Searle, Subsymbolic Functionalism and Synthetic Intelligence

Diane Law
Department of Computer Sciences
The University of Texas at Austin
dianelaw@cs.utexas.edu
Technical Report AI94-222
June 1994

Abstract

John Searle's Chinese Room argument raises many important questions for traditional, symbolic AI, which none of the standard replies adequately refutes. Even so, the argument does not necessarily imply that machines will never be truly able to think. Nevertheless, we must be willing to make some changes to the foundations of traditional AI in order to answer Searle's questions satisfactorily. This paper argues that if we constrain the sorts of architectures we consider as appropriate models of the mind other than the brain, such that they resemble the physical structure of the brain more closely, we gain several desirable properties. It is possible that these properties may help us solve the hard problems of intentionality, qualia and consciousness that Computational Functionalism has so far not been able to address in satisfactory ways.

Introduction

It may seem evident that if one is in basic agreement with the points John Searle makes with his Chinese Room Argument (1980, 1984), that would be reason enough to abandon all attempts at any form of what he calls strong AI. After all, the whole argument is meant to demonstrate the futility of endeavors in that direction. In this paper, I will attempt to show that even if Searle is correct, there is hope for success.

To begin, it is important to note that there are really several different paradigms within the area of Artificial Intelligence, whereas Searle directly addresses only what he calls "strong AI" in his argument. Of the different areas, two in particular (the symbolic and subsymbolic paradigms) seem to fit his definition and thus are possible targets of the Chinese Room Argument (CRA). The purpose of this paper is to examine both of these styles of AI, differentiating them from other types, in order to determine whether they are different in ways that can contribute significantly to the refutation of Searle's argument.

For its first few decades, the discipline of Artificial Intelligence was chiefly, if not exclusively, devoted to the symbolic paradigm. This is the sort of AI that everyone was doing when Searle introduced the CRA. Most of its practitioners have counted themselves as Computational or AI Functionalists in terms of their philosophical affiliation. As such, they have proffered various defenses against Searle's argument. This is hardly surprising, since one of the basic tenets of this philosophy is the idea that we can best characterize the workings of the mind as a series of operations performed on formal symbol structures (Bechtel, 1988; Newell 1976). Adherence to this notion entails a high-level view of the mind in the sense that it dismisses the function of individual neurons (and even of neuronal structures) in the brain as being unnecessary to an understanding of cognitive processes.

In the last ten to fifteen years, however, the field of AI has seen the growth of connectionism as a major paradigm. Although there is great variety in the aims and interests of connectionists, just as there is among the proponents of symbolic AI, there are many who are specifically interested in biologically plausible models of the brain and who feel that even our present-day, admittedly crude models of neural networks may give us new insight into the way the mind works. When Searle introduced the CRA, he was obviously targeting symbolic AI (in particular, the work of Schank's group at

Yale; see Schank and Riesbeck, 1981 for an overview), so it is possible to hypothesize that the CRA may not apply to the subsymbolic paradigm. The proponents of the symbolic approach are quick to point out that connectionists still have not proved that it is possible in practice to use artificial neural networks (ANNs) to model high-level cognitive functions effectively. Nonetheless, this paper will be concerned with defending connectionism from a more theoretical, rather than a purely practical point of view. What we would like to discover is whether connectionism might be able to lead us into new ways of thinking about the mind, productive ways of modeling it and ultimately, whether connectionist systems might be the key to true thinking machines.

Different Sorts of AI

Before we can discuss the possibilities for AI, we must first define exactly what our goals are. As it turns out, various researchers have entirely different aims. Some are not interested at all in anything we might call cognitive plausibility; their work lies more in the realm of engineering than in the cognitive sciences. This is the sort of work that the Department of Defense often contracts and it is generally motivated by a requirement for a very specific result in a very limited domain. We call upon techniques from AI simply because we cannot accomplish a particular goal with the ordinary algorithms that we study in other areas of computer science, such as systems or numerical programming.

Most people intuitively insist that intelligence involves something more than a computer blindly and deterministically executing a program in which a programmer explicitly defines the actions to be taken in each of several (carefully limited) situations. Thus, in order to accept that these programs exhibit something we might call artificial intelligence, we shall have to also accept that the modifier "artificial" places very severe constraints on our normal definition of intelligence. In some sense, this seems consistent with the constraints of meaning that are implied when we place "artificial" in front of other nouns. Surely no one would argue, for example, that there is any significant relationship between an artificial flower and a real flower other than its general physical shape, nor that an artificial limb is a really satisfactory replacement for a natural one. Since work with such concrete and practical aims is in no way concerned with human cognition, we need consider it no further.

There is a second group of AI researchers with a very different objective. As we shall see, we can further subdivide even this group, but for now, if we confine ourselves to consider only their most general high-level goals, we can temporarily put them all into a single category. This group includes all those people whose aim is to somehow emulate the workings of the human mind, whether it be with the intent of simply gaining a better understanding of the brain or with the intent of passing the Turing test or with the much more ambitious desire to eventual build the sorts of perfect androids that are popular in science fiction movies or on futuristic television programs.

As I noted, this latter group is not homogeneous. Searle would argue that those who rely on computer simulations only as tool to better understand the brain are not in the same category as those wishing to simulate a mind in all its aspects. I believe there are relatively few people who truly belong to the former category and who also consider their work part of AI. Most of them work outside the field of Computer Science and they are largely neuroscientists or psychologists. These are people who are concerned with real, human intelligence. As such, Searle would categorize their work as "weak AI" and he explicitly exempts them from his argument.

On the other side, we have a group of researchers who are at least nominally the target of the CRA. These are the people who really are trying to lay the groundwork for the eventual construction of something at least very similar to a human mind, either in silicon or perhaps in some other medium different from the human brain. This paper argues that one can count oneself in this group, but at the same time, be in agreement with most of what Searle says in all the variations of the CRA. This may seem paradoxical, but the fact is that we can still subdivide this second group once again.

On the one hand we have the proponents of what Searle calls "strong AI". His early definition of strong AI is this:

> [T]he computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states (Searle 1980).

A decade later, he makes this definition more concise, saying, "Strong AI claims that thinking is merely the manipulation of formal symbols" (1990). His American Philosophical Association address goes on to further clarify the problems that arise when symbolic manipulation forms the entire basis for programs which purport to think, claiming that since syntax is not defined in terms of anything physical,

> computation is not discovered in the physics, it is assigned to it. Certain physical phenomena [presumably such as the patterns of bits in registers and/or memory, in the case of a von Neumann-type computer] are assigned or used or programmed or interpreted syntactically (1991).

From these remarks, it is clear that Searle's complaints are chiefly with the Computational Functionalists or the adherents of traditional, symbolic AI. Of course there are many standard philosophical criticisms of functionalism (Bechtel, 1988, Churchland, 1986) and it is not the aim of this paper to repeat them all here. Still it is important to note that the foundation of this school of thought is a complete reliance on symbolic processing. These are people who take Newell and Simon's Physical Symbol System Hypothesis (1976) both extremely seriously and very literally. The argument is that we can reduce all intelligence in some way to the manipulation of symbolic tokens. In fact, many in the field of AI seem to define intelligence as just such manipulation. As a result, it follows that intelligence can be realized in a wide range of physical media.

There is yet one more style of research within the field of artificial intelligence: the relatively new school of connectionism. Before we proceed with a discussion of this last paradigm, I want to make it clear that there are all sorts of people who use connectionist models (in fields such as physics, mathematics and engineering) who are merely seeking practical answers to underspecified or intractable problems, just as there are in symbolic AI. No effort is made in this sort of research to ensure biological plausibility; indeed, the learning algorithms often come from formal disciplines such as nonlinear optimization and statistical mechanics. For this reason, we are not particularly concerned here with such work. On the other hand, there is a large number of connectionists whose interest is precisely the study of the way the brain works and who see connectionism as the most reasonable tool to carry out their investigations.

We must still exercise some care, however, when we say that many connectionists use artificial neural networks as a tool to discover how the brain might work, because quite often, they cannot truly count themselves among the proponents of weak AI. On the contrary, many of them see connectionism as the best hope for creating intelligence in some sort of machine other than the human brain. Although it is at present far beyond our capabilities to build an ANN of sufficient complexity to mimic the brain, as far as we can tell, it is not impossible in principle to do so. The idea is that if we could implement such a network, we would have reason to hope that it would "really be a mind," as Searle says.

This sort of work is an attempt at something we might more appropriately call "synthetic intelligence. We can make a distinction between artificial and synthetic intelligence in the same way we make a distinction between artificial and synthetic rubies. An artificial ruby is not a ruby at all and shares none of its chemical properties, while a synthetic ruby is truly a ruby. The difference is simply that a synthetic ruby is man-made in a laboratory, while a natural ruby is not. Given that connectionist techniques continue to improve, providing truer models of both neurons and neuronal structures, we might appropriately consider a connectionist success to be something more similar to synthetic intelligence.

In the original version of the CRA, Searle admits that he was not considering connectionist systems (hardly surprising since work in this area was far from the mainstream in 1980). Nevertheless, he claims that a variant of the argument applies equally to them (1990). He calls this variation the Chinese Gym:

> ...[Consider] a hall containing many monolingual, English speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture...and the outcome would be the same as having one man manipulate symbols according to a rule book. No one in the gym speaks a word of Chinese, and there is no way for the system as a whole to learn the meanings of any Chinese words.

In fact, if we take the system in isolation, Searle is right. This much alone does not improve the situation encountered in the original argument. Without a context, by which I mean some sort of causal connections to the world that allow the system to associate words and concepts with their referents, the network does not understand anything. Still, I find this argument to be somewhat problematic. Searle replaces the artificial neurons with men in a gym and then points out that no one of these men understands Chinese. If we map back now to the neurons the men represent, the argument seems to rest on the idea that to understand Chinese, each individual neuron must understand Chinese. This is not a realistic notion, since we know that, as processors, biological neurons are far too simple to understand a natural language. If our neurons do not individually understand language, then it is not fair to demand that the individual men in the room understand either. Thus, this part of the argument fails. The real problem with the Chinese gym (and the reason that I agree that it is not a system that understands) is that the system is isolated from the environment. The words that it processes have no referents and thus no semantics. If we add real causal connections to the system, in the form of transducers that serve as its only input, and effectors that serve as its only output, then we have the start of a grounded system where semantics play a crucial role. At this point, Searle's argument cannot stand.

Differentiating Connectionism and Symbolic AI

The Chinese Gym argument is not Searle's only reason for believing that connectionist systems are no more adequate for building a thinking mind than are symbolic systems. Despite the parallel processing afforded by ANNs (as well as by the brain), Searle points out that

> Any function that can be computed on a parallel machine can also be computed on a serial machine...Computationally, serial and parallel systems are equivalent...(1990).

Searle is not the only person to make this objection. Actually, it is quite common and has its origins in proofs that all sorts of systems are Turing-computable, connectionist models being one such class. The idea is that if all these systems are able to compute the same things, there can be no substantive difference between them. This is what Smolensky calls the implementationalist view (1988). I will argue, as have many others, that ANNs provide us with something besides Turing-computability.

Despite arguments to the contrary (such as Smolensky's), as recently as 1993, Marinov has devoted an entire paper to defending the proposition that there is no substantive difference between symbolic and connectionist models of cognition.

Marinov's argument relies critically on comparison of one particular sort of ANN (those using the back propagation learning rule) to a variety of standard symbolic machine-learning classification algorithms working on a single, narrowly defined task. His first claim is that such neural networks are strikingly unlike anything we know about the brain. In this he is correct, but since within the connectionist community back propagation is only one of many learning mechanisms (many others exist that are lower level and closer to what little we understand about how the brain learns), it seems as unreasonable to condemn all of connectionism on such a basis as it is to condemn all machines on the basis of the poor performance of the first attempt at powered flight.

Marinov does not attack nor dismiss the idea of biological plausibility as a desirable goal, however, so it is fair to assume that he agrees that this is indeed something worth seeking. Unfortunately, it is at least as difficult to defend the biological (or even the cognitive) plausibility of the standard symbolic machine-learning algorithms. Let us examine them to see why.

Machine-learning algorithms often use a standard technique of building decision trees based on some variation on the following general scheme. The programmer chooses a set of features, the values of which allow categorization of various examples. Quite often the training set consists of positive and negative examples of a single class although they may also represent two or more different classes. The program computes the probability of an example belonging to a particular class, which means that it must see all the examples and count the representatives of each class before it can begin to categorize. The next step is to compute a "gain" for each feature. Roughly speaking, this number is a measure of how much closer we are to being able to say that a particular example belongs to a given class. Computing this term involves a rather complicated summation of probabilities and logarithms. The program chooses the feature that gives us the largest "gain" to be the root of the decision tree. The process is repeated recursively for each feature until the tree is complete.

There are several problems with this, from the point of view of cognitive plausibility. First, the features that allow categorization come from outside the system. If we were to try to relate this to human learning, it would be something like telling someone learning the difference between a teacup and a mug that he or she should look at the ratio between the circumference of the bottom versus the circumference at the top, the thickness of the china, the height versus the width, and the size of the handle. In other words, a great deal of knowledge is built in from the beginning. The program need not discover the features that are important, as is generally the case for humans learning to categorize; they are given.

The second problem comes from the need to see a representative number of positive and negative examples before categorization can begin. This would be tantamount to having to discover the proportion of teacups in the world that actually have the same size top and bottom circumference, the proportion of mugs that has a handle so small that at most one finger will fit into it, etc., before we could begin to understand the difference between the two. Obviously, this is not something that humans need to do.

The third problem may only appear to be a problem; that is, we may find out that we are doing exactly what the machine-learning algorithms do. Still, intuitively, it does not seem quite right to have to solve a lot of complicated equations in order to tell a teacup from a mug, at least not explicitly. It is true that introspection can be a very bad indicator of what really goes on in the brain, but if we can trust it at all, then it appears that we choose features for categorization without so much formal mathematics. In fact, in the case of distinguishing types of cups, it seems that processing is quite parallel. We look at all the features of the example in question and make a decision based on all the information that our senses can give us.

Furthermore, the discussion above really only addresses cognitive plausibility, saying nothing of biology. Machine-learning techniques tell us nothing about how the brain might carry out these processes, whereas the connectionist counterparts at least show that it is possible for a large number of very simple processors working in concert (such as neurons in the brain) to learn to categorize.

A second major point that Marinov makes is that it is a straightforward matter to convert the decision trees that the machine-learning algorithms induce into explicit production rules which humans can easily understand and apply. In contrast, connectionist models store knowledge in a distributed fashion which is difficult, if not impossible to extract from the trained network. Whether or not this is a disadvantage depends to some extent on the goals of the categorization task. If the aim is to provide a set of rules that will allow a human to distinguish a cup from a non-cup (to use Winston's famous example), then there is no contest; machine-learning wins, hands down. On the

other hand, if our goal is to gain some understanding of the way that humans might distinguish members of a category from non-members, the connectionist system may give us a truer picture of the process. After all, it certainly doesn't seem as if we use explicit rules to figure out whether something is a cup or not. It is actually more likely to be a much lower level process, relying on visual perception and a fair amount of parallel processing, leading to simple object recognition. It seems odd that Marinov should demand biological plausibility in one breath, yet reject it in the next, if it turns out that biology doesn't produce the tidy results he desires.

In his response to the Marinov article, Clark (1993) makes some much more pointed distinctions. As he says, although the machine-learning algorithms can indeed employ microfeatures to induce decision trees, the researcher predetermines what those microfeatures will be.[1] We have already mentioned some of the problems that this occasions. On the other hand, ANNs that do not enjoy the benefits of specially prepared input data discover useful microfeatures on their own. Quite often, they do not correspond in any way to those that conscious thought posits, although they produce distinctions just as effectively. Since a great deal of categorization is not the result of conscious deliberation, it is at least worth speculating that perhaps the brain uses exactly such non-intuitive sorts of features in classification tasks. It seems plausible that they might, since the mechanics of processing in the brain bears more physical resemblance to the processing that occurs in ANNs than that of symbolic programs.

Connectionist models have another strength that Marinov ignores, but that Clark mentions. This is their ability to interpolate and to take advantage of what are often called soft constraints (see also Smolensky, 1988). Given that a network is trained on data that includes example I1 (producing output O1) and example I2 (producing output O2), when presented with a novel input I1.5 that lies between the two trained inputs, it will produce an output O1.5 (or possibly O1.4 or O1.6). Now, of course it is possible that such an output is in some way nonsensical, but it is also a way for the network to say, in effect, "well, it's something like I1, but it's also something like I2" Human beings seem to learn new concepts in this way quite often, using what people in educational psychology call "cognitive hooks" upon which to build new understanding. On the other hand, symbolic systems are incapable of handling this sort of input. A novel item either conforms to something it knows or it does not. The difference is simply that connectionist systems are inherently continuous, whereas symbolic systems are just as inherently and unavoidably discrete.

Of course we admit that there are ways to "fuzzify" knowledge representations in symbolic systems, but they typically require the introduction of some probabilistic measures, such as certainty factors, which are difficult, if not impossible to obtain accurately. With a connectionist system, the probabilities are gathered automatically and precisely as the system learns. We may see this as a purely practical problem for symbolic systems since we might argue that we could simply use an ANN to gather statistics and then pass them on to a symbolic system. Since the connectionist system can already make the correct decisions, the advantage to be gained would simply be increased explanatory power. The symbolic system operates with a set of rules that we can print out when a human user wants to know why the system made the decision that it did. As ANNs become more structured and more sophisticated, it is possible that they will be able to give rule-like explanations as well. In that case they would have a great advantage over symbolic systems, since they would not only be able to give explanations of why they produced a certain output, they would also be more explanatory in terms of how the brain does what it does

[1] There are connectionist models that take advantage of the same idea, hand encoding input to make the learning task as fast and simple as possible. It is interesting that a number of connectionists regard these systems as a form of cheating, preferring to concentrate research effort on developing new learning algorithms, new models of "neural" units and new automatic structuring techniques, rather than to have to partially solve the problem before the training ever begins.

and they would have no need to rely on an outside system for any of its computation.

Another important way in which connectionist models differ from their symbolic counterparts is in the way they represent information. As Clark points out, representations in connectionist systems are distributed and superpositional. There are several advantages to this sort of representation. The first seems in some sense to be a purely practical one; a distributed representation makes it possible to store a great many concepts in a relatively small space, since each unit participates in the representation of many different items. Still, this advantage is somewhat more than simply a means to get around having to buy computers with ever-greater amounts of memory. The fact is that the brain itself has a finite number of neurons, and this is one means of explaining how it can store so overwhelmingly many facts, procedures, and episodic memories.

Not only that, but the distributed representation also automatically affords a content-addressable, associative memory. This comes "for free," and seems to be just the answer we need for questions such as how it is that humans can so often bring just the right piece of information immediately to the fore, with no apparent search, or why it is that when we are "thinking, musing or reasoning, one thought reminds us of another" (Dellarosa, 1988).

We have also successfully used artificial neural networks to solve problems for which we have no satisfactory algorithms, most notably pattern-matching tasks. Handwriting recognition is one such area. To my knowledge, there is no symbolic method to solve this problem and, although the connectionist systems that we use to perform this job are not perfect, they are at least able to solve the problem to an extent acceptable for practical applications. Furthermore, even humans sometimes have trouble recognizing non-standard handwriting. This is a case where connectionist systems are definitely capable of doing something that we have not been able to do with symbolic systems Turing equivalence notwithstanding. In cases where we know of no algorithm that can produce the desired computation, ANNs can at least some times give us the solution we require.

There is yet another crucial difference between connectionist and symbolic models which is more important than any of the preceding arguments, since it represents an advantage for which there is no equivalent in any symbolic system: it is relatively straightforward to situate ANNs in such a way as to give them causal connections with the world. Strangely enough, however, few researchers have consciously taken advantage of this feature, even though this is the very thing that saves us from having to throw in the towel, even if we do believe that Searle is basically right. The problem with most connectionist models is that they treat the network as a "brain in a vat," unconnected from the world in which it exists (whether it be the real world or a simulated one, probably doesn't much matter, at least for purposes of early research). As Lakoff put it in his reply to Smolensky's target article,

> Smolensky's discussion makes what I consider a huge omission: the body. The neural networks in the brain do not exist in isolation: they are connected to the sensorimotor system. For example, the neurons in a topographic map of the retina are not just firing in isolation for the hell of it. They are firing in response to retinal input, which is in turn dependent on what is in front of one's eyes. An activation pattern in the topographic map of the retina is therefore not merely a meaningless mathematical object in some dynamical system; it is *meaningful*...One cannot just arbitrarily assign meaning to activation patterns over neural networks that are connected to the sensorimotor system. The nature of the hookup to the body will make such an activation pattern meaningful and play a role in fixing its meaning (1988).

This is a direct reply to Searle's main objection and it is a reply that I see as much more difficult to refute than any that have gone before. In some ways, it is similar to the standard replies, and it falls most squarely within the spirit of the systems reply. Yet the systems reply that I have seen defended so many times is wrong; not for the reason that Searle gives, but because the systems reply makes no requirement for causal connections with the environ-

ment within which a system functions. When it is precisely that environment which provides the input and which the output affects, the system either survives and prospers because it learns to understand the meaning of things and events around it and respond in appropriate ways, or it suffers and fails because it does not.

The reason that a connectionist model can have this special property, whereas a symbolic system cannot, is that inputs to a neural network have a direct causal effect on the state of the network; values on the input units determine the values throughout the system. To put it another way, the network builds its own internal representations of things as an immediate consequence of the input it receives. In this sense, the semantics of the model are an integral part of the system, and as such neither admit nor require arbitrary outside interpretation.

This is very different from the situation we find in symbolic systems. There, human users impose all the meaning of the symbols that a program manipulates by interpreting the output in appropriate ways. The fact that I attach the symbol 'female to the property list of the symbol 'Mary, means nothing to the computer, but when the program prints the symbols: "Mary is a female," it means something to a human observer. Most of us (Searle, quite notably) are not convinced that this constitutes understanding at all. We ask that the symbols mean something to the computer, in the same way that they mean something to us. This is what Harnad calls the symbol grounding problem (1990). He states the problem in this way:

> How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their shapes, be grounded in anything but other meaningless symbols?

Recently, a number of researchers (including Harnad himself, 1991) have begun an attempt to solve this problem (Feldman, Lakoff, et. al., 1990; Stolcke, 1990; Regier, 1991; Sopena, 1988; Jefferson, Collins et. al., 1990; Pfeifer and Verschure, 1991; Nenov, 1991). Without such a solution, we can have no hope for an intentional system. Tellingly, all of the researchers mentioned use neural networks as their tool of choice. This is not simply a coincidence. By design, connectionist systems are uniquely capable of doing what the brain seems to do. First, sensory neurons receive information directly from the environment (or from other neurons, in the case of recurrent networks). This input then directly determines the activation state of the rest of the network. In the case of object recognition, then, the perception of the object induces a particular pattern of activation in the network. This pattern is an internal representation of the object. The network can then attach a symbol (i.e., a name, which is also instantiated as a pattern of activation) to that internal representation. In this fashion, the symbol is grounded. It has meaning. Nothing but this particular meaning produces this particular pattern of activation. It is the object that directly causes it. We are not free to impose arbitrary interpretations on such representations; the representation is a representation of one specific thing and not of any other.

With the above points we have several significant differences between the symbolic and connectionist paradigms. It is important to recognize that they are differences in kind, proven Turing equivalence notwithstanding. The last point demonstrates that, despite the undisputed fact that connectionist systems are still crude and unsophisticated and that they are still weak in practice, in principle at least, they provide a crucial capability that symbolic systems cannot. Turing computability implies only that if an algorithm exists for a procedure that some Turing machine equivalent can carry out that computation. Obviously, since ANNs can provide for grounded meanings and since ANNs are Turing equivalent, some Turing machine can produce the same results, but apparently, it can do so only by simulating an ANN. This implies that ANNs have certain functional capabilities that no other Turing machine has. Although we have no algorithm for enforcing that representations mean something to a machine, we find that there is a way to ensure that they do, if we would only

exploit it. It is unfortunate that even many connectionists do not.

A Somewhat Different Functionalism

At least nominally, believing that it might be possible to construct a synthetic mind implies that one is a functionalist, since it also implies that the actual medium is not a critical feature of the properties that the mind possesses. Yet when we look at the traditional definitions of Computational Functionalism as a philosophical stance, we see that "it views the mind as carrying out formal operations on symbols encoded within it" (Bechtel, 1988). This gives us a great deal of latitude in terms of the sorts of machines we have at our disposal for implementation of the algorithms we hypothesize that the brain is carrying out. Nevertheless, it may be wise to give up some of that freedom, restricting ourselves instead to using machines that are brain-like in important ways (i.e., multiple, relatively simple processors connected in significant and complex ways). If we do, we also gain a certain freedom in the sense that we no longer have to restrict ourselves to purely algorithmic processing. There is certainly no *a priori* reason for specifying that functionalism concern itself strictly with symbolic computation, nor for assuming that the architecture of the brain is immaterial to the sorts of things that it does. We are still left with the notion that we can implement the functions of the mind in some medium other than the brain, even if we have narrowed the field of potential physical realizations of this processing.

What else do we gain if we dispense with the stipulation that all cognition is the result of formal symbol manipulation and that we should therefore model it as such? We can start with the least controversial gains: those that are admitted by those proponents of symbolic AI whom Smolensky refers to as revisionists (1988). Many people readily accept that we can best model certain so-called low-level aspects of cognition with correspondingly low-level connectionist systems. These include such things as olfactory, auditory and visual processing. It is clear that this sort of sensory processing is massively parallel in nature and that it deals routinely with noisy and even contradictory information. Perceptual systems must also have methods of processing analog data. Since these are precisely the specialties of ANNs, even many hard-liners in the symbolic camp admit that perhaps it is best to concede this part of the field to the connectionists.

The parallel aspects of the brain are apparently not confined to processing sensory information, however, and it seems unwise to suggest that the only explanatory power that connectionist systems might have is in the realm of these low-level processes. Indeed, there seem to be many activities that go on in parallel which a traditionalist would consider purely symbolic processing. An example of such parallel activity is the familiar phenomenon that occurs when we cannot remember a given fact such as a name, or solve a particular problem. We think very hard about it at a conscious level without success, but quite often, once we have abandoned the conscious mental search, the answer comes to us, seemingly "out of the blue." Minsky hypothesizes that this might be the result of demons that we set to work when we originally come across the problem (1986, p. 274). This might be a satisfactory high-level explanation, but the question remains, what exactly *are* these demons (other than programs) and how do they go about their work in the subconscious while we engage in unrelated mental activity at the conscious level?

In general, we will need to explain high-level processes in terms of lower-level ones. Since what we know about the brain indicates that all its activity consists of massively parallel neuronal firing, it is obvious that at some point we will need to explain how the activity of massive numbers of relatively simple processors can account for all of cognition. It may turn out that we cannot find any way that the brain could implement certain high-level theories. We may also discover that such an implementation is far more complex and unnatural than that required by a lower level, connectionist explanation. In either case, we will have to admit that even though our high-level explanations *seem* to explain mental activity, in the end, they are at best speculative.

Even though it may be the case that we can find symbolic, rule-like explanations for much of mental processing or that we can identify features that allow for effective categorization, it does not necessarily follow that the brain uses them in its own processing. Indeed, as we saw above in the discussion of Marinov's paper, a connectionist system may do at least as good a job of classification using features that are not only counter-intuitive, but which are sometimes completely opaque to conscious understanding. Thus it would seem to be a mistake to assume that the brain *must* do things (particularly those things that we do unconsciously) in the way that we can most easily describe. We can generally accept that introspection does not always lead to correct theories of the mind.

The previous paragraph sounds suspiciously like something the eliminative materialists might suggest and indeed, they may turn out to be right. On the other hand, there is so far no reason to believe that all the explanations for mental phenomena that cognitive science and symbolic AI have theorized are wrong. It may be that we *will* have to revise or replace at least some of them. Still, the fact that connectionist systems are Turing equivalent indicates that we could find a way to implement many of these theories in connectionist architectures, even though several unsolved problems currently stand in the way. That is not to say that following the implementationalist path is the proper thing to do, since we may find more explanatory theories by simply allowing ANNs to find their own ways of solving problems.

At the same time, there is one important argument against trying to reduce all of cognition to rule-following behavior. Considering the amount of attention that we must pay to the procedure of much of conscious explicit rule-following behavior and the care that we must take to perform the correct steps accurately and in the right order, it is difficult to explain how such behavior can take place in the subconscious with rules that we quite often cannot even state. Laird, Newell and Rosenbloom (1987) have proposed "chunking" as one possible solution to this problem and Anderson's Act* (1983) proposes the compilation of "macro-operators" as a similar solution. The problem with these models is that they assume an initial phase of explicit reasoning, following equally explicit symbolic rules as a prerequisite to building more automatic means for solving problems. Yet we find no evidence of such explicit behavior for much of what we do (using our native language is a case in point), nor evidence of explicit knowledge of the rules that we might conceivably have used to "reason through" the problem initially. It may be more productive to proceed with the idea that these "rules" are not at all like the symbolic rules that a production system (for example) uses. It is difficult to imagine exactly what these rules might be like unless we are familiar with connectionist systems, where we find that behavior that we can describe with rules occurs routinely without any rules in any sort of symbolic form being present in the system. If it makes us more comfortable, we can simply say that the "rules" are distributed among the weights in the system, just as the representations for other entities are.

This is not to suggest that we should simply discard all theories that rely on rule-following behavior, without further ado. For one thing, it is obvious that we at least seem to use explicit algorithms for conscious problem-solving. Thus, any theory of mind must be able to explain this phenomenon in some way. For this reason, current, ongoing research on the variable-binding problem in connectionist systems, although still preliminary, is exceedingly important (see Shastri and Ajjanagadde, 1993; Smolensky, 1990), since as I mentioned above, we will ultimately need to explain all such high-level behavior in terms of the sorts of processing that the physical brain can perform. Furthermore, it may well be that for certain purposes, the more transparent explanations that explicit-rule based theories offer will be more useful and easier to manipulate. Such purposes would be ordinary folk-psychology predictions of the behavior of our fellow human beings and methods to deal with problematic behavior such as learning disabilities, neuroses, and antisocial behavior. It doesn't seem helpful in such cases to simply say, "well, Johnny's brain is just wired up wrong. Short of adjusting all his synapses, there is nothing to be done!"

Difficult Problems for Synthetic Intelligence

There are many mental phenomena that neither traditional AI nor Computational Functionalism has been able to explain. Searle says that there are basically four features of mental phenomena that make the mind-brain problem intractable: consciousness, intentionality, subjectivity and mental causation (Searle, 1984). These are the really hard problems for synthetic intelligence, for Cognitive Science and for philosophy in general. Some Computational Functionalists (along with the eliminative materialists) have "solved" them by denying their existence to some extent or another; others have presumed that a machine running the "right" program would somehow produce them by some as yet unspecified means. Still others are satisfied with a purely behavioristic test of intelligence such as the Turing test, saying in essence that whether a machine simulation of the mind actually includes these features is immaterial, as long as the machine produces the appropriate outputs.

Yet these problems do not seem so easy to dismiss. Certainly, the average "man in the street" feels that these are important aspects of the mind, that they are at the heart of the "mark of the mental" and that without them, we cannot grant that a machine is truly intelligent. We evidently must deal with these features in a more constructive way if we are to satisfy our most intuitive requirements for intelligence.

We have already seen part of a solution to the problem of intentionality, when we considered the importance of connecting the system to the environment. Nonetheless, we have not yet solved the problem entirely. For one thing, at best, merely adding causal connections between machine-mind and world can only provide referents for concrete objects and events. Obviously, this does not take care of referents for things that do not exist, although we might reasonably surmise that many of these things are composites of things that do exist (e.g., unicorns or Santa Claus). Forming such composites is a strong point of connectionist systems which excel at interpolation tasks. We cannot dispose of other sorts of referents so easily. Many of the things for which every natural language has terms are simply not so concrete nor so compositional.

One major class of referents that belong to this group are the referents for subjective sensations or internal mental states. All of our terms for emotions, for example, fall into this category. Notwithstanding the idea that we may learn critical notions about such states as pain from the behavior they produce, as Wittgenstein argues (1953); intuitively, our most intimate understanding of the word "pain" comes from personal experience of painful feelings. We can make similar arguments for other words that refer to internal states, whether they be sensations or emotions. It is not at all clear that we shall ever be able to make a machine feel pain (or anything else for that matter), and thus there may not be any way to ground such terms. Yet it seems important to attempt to do so, since such feelings apparently have causal powers.

I have no easy answer to this problem, but I do have some hope that we shall find an answer. We are generally reluctant to grant that single-celled organisms feel even primitive sensations such as pain or even hunger (supposedly unlearned responses, Rolls, 1990) and of course it seems odd to imagine that they might feel happy or jealous. It is less strange to think about vertebrates feeling pain, although most people are still not willing to attribute the full range of human emotions to any animals other than humans themselves. The difference seems to be that as we observe more and more evolved organisms, we can more easily imagine that they are capable of an ever wider range of feelings.

If there is anything to our intuitions and if it is indeed the case that certain primitive emotions are innate, rather than learned, then it would seem that the most productive course to follow would be the course of evolution. In the absence of a fully explanatory theory of subjective sensation, the use of genetic algorithms (Holland, 1975; Goldberg, 1989) may be our best bet in an effort to produce artificial neural networks that function just as people do when they experience a sensation such as pain. According to theories extant in experimental psychology, we learn more "sophisticated" emotions, such

as fear, through the mediation of primary reinforcers such as pain (Rolls, 1990). If this is the case, then there is some hope that if we can achieve primitive internal states in an ANN through evolutionary processes, then other emotional states could follow through learning.

In some sense, this is not a completely satisfactory solution, since it is possible to imagine an embodied network that does "all the right things" when it runs into a table at full speed, for example, and yet which *feels* nothing, despite all its sensors. It might indeed be possible to use genetic algorithms to produce such robots, selectively reinforcing those networks that avoid painful situations whenever possible and which react in convincing ways (shouting and nursing the injured part, for instance) when avoidance is not possible. Still, we have no way to tell if they really experience something subjectively awful or if their reactions are purely behavioristic. On the other hand, we have no way to tell that about each other, either. If we ask why we are the way we are, we have no better answer than to say that in all probability we are that way because evolution made us so. If we use genetic algorithms to produce (perhaps only part of) synthetic minds, our answer to the question of why they behave the way they do is exactly the same.

In a more positive vein, it is difficult to conceive of an evolved ANN that would behave in convincingly appropriate ways while responding to combinations of stimuli[2] and still not be in some special state that we might reasonably identify as a pain state. If pain is instantiated in humans via particular brain states, then we are at least close to an answer. Furthermore, if we consider pain or other subjective experiences to be particular mental states with causal efficacy (i.e., able to cause other mental states to occur or able to provoke physical reactions), then we may identify the states produced in an artificial neural network[3] as just such states. Another way of looking at this is to ask what is happening in the brain of a human who feels something in particular, such as pain. There are certain physical changes that take place. For example, there are changes in the concentrations of certain hormones and transmitters and the firing rate of certain neurons changes (Rolls, 1990). The exact meaning of all these changes is not completely clear, but certainly, if an ANN were to undergo similar changes of state (ignoring for the moment hormonal changes; we can treat changes in levels of transmitters as changes in connection strength) in response to external or internal stimuli, it would seem fair to surmise that the network might actually be feeling something. At least, it is definitely in a state that is not normal and it has undergone the same sorts of state changes that occur in the human brain. If appropriate behavior accompanies these state changes, then we have a reasonably strong reason to believe that internal states similar to our own exist. At the very least, the system can now ground the word "pain" in its more social meaning. Furthermore, an agent capable of these state changes could understand at least something of what others go through in similar situations. If we were to expand our understanding of computers to include chemical processes as well as electrical ones, and we could show that the state changes are similar to those of humans, the claim grows even stronger. Whether evolutionary techniques can actually produce these effects is an empirical question.

With the last few paragraphs, I have outlined some concrete ways to try at least to deal with both intentionality and subjectivity, methods that have no direct counterparts in purely symbolic processing. According to Searle, we still need to consider consciousness and mental causation. When he speaks of the latter, we can

[2] We can imagine a creature engaged in some intensely pleasurable pursuit hardly noticing a mildly painful experience while the same creature might react violently to the same stimulus if it were already tired, frustrated or otherwise stressed.

[3] We must assume that the network in question is specifically one that is evolved through genetic algorithms and that we determine the fitness of such networks on the basis of their appropriate reactions to events that would cause particular subjective internal states in humans. Of course this is a behavioristic measure, but I see no other alternative.

presume he is talking about various causal powers of the mind. For instance, certain mental states can lead to other mental states or to motor action that has an effect on the environment. If this is all, then this is the easiest to attain of the four properties. Since the states of connectionist systems are by nature associative, it is obvious that certain states would lead naturally to other related states. This accounts for internal causality. On the other hand, if an ANN controls effectors as we have outlined above, then surely we cannot deny that the capability exists for the state of the artificial mind to alter facets of the environment.

Consciousness, is of course the most difficult of all. It is of necessity the hardest problem, if for no other reason than that we really don't have anything more than a vague intuitive notion of what it is. Patricia Churchland (1986, p. 370) relates a surprising story of the famous patient H.M. in which she notes that although he can solve the Towers of Hanoi Puzzle, he does not remember having done it before and he does not realize that he has the skills necessary to do it. It is, she says, "as though some part of H.M.'s nervous system knows what he is doing and has the relevant complex intentions, but H.M. does not." This seems extremely odd to us because it flies in the face of our intuitions about consciousness. How can we be "aware" (at the level of the nervous system) without being consciously aware? We talk about dreams as being the product of the subconscious mind, opposing it to the conscious mind, and yet while we are dreaming it seems very much like the sort of thing that our minds do while we are conscious. Indeed, the brain waves of dreaming subjects are very similar to those produced by alert subjects and quite unlike the brain waves produced in other states (Shepherd, 1983). We can take it for granted that, at least while we are alert, our brains are doing many things in parallel, processing all sorts of sensory information while we think consciously about an upcoming beach vacation or try to remember what else we needed at the grocery store. Perhaps at the same time, we have a nagging feeling that there is something else important to which we really should be attending. Indeed, we cannot "turn off" the train of conscious thought as long as we are awake, no matter how hard we try.

I suspect that consciousness has a great deal to do with attention, or perhaps it even is identical to attention. To put it in terms of structured artificial neural networks, we can imagine that we might have a connectionist system built out of many interrelated "specialist" networks, each of which performs specific tasks. Some of these networks are gating networks that determine which other networks can contribute their outputs to some larger process or computation. We can imagine heterarchies of such gating networks[4] that compete for dominance. Emergency situations would take immediate precedence, for example, while problem-solving would involve gating the outputs of inference-performing networks. These networks in turn feed and are fed by an appropriate associative memory module. It may be that our conscious thought reduces to nothing more exotic than this. If this is the case, then it appears quite possible that we could account for conscious thought via assemblies of ANNs.

Conclusions

Clearly we face many problems and uncertainties in the quest for a theory of mind. It is possible that some of the things that seem so hard are difficult simply because we are looking at them in the wrong way. Since the human brain is the only intelligent machine with which we are familiar, it does seem unwise to try to divorce intelligence from it completely and attempt to study cognition with purely abstract and formal methods. As the maxim goes in the field of aesthetic design, "form follows function." We know that evolution is a satisficer rather than an optimizer, but it does seem worth considering that the architecture of the brain is the way it is for some good reason.

One problem for the field of Artificial Intelligence is the way we go about designing our programs. We do it (just as they teach us in our

[4] Some systems like this exist already, with the gating networks being trained to do their jobs along with the other networks (e.g., see Jacobs, Jordan and Barto, 1991; Jordan and Jacobs, 1993).

first year programming courses) top-down. We are trying to simulate very high-level mental phenomena, but (ignoring what they teach us at school), we never bother to decompose the problem down to its low-level details. Of course there is one very good reason for this. If we did continue our design work down to the lowest levels, there would undoubtedly be several generations of us who would never get out of the design phase. That would mean several generations of researchers who would rarely have any reason to publish papers, which would indeed be a dire circumstance! Fortunately, there is a simple alternative: we simply do not start up so high. It is hard to imagine that we will be very successful if we keep trying to set a high level process on top of a void. In some sense, Marvin Minsky's late 60's program for stacking blocks did just that: it kept trying to place the top block first (Minsky, 1989).

It seems likely that we will not only need to think about the problems in more bottom-up fashion (some might argue that this is a giant step backwards), but we will probably have to change our emphasis in terms of the tools we use as well. Obviously, I think that we will probably find it useful to increase our reliance on neural networks; but I also believe that we cannot afford to just keep using the very crude models we have at present, but will have to continue to refine them and find ways to make them more realistic. Some work is already being done in this respect, for example Nenov (1991) has recently built a much more sophisticated neural model of memory than anything we have seen so far and is now working on biologically inspired models of attentional mechanisms. Shepherd et al. (1989) have similarly shown that more realistic neural models of cortical pyramidal neurons have significantly greater computational powers than the usual artificial neuron. We may also need to rethink our ideas about computers themselves, incorporating chemical processes as well as electricity. I also believe that it will be necessary for us to fall back on less deterministic methods. After all, the human brain was not built over the course of a few months or years, nor was it designed first and then implemented. My feeling is that we are not likely to be able to do a great deal better than Nature, and so I would guess that we will need to let evolution play a large role in shaping the architectures of mind that we employ. It is quite likely that if we can succeed in using this sort of tool to implement true synthetic intelligence, we won't end up with a copy of the human brain, but that may be so much for the better, since if we find that the exact structure of the brain is not essential for intelligence, that in itself will tell us a very great deal of what we would like to know. It is my hope that we will not have to go so far as to simulate a brain neuron for exact neuron, but I am fairly certain that we do need to begin at a level that is much closer to the neuron than to the symbolic representation of abstract ideas. Whether we succeed or not is still very much an open question, but it seems obvious that if we are to do so, we shall need to avail ourselves of many of the solutions that Nature has already derived.

References

Anderson, J.R. 1983. *The Architecture of Cognition. Cambridge.* MA: Harvard University Press.

Bechtel, W. 1988. *Philosophy of Mind: An Overview for Cognitive Science*. Hillsdale NJ: Lawrence Erlbaum Associates.

Churchland, P.M. 1986. *Matter and Consciousness*. Cambridge, MA: MIT Press/Bradford Books.

Churchland, P.S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, MA: MIT Press/Bradford Books.

Clark, A. 1993. Superpositional Connectionism: A Reply to Marinov. *Minds and Machines*, 3:3 pp. 271-281. Kluwer Academic Publishers.

Dellarosa, D. 1988. The Psychological Appeal of Connectionism. *The Behavioral and Brain Sciences* 11:1 pp. 28-29. Cambridge University Press.

Feldman, J.A., G. Lakoff, A. Stolcke and S.H. Weber 1990. Miniature Language Acquisition: a Touchstone for Cognitive Science. *Proceedings of the 12th Annual Meeting of the Cognitive Science Society*.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42:1-3 pp. 335-346.

— 1992. Connecting Object to Symbol in Modeling Cognition. In A. Clarke and R. Lutz. (eds.) *Connectionism in Context*. Springer Verlag.

Harnad, S., S.J. Hanson and J. Lubin 1991. Categorical Perception and the Evolution of Supervised Learning in Neural Nets. Presented at the American Association for Artificial Intelligence Symposium on Symbol Grounding: Problem and Practice. Stanford University, March. 1991.

Holland, J.H. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

Jacobs, R.A., M.I. Jordan and A.G. Barto 1991. Task Decomposition Through Competition in a Modular Connectionist Architecture: The What and Where Vision Tasks. *Cognitive Science* 15, pp. 219-250.

Jefferson, D., R. Collins, C. Cooper, M. Dyer, M. Flowers, R. Korf, C. Taylor and A. Wang 1990. Evolution as a Theme in Artificial Life: The Genesys/Tracker System. TR-UCLA-AI-90-09.

Jordan, M.I. and R.A. Jordan 1993. Hierarchical Mixtures of Experts and the EM Algorithm. A.I. Memo No. 1440/ MIT.

Laird, J.E., A. Newell and P.S. Rosenbloom 1987. Soar: An Architecture for General Intelligence. *Artificial Intelligence* 33:1 pp. 1-64.

Lakoff, G. 1988. Smolensky, Semantics and the Sensorimotor System. *The Behavioral and Brain Sciences* 11:1 pp. 39-40. Cambridge University Press.

Marinov, M.S. 1993. On the Spuriousness of the Symbolic/Subsymbolic Distinction. *Minds and Machines*, 3:3 pp. 253-271. Kluwer Academic Publishers.

Minsky, M. 1986. *The Society of Mind*. N.Y.: Simon and Schuster.

— 1989. The Intelligence Transplant. *Discover*. 10:10, pp. 52-8.

Nenov, V.I. 1991. *Perceptually Grounded Language Acquisition: A Neural/Procedural Hybrid Model*. TR-UCLA-AI-91-07

Newell, A. and H.A. Simon 1976. Computer Science as Empirical Inquiry: Symbols and Search. Reprinted in J.L. Garfield (ed.) *Foundations of Cognitive Science: The Essential Readings* 1990, pp. 113-138. N.Y. Paragon House.

Pfeifer, R. and P. Verschure 1991. Distributed Adaptive Control: A Paradigm for Designing Autonomous Agents. In F. J. Varela and P. Bourgnine (eds.) *Proceedings of the First European Conference on Artificial Life: Toward a Practice of Autonomous Systems*. Cambridge, MA: MIT Press/Bradford Books, pp. 21-30.

Regier, T. 1991. Learning Perceptually-Grounded Semantics in the L_0 Project. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*.

Rolls, E.T. 1990. A Theory of Emotion, and its Application to Understanding the Neural Basis of Emotion. *Cognition and Emotion* 4:3. pp. 161-190.

Schank, R.C. and C.K. Riesbeck 1981. Inside Computer Understanding. Hillsdale, New Jersey: Lawrence Erlbaum.

Searle, J.R. 1980. Minds, Brains and Programs. *The Behavioral and Brain Sciences.* 3, pp. 417-58.

— 1984. *Minds Brains and Science*. Cambridge, MA: Harvard University Press.

— 1990. Is the Brain's Mind a Computer Program? *Scientific American* Jan. 1990 pp. 26-31.

— 1991. Is the Brain a Digital Computer? *Proceedings of the American Philosophical Association* 64:3 pp. 21-37.

Shastri, L. and V. Ajjanagadde 1993. From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables and Dynamic Bindings Using Temporal Synchrony. *The Behavioral and Brain Sciences* 16, pp. 417-494.

Shepherd, G.M. 1983 *Neurobiology*. N.Y. Oxford University Press. p. 478.

Shepherd, G.M., T.B. Woolf and N.T. Carnevale 1990. Comparisons Between Active Properties of Distal Dendritic Branches and Spines: Implications for Neuronal Computations. *Journal of Cognitive Neuroscience*. 1:3 pp. 273-286.

Smolensky, P. 1988. On the Proper Treatment of Connectionism. *The Behavioral and Brain Sciences* 11:1 pp. 1-74. Cambridge University Press.

— 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* v. 46 pp. 159-216.

Sopena, J.M. 1988. Verbal Description of Visual Blocks World Using Neural Networks. UB-DPB-8810. Universitat de Barcelona.

Stolcke, A. 1990. Learning Feature-Based Semantics with Simple Recurrent Networks. International Computer Science Institute, Berkeley, CA. TR-90-015.

Wittgenstein, L. 1953. *Philosophical Investigations*. N.Y: Macmillan.