that humans go much further: we can use information about our own states to guide our actions. He therefore proposes that what is needed for cognitive systems to be conscious is that they possess a recursive ability to model (albeit partially and incompletely) what is going on in themselves, and to use such models in controlling future activity (mental and physical).

The strategy employed by Johnson-Laird (himself a psychologist, albeit one with a strong computational perspective) and those AI researchers who have taken consciousness seriously has been to focus on the functions of conscious: how is it that our information processing is influenced by being conscious. They then try to capture these functional elements in either schematic designs or actual AI systems. Until recently most theorists who took up the question of how an AI system might exhibit consciousness operated within the framework of symbolic AI. A symbolic AI system is one in which explicit symbol structures within the computer represent pieces of information and the system employs rules to transform these rules. Symbolic AI has been quite successful in modeling a variety of cognitive activities but has also exhibited some clear limitations. Those frustrated with the problems of working in traditional AI have recently brought about a renaissance in another form of AI modeling, one that is not grounded on performing formal operations on complex representations. Drawing their inspiration from the basic conception of how a brain works, these theorists develop models using processing units which can take on activations and, as a result, excite or inhibit other units. Systems of such processing units are referred to as *connectionist* or *neural network* systems. (For introductions to connectionist modeling, see Rumelhart, McClelland, and the PDP Research Group [29] and Bechtel and Abrahamsen [2].) As we shall see below, connectionism provides quite different resources and challenges for someone seeking to explain consciousness. But for some connectionists, the enterprise is much like that for practitioners of more traditional AI: identify functional features of consciousness and show how they might be explicated using a connectionist model.

One connectionist who has approached consciousness in this way is Paul Churchland [5]. Churchland identifies the following features of consciousness:

1) Consciousness involves *short-term* memory.

2) Consciousness is *independent of sensory inputs*.

3) Consciousness displays *steerable attention*.

4) Consciousness has the capacity for *alternative interpretations* of complex or ambiguous data.

5) Consciousness *disappears in deep sleep*.

6) Consciousness *reappears in dreaming*, at least in muted or disjointed form.

7) Consciousness harbors the contents of the several basic sensory modalities within a *single unified experience*.

His strategy is then to show that connectionist models of a particular kind have the resources to explain these features of consciousness. The sort of model he appeals to is a recurrent networks, a network in which activation largely flows forward from input units to output units through one or more intermediate layers, but in which there is feedback from units later in the processing stream to those earlier [13]. One thing such feedback provides is a way to make processing of inputs supplied later sensitive to what happened in the processing of earlier inputs. That is, recurrent networks exhibit the first feature of consciousness, a kind of short-term memory. Second, these networks can continue to processes recurrent stimulation even when no new input is provided, and thus are independent of sensory inputs. Third, recurrent pathways provide a means of influencing the processing of a new input, in particular steering the network to attend and respond to certain features of the input rather than others. Fourth, recurrent pathways can cause the network to respond to the same input in multiple ways, thus exhibiting the capacity for generating multiple interpretations.

To explain other features of consciousness Churchland combines the appeal to artificial neural networks with information about the brain, in particular Llinas' discovery of 40 herz oscillations in the neural activity of the cortex, which seem to be due to recurrent neural pathways from the intralaminar nucleus, whose neurons have an intrinsic tendency to oscillate at 40 herz. Overlaying this oscillation are a variety of specific activation patterns which seem to constitute the brain's response to particular stimulations. This amplitude of this oscillatory pattern is greatly diminished in deep sleep when the neurons in the intralaminar nucleus are inactive. But it reappears in REM sleep, along with the overlaying activation patterns, which now, however, are not correlated with external inputs [20]. Churchland proposes that the recurrent pathways from the intralaminar nucleus can account for both the disappearance of consciousness in deep sleep and its muted or disjointed reappearance in REM sleep. Finally, Churchland proposes that the recurrent processing through the intralaminar nucleus explains the unified character of consciousness--all processing must go through the bottleneck of the intralaminar nucleus, and this imposes unification on the processing.

## 2. Perceived Problems with Computational Approaches

As the previous section suggests, there is reason to be optimistic that computational models can capture many of the functional features of consciousness. To many critics, though, computational attempts to model consciousness, whether in more classical symbolic systems or in connectionist systems, miss the point. These critics contend that what is special about consciousness is not its functional aspects, but its qualitative character (see Chalmers [6]). The worry about the qualitative character of consciousness was presented crisply by Thomas Nagel

[23] when he posed the question "What is it like to be a bat?" He contended that although we might come to understand the neurophysiology of bats and work out in detail the manner in which they process sonar information, we will never understand what it is like to experience the world as a bat does. This information simply is not there in the physical or functional information about bats. One must be a bat to appreciate how bats experience the world. And since it must be experienced to be appreciated, it cannot be explained in terms of the neurophysiology of the bat. What goes for bats also go for us: the qualitative character of our conscious states cannot be understood by knowing the physical processes occurring within us. If nothing about the neurophysiology of a bat or us could account for the way a bat or we experience the world, the consequences for AI, symbolic or connectionist, are rather dire: an important aspect of the cognition of humans and other animals may simply lie outside of the scope of AI (that is, outside of the scope of any computational model).

To undercut the potential of computational models to explain consciousness one need not adopt even such a strong position as Nagel's. It is common to differentiate the functional and physical aspects of events within physical systems. Functional features of a system have to do with the operations components of the system perform and the way these are coordinated. The physical features concern the particular physical devices that perform those operations. This distinction corresponds to the distinction between hardware and software in the case of computers, but can be generalized to other complex systems. For example, one can differentiate the catalyzed reactions that are performed in different metabolic processes from the specific enzymes that catalyze the reactions. What Nagel and others who others who share his concern claim is that the qualitative character of mental events is not due to either the functional or physical character of cognitive systems but is in some way beyond the physical and functional dimension. However, to argue that AI is incapable of explaining consciousness, it is sufficient to argue that conscious is not a functional property but depends on the physical realization of our cognitive system. That is, one only needs to argue that consciousness can only be explained by details of brain physiology, and not the functional properties of the brain (its procedures for processing information). Consciousness would then be *program resistance* in contrast with the program receptive properties of sapience (Gunderson, [16]). Such a view has been advocated by John Searle [30] in what many have regarded as one of the most serious criticisms of AI.

Searle focuses his argument on intentionality, a feature which Brentano [4] identified as separating mental systems from physical ones. Intentionality is the characteristic that mental states enjoy of having content, or being *about* something. In contemplating a rose, one's thoughts are about the rose. In developing a theory of consciousness, one's thoughts are about consciousness. Mental states seem, in some way, to be linked to their objects. But this link is not, and cannot be, an ordinary physical relation. We can have thoughts about things that do not exist, such as Ponce de Leon's fountain of youth. And our thoughts of the fountain of youth are not the same as our thoughts of other nonexistent objects. This is not an idle worry, since many of our thoughts are directed at possible future states of affairs, many of which will never be realized. For Brentano, the fact that mental states seemed relation-like, but not actually relations, was an argument for dualism; for him, our mental states behaved differently than any phenomena in the physical universe. (One common response to Brentano's problem is to hold that in mental states we are not related to objects in the world, but to mental representations of those objects. There is nothing problematic, it seems, in having mental representations of nonexistent objects. But this fails to solve the problem. In order for mental states to account for our ability to function in the real world, for example, to allow us to make plans that lead to real actions, our mental states must be *about* the objects of the world, not just our representations. If one accepts the need for mental representations, the problem of intentional can be stated as the problem of how our mental representations relate to objects in the world. Now the problem that intentionality seems to be relation-like but not a relation arises again.)

Searle's contention is that real intentionality, which he calls *intrinsic intentionality* cannot be exhibited by computers because having intentionality is not a matter of having the right program. He argues for this claim by means of a thought experiment. Assume that we have a program that proposes to account for how a person engages in a conversation in Chinese and that a computer running this program is behaviorally indistinguishable from an actual Chinese speaker. The program takes in Chinese text, performs some formal operations on it (which might involve writing down other characters, storing them, retrieving them, etc.). Now, to show that the program does not account for intentionality, allow Searle to execute the program. He will be locked in a room with English directions that specify the steps in the program. Some Chinese text is slipped under the door, and Searle executes the operations specified in the program. After a time the operations specify that he should slide some new text back under the door, and he obliges. This process is repeated and Chinese speakers outside the room are satisfied that they are engaging in conversation with a fellow Chinese speaker. The only problem, Searle contends, is that he does not understand a word of Chinese (he did not even know he was working on Chinese text) and hence had no idea that he was carrying on a conversation or what it was about. Since he did not know the content of the conversation, he did not exhibit intentionality concerning the Chinese conversation. Nor, Searle contends, does a computer which is running the program.

Searle anticipated several obvious lines of response to this argument, one of the most compelling of which he labeled the *systems* reply. This response contends that when Searle is in the Chinese room he does not understand the Chinese text. But that is because the program and some of the intermediate states of processing are external to Searle, stored in the program which Searle consults and in the intermediate representations he stores on paper. Searle, however, revises the scenario so as to answer this. He proposes that he memorize the program and carry out all of the operations in his head. Now everything is inside Searle. Yet, he contends he still does not understand the Chinese conversation in which he is engaged. A real Chinese speaker, on the other hand, knows what he or she is talking about. Thus, Searle's simulation of the Chinese speaker fails to capture the intentionality of the Chinese speaker's mental states. It has failed to capture the Chinese speaker's conscious awareness of what he or she is talking about. Since any computer simulation would possess no more than Searle possesses, it too would fail to exhibit such conscious awareness of what its conversations were about. (The advantage of the thought experiment over any computer implementation, Searle contends, is that while we cannot get a reliable report from the computer as to what it is like to engage in the conversation, Searle, who already knows what it is like to engage in a real conversation in which he knows what he is talking about, can tell us that in the thought experiment he would have no such knowledge. Indeed, the computer might put out a response that it does know what it is talking about, but if the program Searle was executing directed him to produce such an output in English, he would be able to testify that it was an erroneous statement.)

Searle's diagnosis of the problem is that the computer simulation fails to capture the important causal element in conscious human performance. Intentional states have real causal roles to play in us, and they play this causal role in virtue of their intentionality. Computational models, being purely formal, fail to capture this causal role. The only way to capture this causal role is to turn to something that has true causal powers: the biological states in our brains. Just as plants generate oxygen from carbon dioxide as a result of possessing chlorophyll, our brains produce behavior because they have intentional states within them. This intentionality is, for Searle, an emergent property of the neural structures in our brains. Other natural phenomena might generate intentional states, but a purely formal analysis, such as is provided in an AI program, cannot realize the causal agency found in real brains. Thus, intrinsic intentionality is conceptually beyond the capacities of AI programmers. (It seems quite reasonable to object to Searle's appeal to biology by noting that functional explanations are not limited to psychology but appear in physiology, biochemistry, etc. Moreover, when we explain a feature of some system, we do so by finding a level of organization at which functional properties of the system's components can account for the features of the system. Thus, if Searle's rejection of functional or computational accounts of intentionality were correct, they would also rule out biological explanations of intentionality.)

More recently, Searle [31] has raised the stakes of his position with respect to consciousness. One might have responded to Searle's earlier position by agreeing that indeed AI simulations could not capture the intrinsic intentionality or conscious awareness of mental content exhibited in our conscious mental states, but that at least they could account for a great deal of mental life. A common assumption of many cognitive scientists is that most mental processes are non-conscious and these should be within the range of computational explanations. We can assign representational content to underlying mental states and explain behavior in terms of processes operating over these contentful states. But Searle contends that in order for a state to be representational it must possess what he refers to as its *aspectual* shape: a particular perspective on the object or event represented. For example, a representation of Clinton may represent him as a jogger, not as President. Since the referent of the representation remains the same for both the jogger and President representation, this aspectual shape is not captured by physical relations to Clinton but can only be identified and employed by us when we are conscious of it. (This claim seems problematic at best. Computational theorists would seek to capture the particular aspectual shape of a representation in terms of relations to other representations, not just to possible referents in the world.) Hence it makes no sense to attribute it to states that are inherently non-conscious and so it makes no sense to posit inherently non-conscious intentional states and to try to explain behavior in terms of them. When behavior arises without consciousness, the explanation for it must be purely neural, and make no reference to intentional mental representations. If Searle is right, therefore, not only do conscious mental states lie outside the scope of AI, but since only conscious or potentially conscious mental states can figure in real psychological explanations, so does any psychological phenomenon.

## 3. Strategies for Capturing Consciousness in Computational Systems

For the remainder of this paper I will take Searle's arguments to pose a serious challenge for anyone seeking to provide a computational account of consciousness and sketch ways in which computationalist might counter his arguments. The first step is to overcome the assumption that consciousness is a single, holistic phenomenon that is either present or not. This sense lies behind not only Searle's arguments but many attempts to argue that consciousness is a special property that we cannot hope to explain. It lies behind the claim that what differentiates AI systems, robots, or zombies from us is that in the first three cases there is no one home, while in our case there is a someone who is home. If conscious were such a phenomenon that it is either present or absent, then it is difficult to show how an underlying physical system could account for it. On the other hand, if the phenomenon of consciousness can be decomposed, and some aspects of it realized while others are not, then one has a basis for identifying underlying mechanisms whose presence or absence correlates with the features of consciousness that are present or absent (see Bechtel and Richardson [3] for an account of the development of scientific explanations via the strategies of decomposition and localization). Dennett [11] offers a sustained attack on the idea of consciousness as a unified or atomistic phenomenon, while the strategy of trying to show how different components of consciousness might be explained by different physical or functional processes and how we might use defects in consciousness to identify responsible mechanisms is well described in the work of Flanagan ([14] and this volume).

Inspired by Searle, I shall focus on three features that seem particularly prominent when we think of our conscious mental states. First is the intentionality of mental states. Intentionality is often viewed as principally a feature of cognitive states such those characterized in terms of propositional attitudes like belief and desire. But it is also characteristic of more purely phenomenal states such as perceiving colors or hearing tones: we see a red circle or hear an off-pitch tone. Second, we are aware of these states and their content; we have what appears to be reliable first person privileged access to their contents. Third, there is a distinctive qualitative character to each of these states. This is especially true of perceptual and imagistic states: seeing red feels qualitatively different from seeing blue. But it is also true of propositional attitude states: thinking today is a holiday feels qualitatively different than thinking that consciousness can be explained computationally. The question is whether any or all of these features of conscious states can be realized in an AI system.

In one respect, intentionality seems to be obviously present in symbolic AI systems. The symbols employed in these systems, over which the formal operations specified in the program are then performed, are understood to be representations. It is insofar as these symbols are interpreted as representing features of the world that the systems can be thought of as having knowledge of the world and figuring out solutions to the problems presented to the systems. But this intentionality is illusory. There is nothing about the representations in symbolic computer systems that makes them have specific content. This is easily seen by recognizing that a program designed to perform one task can be put to work to perform another task as long as the formal operations specified in the program are correct. A program to play tick-tac-toe, for instance, is not intrinsically a tick-tack-toe program. The program will do an equally good job of Simon's game of number scrabble, which consists of laying cards with values 1 through 9 on a table face up, and allowing two players to alternately choose one until one player has a set of three cards that add up to 15 ([32], p. 76). In the case of traditional symbolic AI programs, all of the work is done by the set of formal operations in the system, operations defined in terms of the formal composition of the symbols themselves (their syntax). If these operations are appropriate, then the program will provide the right responses when its symbols are interpreted as referring to entities in the real world. In Dennett's [9] characterization, a symbolic AI device is a syntactic engine which performs as a semantic engine. But all of the intentionality is supplied by the human being who interprets the AI system. This is what Searle refers to as *as if* intentionality: the system is behaving as if it were an intentional system, but it is not really such a system. (Searle holds much the same position with respect to the words of a natural language--in themselves they do not possess meaning or intentionality, but only insofar as they are interpreted as having particular meaning by human beings.)

One diagnosis of why traditional symbolic systems do not really possess intentionality is that there is not the right kind of connection between the representations within the system and what they represent. Some philosophers of language have attempted to explain the specific meanings of words in natural language in terms of historical links to occasions in which they were used to refer to particular objects. This approach encounters a number of problems, especially in terms of explaining how words could refer to nonexistent entities or to entities not directly encountered (e.g., words used to refer to theoretical posits). Another, related approach, focuses not on how words or mental representations were first connected with objects, but how they are adaptations which enable our cognitive system to better adapt to the environment we encounter. The model for this is the function of a biological trait. The function of a biological trait is usually construed as either the activity whose performance led to the selection of that trait by an organism or that which is now enabling it to meet selection forces [35, 36]. Dretske [12] and Millikan [22] are two philosophers have advocated pursuing such a perspective with respect to mental states, arguing that it is the manner in which they have been selected within the system that determines their semantic content. Dretske, in particular, emphasizes the need for mental representations to be the product of a learning system (learning being, at least in part, a selective process). (The treatment of cognitive systems as adaptive provides a suggestion as to how we might explain our ability to refer to nonexistent entities. Our system, like any biological system, is only imperfectly adapted to its environment. At points of misfit our cognitive system may mischaracterize entities or represent nonexistent ones.)

Many traditional symbolic AI systems have not been learning systems, although recently there has been considerable interest in developing symbolic learning systems. But connectionist systems are generally learning systems. Learning occurs as a result of adjustment of weights within the system as the network performs the tasks assigned to it. One of the crucial advances leading to the reemergence of connectionist modeling in the 1980s was the development of a general learning algorithm, backpropagation, for networks consisting of multiple layers of units with feedforward processing [28]. It is common when analyzing connectionist networks to view various layers within the network as developing specific representations of the information provided at the input layer, representations that facilitate the network in performing the task it is being trained to do. This is clearly illustrated in a network Hinton [18] trained to respond to queries about relationships in two family trees. The two trees were isomorphic with each other, but one consisted of individuals with British names and the other with individuals with Italian names. Each tree consisted of three generations. The network was structured so as to receive as input a localist representation of one individual in the tree and of a relationship, and the network was trained to identify the other individual or individuals who stood in that relationship. (A localist representation is one in which a single unit, when activated, is designated to represent a particular entity or relation.) The network processed this information through three hidden layers. The first hidden layer consisted of two sets of six units, one of which received inputs from the units representing input individuals, the other from units representing input relationships. Hinton analyzed the connections leading from the input units to these hidden units, and was able to establish that these hidden units had learned to represent selected features of the input. For example, different hidden units receiving input from the units representing individuals learned to discriminate the generation of the individual, the nationality of the individual, and whether the individual was on the left side or right side of the tree. (Of course the network never saw the tree; it had to extract the relevant information from the problems it was given. Nonetheless, it determined that it needed to use its six hidden units to categorize the input information on different dimensions in order to solve the problem on which it was being trained.)

Analysis of the sort Hinton conducted is not generally possible, but connectionists have employed other techniques, such as cluster analysis [18] to determine how whole patterns of activation capture information supplied in the input. The success of these analyses makes it seem plausible to view these learned patterns as possessing real intentionality. The patterns of activation on hidden units result from the adjustment of weights between units as the network learns to respond correctly to inputs and these patterns of activation then supply all of the information about the input that is available to units later in the system. The assignment of a content to a particular pattern of activation is not simply an act of interpretation by the network designer. Rather, it is an interpretation that captures the causal relations within the network, particularly the causal relations that led to the pattern of activation occurring in the system in response to a particular input. The pattern of activation is as it is due to the causal history of the network, and it will affect the future behavior of the network as a result of possessing those causal powers. In this sense, the intentionality of network representations seems real.

When Searle argued that real cognitive systems possessed intrinsic intentionality, not mere *as if* intentionality, a major part of his case was that he, as a real cognitive system, knew the contents of his representations whereas the AI simulation did not. Only from an external perspective could one attribute content to the representations in the AI system. It is more difficult to run Searle's thought experiment in the case of a network since one of the feature of networks is that there is no central processing unit in which all operations are performed. Thus, there is no one set of tasks Searle might be imagined to perform in his head and evaluate whether, when he performs them, he achieves real intentionality with respect to that task. But even so, it seems plausible to argue that there is no awareness of what is being represented within the network. So, even if the above argument succeeds in showing that the attribution of intentionality in learning systems is capturing something real about the system, it still seems as if the AI system is fundamentally different from us in this respect. Hinton's network does not know what its representations represent. Part of this is due to the nature of the inputs and outputs of Hinton's network: it learns only to deal with symbolically encoded information, and its generates a symbolic representation. It never encounters members of the two families and never interacts with them. But that can be accommodated in a straight-forward manner: a network can be established inside a robot and receive its input from the sense organs of the robot and generate its output through the robot's motor system (see Nolfi, Elman, and Parisi [24] for a suggestive simulation of this sort). But even when the network is hooked up to the world in a more realistic manner, it does not seem likely that it will acquire the direct, first person awareness of the contents of its representations that humans seem to possess.

Accounting for the direct, first person awareness of content seems, at first, to be impossible in a computation system. The reason is that we have no plausible models of how such direct awareness might be accomplished in a physical system. A strategy one might pursue in this situation is to argue that such awareness is illusory (the strategy I am proposing here is similar to that Dennett [11] proposes for dealing with the qualitative character of mental events). But the feeling that we know the contents of our own conscious mental states is sufficiently powerful that if we are to hold that it is illusory, we are obliged to give a plausible account of how it arose. One suggestion is that it is due to the fact that we are language users. If one views language as a way of giving public expression to internal mental representations, and so dependent upon them, then appeal to language here cannot help us. But another possibility is that language is a relatively autonomous representational system, one which is learned in a public environment as a means of acquiring information from other and influencing their behavior (Bechtel [1]). Vygotsky [37] argues that only after language is learned in this public manner is its use internalized in private thinking. As we learn to use public language, one of the things we find useful is to characterize the mental states of ourselves and others. The activity of describing our mental lives in language is a separate activity from having the mental state. (Rosenthal [27] has argued for the claim that conscious awareness of a mental state is the result of having a second mental state directed at the first mental state.) We have information about our mental states that is not accessible to others as a result of them being states within us. We thereby account for the privileged access we have to our own mental states. But, just as importantly, this access is not direct in the way that would make it infallible. Having learned to use language to refer to events in the world, we must then learn to characterize the contents of our mental states in terms of features of the external world (see Gopnik [15] for supporting experimental evidence from developmental psychology for this suggestion).

So, perhaps we have a form of privileged access to our own mental states, and one that is more direct than other people are able to have, but not the less not direct and infallible access. We learn to use internal and perhaps behavioral cues to determine the content of our mental states and to express this content in language. If this account is correct, though, then it would seem possible for computational systems to achieve the same thing. We don't yet have computers fully capable of operating in a natural language, but progress is being made. What will be required, according to the analysis offered above, for computers to exhibit the same sort of access to mental states as we seem to have is for them not only to use language to characterize their environment, but also to characterize their own mental states in terms of how they represent the world. There do not seem to be any insuperable difficulties in developing computational systems of this kind, and one can even see a benefit of having such systems: by being able to represent their own mental states to themselves, such systems might be able to develop ways of revising their mental activity.

The final feature of consciousness that I set out above is the one that has been of most concern to philosophers. Mental states have distinctive qualitative characteristics. Seeing a blue object, for example, is a very different experience than seeing a red object. It is not just that we are able to differentiate the two experiences and so produce different verbal responses to them. They feel differently to us as we have them. Philosophers refer to the qualitative characteristics of mental states as *qualia*. Some philosophers (e.g., Dennett, 1988) deny the existence of qualia, arguing that nothing fits the definition of qualia as ineffable, intrinsic (atomic and unanalyzable), private, and directly accessible in consciousness. Dennett argues for this by a set of thought experiments or intuition pumps that attempt to break these properties apart and to show that to the degree any of the properties is satisfied it is not by means of mysterious entities known as qualia. Others have tried to show how qualia might be brought within the scope of physical explanation by showing how qualia are more complex than usually thought, and have features which co-vary with features of our underlying physical brain, such as features of our color or other sensory processing system (Hardin [17], Clark [7]). Both of these represent important advances towards developing a reductive explanation of our qualitative experience in terms of physical processes in the brain.

In this spirit, Lloyd [21] tries to show how a connectionist network might account for salient features of our phenomenal states. In addition to the four properties cited in Dennett's definition above, he claims that phenomenal states exhibit what he refers to as *phenomenal* *superposition* insofar as different features of objects presented to us are all integrated in our representation of the object. For example, our awareness of a red chair integrates our awareness of it as a chair and as red. The two features are not separately recorded as, for example, they are in a sentence describing the chair. The result is that subtle differences between two objects (e.g., subtle changes in shape or color) may result in similar but slightly different phenomenal states. The same characteristic is found in connectionist networks that distribute representational functions over sets of units so that a pattern of activation over a set of units serves to represent the relevant aspects of the input information and that same pattern represents a variety of pieces of information about that input. Thus, information is *superimposed* rather than stored discretely. This allows the network, for example, to respond subtle differences between two very similar inputs by generating very similar outputs. This similarity in representational features suggests to Lloyd that connectionism has the potential to explain the phenomenal features of consciousness.

For some (e.g., Chalmers [6], Cottrell [8]), however, these steps are not sufficient. There remains the brute fact that seeing red has a certain qualitative character that we are aware of and which has not been explained. Why do the physical or functional processes yield phenomenal states of this character? There seems to be nothing about the physical or functional processes occurring in us that causes seeing red to have the qualitative character it in fact has. If, in fact, this explanatory gap cannot be closed, it seems difficult to figure out what one might do to make an AI system experience the right qualitative character. But it is worth attending to what makes it seem that mental states have a qualitative character that escapes physical or functional explanation. Historically scientists have found ways to bridge gaps between concepts that did not fit into the same explanatory scheme. For example, genetic factors and chromosomes were not part of the same conceptual domain, but earlier in this century researchers noted a few important correlations between the state of chromosomes and genetic effects, proposed that genetic factors were in fact located on chromosomes, and proceeded to develop a powerful explanatory framework that linked other features of genetic factors and chromosomes (see Wimsatt [35]). Cottrell allows that eventually a new framework integrating the processing states and qualitative characters might be developed. But to the critics of computational and physical accounts of consciousness, such as Chalmers, the case of qualia is inherently different. The reason is that we are aware of the qualitative character from a first person perspective, while the physical and functional properties of the brain are known from a third person perspective.

If this is right, then the challenge to the computationalist is great since there seems to be nothing about the subjective character of these states that the computationalist can model. If the computationalist is to succeed, he or she needs to question the nature of our first person perspective on our mental states. Perhaps, as I suggested above, it is not as direct as it seems. Perhaps we seem to have direct access to the qualitative character of our mental states only insofar as we use other mental states, ones that figure in linguistic production, to characterize them. Then the problematic character of qualitative experience may be dismissed as an obstacle. We only need to explain what leads us to make the sorts of reports about our mental states that we do and the path may be open to a computational account of the sort Lloyd proposes. We can develop systems whose information about its mental states is the same as ours and which can therefore describe their mental states in the same way. But denying that the subjective characteristics of our mental states are really there and experienced by us seems too radical for many. If it is too radical, however, the prospects are dim for a computational account of the subjective character of our conscious experiences.

## 4. Conclusions

Can computational models of thought offer accounts of consciousness? While there has been limited work in AI on consciousness, some features of consciousness seem to lend themselves relatively easily to computational modeling. This applies to some of the features of consciousness I extracted from Searle's criticism of AI, that conscious mental states are intentional and that we are aware of their contents. I offered a connectionist simulation that suggests how intentionality might be captured in computational systems. The latter feature seems problematic insofar as we tend to think of conscious agencies of having direct first person access to the contents of their states, but I suggested that this may be an illusion and that we have simply learned to describe some of our mental states through other mental activity involved in language use. The least tractable feature of consciousness from a computational perspective would seem to be the qualitative character that seems to attach to conscious states, e.g., the qualitative character associated with seeing red. While the functional features of such states might be analyzable, there seems always to be the additional question of why they have the qualitative character they do. What makes the qualitative character difficult to account for computationally is that we are only aware of it from the first-person perspective. For computational accounts to get a foothold in explaining this feature of consciousness, it seems that they need to deny that we really have experiences with these subjective characteristics and claim only that we so describe our conscious states. If such a move is rejected, then the subjective character of mental states my be program resistant.

References

1. Bechtel, W. Decomposing intentionality: Perspectives on intentionality drawn from language research with two species of chimpanzees. *Biology and Philosophy*, 8, 1-32, 1993.

2. Bechtel, W. and Abrahamsen, A. A. *Connectionism and the Mind.* Basil Blackwell, Oxford, 1991.

3. Bechtel, W. and Richardson, R. C. *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton University Press, Princeton, 1993.

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

5. Churchland, P. M. *The engine of reason*. MIT Press, Cambridge, MA, in press

6. Chalmers, D. *Toward a theory of consciousness*. MIT Press, Cambridge, MA, in press

7. Clark, A. *Sensory qualities*. Oxford, Oxford University Press, 1993

8. Cottrell, A. Tertium datur? Reflections on Owen Flanagan's *Consciousness reconsidered. Philosophical Psychology*, in press.

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

10. Dennett, D. C. Quining qualia. In *Consciousness in contemporary science*, A. J. Marcel and E. Bisiach (Editors). Oxford University Press, Oxford, 1988.

11. Dennett, D. C. *Consciousness explained.* Little, Brown, and Company, Boston, 1991.

12. Dretske, F. *Explaining behavior*. MIT Press, Cambridge, MA, 1988.

13. Elman, J. Finding structure in time. *Cognitive Science*, 14, 179-211, 1990.

14. Flanagan, O. *Consciousness reconsidered.* MIT Press, Cambridge, MA, 1992.

15. Gopnik, A. How we know our minds: The illusion of first-person knowledge of intentionality. *The Behavioral and Brain Sciences* 16, 1-14, 1993.

16. Gunderson, K. *Mentality and machines*. Anchor Book, Garden City, NY, 1971.

17. Hardin, C. L. *Color for philosophers: Unweaving the rainbow*. Hackett, Indianapolis, 1988.

18. Hinton, G. E. Learning distributed representations of concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 1-12. Erlbaum, Hillsdale, NJ, 1986.

19. Johnson-Laird, P. N. *Mental models*. Harvard University Press, Cambridge, MA, 1983.

20. Llinas, R. R. and Pare, D. On dreaming and wakefulness. *Neuroscience* 44, 521-535, 1991.

21. Lloyd, D. Consciousness: A connectionist manifesto, in preparation.

22. Millikan, R. G. *Language, thought, and other biological categories*. MIT Press, Cambridge, MA, 1984.

23. Nagel, T. What is it like to be a bat? *The Philosophical Review* 83, 435-450, 1974

24. Nolfi, S., Elman, J. L., and Parisi, D. Learning and evolution in neural networks. Technical Report 9019, Center for Research in Language, University of California, San Diego, 1990.

25. Osherson, D. N. and Lasnik, H. (Editors) *An invitation to cognitive science*. MIT Press, Cambridge, MA, 1990

26. Posner, M. (Editor) *Foundations of cognitive science*. MIT Press, Cambridge, MA, 1989.

27. Rosenthal, D. M. Two concepts of consciousness. *Philosophical Studies* 49, 329-59, 1986.

28. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323, 533-536, 1986.

29. Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (Editors). *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: *Foundations*. MIT Press, Cambridge, MA, 1986.

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

31. Searle, J. R. Consciousness, explanatory inversion, and cognitive science. *The Behavioral and Brain Sciences* 13, 585-596, 1990.

32. Simon, H. A. *The sciences of the artificial*. MIT Press, Cambridge, MA, 1969

33. Stillings, N. A., Feinstein, M. H., Garfield, J. L., Rissland, E. L., Rosenbaum, D. A., Weisler, S. E., Baker-Ward, L. (Editors) *Cognitive science: An introduction*. MIT Press, Cambridge, MA, 1987.

34. Wimsatt, W. C. Teleology and the logical structure of function statements. *Studies in the History and Philosophy of Science* 3, 1-80, 1972.

35. Wimsatt, W. C. Reductionism, levels of organization, and the mind-body problem. In *Consciousness and the brain: A scientific and philosophical inquiry*, pp. 205-267, G. Globus, G. Maxwell, and I. Savodnik (Editors). Plenum Press, New York, 1976

36. Wright, L. *Teleological explanations: An etiological analysis of goals and functions*, University of California Press, Berkeley, 1976.

37. Vygotsky, L. S. *Language and thought*. MIT, Cambridge, MA, 1962.