Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Minds, Machines and Searle

Stevan Harnad
Cognitive Science Laboratory
Princeton University
221 Nassau Street
Princeton NJ 08544-2093
harnad@cogsci.soton.ac.uk

ABSTRACT: Searle's celebrated Chinese Room Argument has shaken the foundations of Artificial Intelligence. Many refutations have been attempted, but none seem convincing. This paper is an attempt to sort out explicitly the assumptions and the logical, methodological and empirical points of disagreement. Searle is shown to have underestimated some features of computer modeling, but the heart of the issue turns out to be an empirical question about the scope and limits of the purely symbolic (computational) model of the mind. Nonsymbolic modeling turns out to be immune to the Chinese Room Argument. The issues discussed include the Total Turing Test, modularity, neural modeling, robotics, causality and the symbol-grounding problem.

1. Introduction

Searle's (1980a) "Chinese room" argument against "Strong AI" has had considerable influence on the cognitive science community (e.g. Edelson 1982; McDermott 1982; Harvey 1985; Searle 1982a, 1983, 1985). Like its precedessor, Lucas's "Goedel-unprovability" argument (Lucas 1961; cf. Slezak 1982), it has challenged the computational view of mind and inspired in many respondents the conviction that they have come up with decisive, knock-down counterarguments (e.g. Abelson l980; Block 1980, etc., Searle 1980b). Yet the challenge does not seem to want to go away (see Russow 1984; Carleton 1984; Hanna 1985; Rey 1986). Indeed, some have gone so far as to to define the field of cognitive science as the ongoing mission of demonstrating Searle's argument to be wrong[1] (Hayes, p. 2, in Lucas & Hayes 1982). The debate has at times even overflowed into the general intellectual press (Searle 1982b, c; Dennett 1982). This inaugural issue of a new journal devoted to experimental and theoretical AI seems a natural place for someone who has abstemiously umpired the debate now for almost a decade in the pages of "Behavioral and Brain Sciences" to air his own attempt, if not to settle the matter once and for all, then at least to make explicit those points -- empirical, theoretical, and intuitive -- on which we can all agree to disagree, pending more ideas or data.[2]

2. Searle's Chinese Room Argument

2.1 Simulating Simulation.

Searle formulates the problem as follows: Is the mind a computer program? Or, more specifically, if a computer program simulates or imitates activities of ours that seem to require understanding (such as communicating in language), can the program itself be said to understand in so doing? Searle's argument is based on a very simple "simulation" of his own:

First, suppose there is a computer program that can simulate the understanding of Chinese by examining all the Chinese symbols it receives as input and consulting an internal look-up table that indicates what symbols it should send back as output. If such a program could perform (one might say dissimulate) very well -- well enough, let's suppose, to convince a Chinese speaker at a teletype terminal that he was telecommunicating with another Chinese speaker rather than with a machine -- it would have passed "Turing's Test" (Turing 1964). According to that test, we should stop denying that a machine is "really" doing the same thing a person is doing if we can no longer tell their performances apart. That is, if a person does understand Chinese, and the machine performs so well that we can't tell whether or not it's a person, then the machine must understand Chinese too.

Now Searle "simulates" the simulation: Suppose that, instead of a computer program, Searle himself (who understands no Chinese) receives the Chinese symbols, does the look-ups in the table, and sends back the requisite output symbols. Since Searle obviously would not be understanding Chinese under those circumstances, neither could the computer simulation he was "simulating" have been. And neither, by extrapolation, could any computer simulation of anything. So much for the Turing Test, and the mind as a computer program.

On the face of it, this argument looks valid. It certainly works against the most common rejoinder, the "Systems Reply" (e.g., Wilensky 1980), which claims that, even though Searle is obviously not doing any understanding under Chinese-Room conditions, there is some superordinate "system," of which Searle is merely a part, that is really doing the understanding. In many respects it is just this sort of uncritical hand-waving (akin to claiming that more of the same -- more complexity or more capacity or more speed -- will add up to understanding) that Searle's argument was formulated to discredit. For if I myself am not understanding under these conditions, Searle argues, then what could there possibly be about the symbol-tokens themselves, plus the chalk and blackboard of the lookup table, plus the walls of the room, that could be collectively "understanding"? Yet that is all there is to the "system" besides me! Searle means to show, using the "intuition pump" provided by his Chinese Room, that nothing more substantial than this counterintuitive act of faith underlies the belief that if the right software is running on the hardware, the system will understand. My own critique of Searle's critique will attempt to decompose the problem in a more perspicuous way than the Systems Reply has been able to do. In the process, perhaps some of the intuitions underlying the Systems Reply will be resurrected in a more explicit and defensible form. Let us now take a much closer look at Searle's argument.

2.2 Simulation versus Implementation.

A large portion of Searle's argument seems to revolve around the notion of "simulation," so we must first indicate much more explicitly just what a simulation is and is not: Suppose that computers had been invented before airplanes (and that a lot of aerodynamic theory, but not enough to build a successful plane from scratch, was likewise already known at the time). Before actually building a prototype plane, engineers could under those conditions save themselves a lot of trial and error experience by simulating flight -- i.e., putting into a computer program everything they knew about aerodynamic factors, lift, drag, etc., and then trying out various designs and materials by merely simulating these "model" airplanes with numbers and symbols.[3]

Now suppose further that enough of the real factors involved in flight had been known or guessed so that once the computer simulation was successful, its first implementation -- that is, the first prototype plane built according to the principles that were learned from the simulation -- actually flew. Note, first, that the sceptics who had said "simulation ain't flying" would be silenced by the "untested" success of the first implementation. But an important distinction would also become apparent, namely, the difference between a simulation and an implementation. A simulation is abstract, an implementation is concrete; a simulation is formal and theoretical; an implementation is practical and physical. But if the simulation models or formalizes the relevant features (Searle calls them "causal" features) of the implementation (as demonstrated in this case by the successful maiden voyage of the prototype airplane), then from the standpoint of our functional understanding of the causal mechanism involved, the two are theoretically equivalent: They both contain the relevant theoretical information, the relevant causal principles.

The idea of a mechanism is really at the heart of the man/machine (mind/program) problem. We must first agree on what we mean by a mechanism: A mechanism is a physical system operating according to causal, physical laws (including specific engineering principles). A windmill is a mechanism; so is a clock, a plane, a computer and, according to some (including this writer), an amoeba, a rodent, and a man. We must also agree on what we mean by understanding a mechanism: A mechanism is understood if you know its relevant causal properties. How can you show that you know a mechanism's relevant causal properties? One way is by building it. Another is by providing a successful formal blueprint or program for building it.[4]

Now it seems obvious that there are circumstances in which it would make sense to try to simulate a mechanism and circumstances in which it would not. It would make sense to simulate if you did not know the relevant causal properties, or did not know them all, or were not sure you knew. In all these cases, getting a computer model to work would be informative.[5] On the other hand, it would make little sense to simulate if you already knew all the relevant causal properties of a mechanism, in that, say, you had already built it. So simulation is informative before successful implementation, but trivial after.

2.3 Symbolic Functionalism versus Robotic Functionalism: The Hardware Fallacy.

Note that the word "implementation" can have two rather different meanings. It is important to be clear about which one is intended here: "Software" (a program) is sometimes said to be "implemented" when it is actually run on hardware (a computer). Let us call this kind of implementation "c-implementation." This is not how the word implementation is being used here (although it is often what philosophers, especially "functionalists," have in mind when they talk about a computer being in certain "states" as a function of the software running on it; Putnam 1975). What is meant in this paper by an implementation -- to be referred to as "p-implementation" in this section only, so as to contrast it with c-implementation -- is the actual physical device that is built on the basis of the causal principles formally encoded and tested in a computer simulation. By way of an extremely simple example, a thermostat can be simulated by a program that takes certain numbers ("temperatures") as inputs and does a comparison to see whether the input is less than or equal to a certain number, putting out a "1" ("turn on furnace") if it is, and a "0" ("turn off or leave off furnace") if it is not.[6] The p-implementation is not the running program, but an actual thermostat (and furnace) built according to these formal principles, with a thermometer measuring temperature and a switch triggered when the mercury drops to a specified height (and a furnace activated by the switch, etc.). The formal features of the simulation become physical, causal ones in the p-implementation.

A computer is of course a physical, causal system, too, but the causality involved is just the switching of electrical circuits according to circuit-logic.[7] Hence a c-implementation is in reality a very special case of a p-implementation. Moreover, it is important to note (and will come up again later in this discussion, section 3.2) that in most p-implementations, part, "but not all" , of the physical mechanism could actually be replaced by a digital computer (i.e., by a c-implementation): In the case of the thermostat system, the decision mechanism could be a computer, but the thermometer and furnace could not.[8]

It should also be pointed out that the cognitive-science counterpart of the functionalist philosopher's pet topic, the disembodied "brain in a vat" (Dennett 1981) is not an all-purpose computer in a certain machine state (i.e., not a c-implementation) but a disembodied, implemented mechanism -- one that lacks its peripheral input-output devices but retains all of its (albeit short-circuited) central ones (such as, perhaps, a real internal thermometer). In other words, the brain-in-a-vat is either a p-implementation or a hybrid p/c-implementation. The tendency to pre-emptively equate the brain instead with a general computer (i.e., a pure c-implementation) can be dubbed the "hardware fallacy" (Harnad, pp. 79 - 86, in Lucas & Hayes 1982, and Harnad, in Donchin, forthcoming).[9]

The form of functionalism underlying the hardware fallacy is better described as "symbolic functionalism" : the belief that mental function is really just symbolic (e.g., verbal, inferential, computational) function; that the mind manipulates symbols the way a Turing machine does, and hence that the brain just supplies the hardware for doing computation -- just a c-implementation (Fodor 1981; Pylyshyn 1984). Symbolic functionalism is to be distinguished from a rival form of functionalism that will be argued for here: "robotic functionalism" (Harnad 1987c). For the robotic functionalist, not only are nonsymbolic functions (e.g., sensory, motor, analog, and associative functions) potentially just as mental or cognitive as symbolic functions, but they may even be primary, with the symbolic functions "grounded" in the nonsymbolic ones (Harnad 1987a, b). The distinction between symbolic and robotic functionalism will become clear later in this paper.

2.4 Modeling Parts versus Modeling Wholes: The Total Turing Test.

The first answer to Searle's Chinese-room argument is therefore this: The simulation of understanding Chinese does not understand Chinese any more than the simulation of flying flies. It is the implementation of a successful simulation that would understand (or fly). But notice that I said successful simulation. It has already been suggested that getting a simulation to work is not a trivial task (no matter how it may look from the philosophical armchair) -- in fact, getting one's model to work is the name of the game in the study of Artificial Intelligence. Notice that what Searle does is first to conjecture that the way to simulate the understanding of Chinese is to use look-up tables that tell you what symbol to match with what. Then he "simulates" that simulation by substituting himself as the looker-upper. Then he conjectures further: "Suppose that after a while the programmers [!] get so good at writing the programs and I get so good at manipulating symbols that my answers are indistinguishable from those of native Chinese speakers. I can pass the Turing Test for understanding Chinese...(1982a, p. 5)." This would be truly remarkable! Without so much as a trial run of a program, Searle has proposed a solution to the problem of language comprehension, production and translation (a problem that has defied the efforts of a generation of computational linguists): Use look-up tables. And this solution is supposed not only to work, but to work well enough to pass the Turing Test. This would be quite a feat. Unfortunately, there is no evidence at all that it would work (and plenty of evidence that it would not). "No matter," Searle may reply, "pick whatever kind of program you think will work and I'll `simulate' your simulation in the same way."

Fine. The only problem is that a successful simulation may have to do a lot more than the circumscribed Chinese symbol manipulation and look-ups that Searle has prescribed. If the first shortcoming of Searle's argument had to do with simulation versus implementation, this second shortcoming concerns simulating parts versus simulating "wholes" : The Turing Test, after all, allows you to tele-interrogate the computer in any way you like in order to test whether it is a person: You can ask it about its past, its inner life, its hopes, its fears: anything you could discuss with a person.[10]

Needing to be able to handle any reasonable linguistic interaction that a person could handle puts quite a burden on look-up tables and symbol-matching. It is true that the theoretical Turing Machine (of which a real digital computer is, by the way, a p-implementation) has only primitive operations such as symbol-matching and look-up, but no one has suggested that there is any way to bootstrap to a working model of language translation (let alone mind) using only such primitive operations. A program, which is just a formalization of a theory, can simulate just about any kind of operation the theorist can invent, including the transduction and digitization of sensory input, the manipulation of analog images, the extraction of sensory features, the construction of hierarchies of abstract categories, the storage and execution of movement programs, etc. Who is to say that the Turing Test, whether conducted in Chinese or in any other language, could be successfully passed without operations that draw on our sensory, motor, and other higher cognitive capacities as well? Where does the capacity to comprehend Chinese begin and the rest of our mental competence leave off? Searle has made the implicit assumption here -- one that he happens to share with his opponents in AI! -- that there could exist a self-sufficient "module" that was able to pass his purely verbal Turing Test without simultaneously being able to do "everything else" we can do, i.e., without also being able to pass the Total Turing Test. But despite the popularity of the modular view of mind these days (Fodor 1985), there is no evidence that a macromodule of that particular kind (i. e., an autonomous language-comprehension/production module) could work -- and certainly not with solely the impoverished repertoire of operations Searle has granted to it.

Searle's underestimation of the significance of the problem of actually getting a model to work is further illustrated by his cheerful confidence about the imminence of a solution to the chess-playing problem: "An oddity of artificial intelligence, by the way, is the slowness of programmers [!] in devising a program that can beat the very best chess players. From the point of view of games theory, chess is a trivial game since each side has perfect information about the other's position and possible moves, and one has to assume that computer programs will soon be able to outperform any human chess player" (1982a, p. 6).

The problem is that although all possible chess moves are indeed known, the brute-force "combinatorial" solution -- that of trying every possible move -- won't work, because (like waiting for the proverbial monkey to type Shakespeare) it requires unrealistically much computation and time. To find a realistic "short-cut" that avoids "combinatorial explosion" is in fact the definition of the problem in chess simulation, and indeed in mind simulation as well (Harnad 1982a). The data on our "competence" -- our cognitive performance capacity -- are already "in" , so to speak (Harnad, p. 86 in Lucas & Hayes 1982 and Harnad, in Donchin, forthcoming): We already know that we are able to discriminate, manipulate, sort, name and describe certain "objects" (concrete and abstract, including events, names and descriptions). Now the theoretical burden is to explain how the kind of device we are -- or any kind of device -- can generate such a performance capacity. Yet even having this rough idea of what people can do (all our possible "moves") unfortunately does not guarantee any quick solution to the problem of how we do it, because combinatorics can't be the way. Neither brains nor computers have the capacity to store and operate on "every possibility" (whatever that may mean). And reducing combinatorics to realistic size is anything but a trivial problem (Harnad 1976). Moreover, chess-playing, like language understanding, may not be modular; i.e., it may not be isolable from the rest of our cognitive capacity; so their respective solutions may converge.[11]

2.5 Theory-Testing versus Turing Testing: The Convergence Argument.

Hence it turns out that it is not real simulation that is trivial, but "thought-simulation," in which someone says "Suppose we could accomplish X by doing a and "b" ..." (e.g., suppose we could trisect an angle by using a compass and straight-edge). For obviously this statement (and whatever follows from it) cannot be very informative if one in fact cannot accomplish X by doing a and "b" . Artificial Intelligence is a kind of put-up-or-shut-up discipline. If you want to talk about what a model or a simulation can or cannot do, first get it to run (see Harnad 1982a).

Note that there is also an important difference between two senses of the word "test" that Searle tends to run together. One involves the formal testing of a theory by modeling its principles on a computer and testing whether, given the appropriate input data, it will produce the predicted output. The other is the Turing "test" in which a person tests informally whether he can tell if he is dealing with a computer or a person when they interact (i.e., receive one another's output as input). Of course, to pass the informal Turing Test (i.e., to convince a person), a candidate model must first pass the formal Turing Test (i.e., be able to generate our performance capacities).

One must agree with Searle, however, that most actual programs in contemporary AI are trivial, and that when you look at how they accomplish the superficially impressive feats they can perform (playing simple games, recognizing and manipulating simple objects, describing simple scenes, engaging in simple conversations, solving circumscribed problems), they turn out to be a bag of ad hoc tricks, specialized for the specific tiny task they were designed to perform, with little or no generality from one task to another, and no apparent resemblance whatever to the way human beings actually accomplish those same tasks. It is partly for this reason that these have been dubbed "toy" problems, by way of contrast with the real thing. And yet, for all their simplicity, and despite being based on ad hoc tricks, these toys "work" , whereas Searle's hypothetical model does not. And whereas Searle's thought-experiment and scepticism lead nowhere, there is every reason to believe that AI's toys will grow and grow, until some day they will begin to look more like the real thing, and some of them may even begin giving the Turing Test a run for its money.

This amounts to what might be called the "convergence" argument for the psychological reality of AI models (Harnad 1982b, 1987b), namely, that there may indeed be many arbitrary ways of simulating tiny parts of the performance capacity of a mechanism as complex and powerful as the mind, but as you become more ambitious (and successful) and begin to model something that is closer to the mechanism's total competence, specialized tricks should become less and less effective, and more general, all-purpose solutions will have to be discovered, if for no other reason than (1) parsimony and (2) generality (the capacities of the whole must subsume all the capacities of its parts). More general models must converge on more general principles, supplanting special-purpose tricks, and narrowing the options (as well as the gap between brain and mind, intelligent machine and man).

To deny the psychological reality of AI models a priori with the argument that a theory may fit all the behavioral data (i. e., all the observable evidence) yet fail to be "true" of the organism -- i.e., the canard that AI may never converge on "the way we actually do it" -- seems equivalent to arguing that although a physical theory fits all the physical data it may still not be "true" of the world. Chomsky's (1980, pp. 11 - 12) argument for the psychological reality of linguistic theory on the ground that all nontrivial theories are underdetermined by data seems pertinent here, as do functionalist arguments about the implementation-independence of cognitive theory (Pylyshyn 1984). Only better rival theories can provide a rational basis for doubting a successful theory, not a priori scepticism about "mere simulation."

It is a corollary of the convergence argument that, just as there are likely to be few viable substantive options (if any) among the possible ways of successfully designing a whole person, so it is highly unlikely that one of those options will turn out to be as radically different as a complete, insentient automaton: a robot that is totally indistinguishable from us in everything it does and says, yet fails to feel as we feel, having no subjective life, no perception, no understanding, no consciousness, just behavior (Harnad 1982b). In any case, such a possibility, though often contemplated by philosophers (e.g. by Searle 1982b, p. 57: "In general two systems can produce the same external effects while working on quite different internal principles"), surely represents an untestable, and hence indeterminate, conjecture. We can stretch "behavior" to include not only everything a person does, but also everything his brain does, and even everything the molecules in his brain do. But if that isn't enough -- if internal "effects" are still more elusive than that -- then one must agree with Dennett (1982, p. 56) that there seems to be no basis for continuing to call such effects "physical" at all; and that even though Searle continues to protest that he regards them as "caused by" and "realized in" " the physical brain (p. 57), he must, as Dennett surmises, be "some sort of dualist" (p. 56) after all.[12]

2.6 Brain-Modeling versus Mind-Modeling.

At this point Searle would perhaps interject that in holding out for an understanding of the "causal powers" of the real brain as opposed to the abstract principles of a computer model of mind he is himself opting precisely for the whole as opposed to the parts. That sounds reasonable, but the problem is that brains can (and must) be modeled too, if we are to come to understand them (and, in practice, the growing discipline of neural modeling is doing just that). Hence, to turn to the brain is not necessarily to turn away from computer simulation. Moreover, although according to the convergence argument the degrees of freedom (the number of alternative ways) for modeling the total capacity of a mechanism are much fewer than the degrees of freedom for modeling parts of it (equiparametrically), it is still true that the degrees of freedom of an implementation will always be more than those of a simulation,[13] if only because not all "causal powers" will be relevant (cf. a Boeing 707 versus a DC 10, a man versus a woman, me versus you). So there is likely to be more than one way to implement a human brain, perhaps even radically different ways, including ways that dispense altogether with the biochemical medium that Searle seems to find so compelling. Perhaps a better way of referring to the simulation vs. implementation distinction would be as simulation vs. "synthesis" . This is, after all, the science of "artificial" intelligence, and if we can synthesize hearts, lungs and blood with various different materials, why not brains and minds too? Hence if Searle wants to insist that he will be able to generalize his Chinese Room Argument to "simulate" any succesful simulation of the whole mind, then he will have to agree that he will likewise be able to "simulate" any successful simulation of the brain, and hence that both approaches are wrong-headed. (Recall that a successful simulation must always formalize the relevant causal principles that will make the implemented mechanism work.)

The root problem, I think, is this: For Searle, simulation is mere imitation, a kind of trivial and superficial aping of something, whereas in reality simulation is an extremely powerful means of developing a full theoretical understanding of a mechanism, indeed, a causal understanding of it. Even Searle's distinction between "Strong" and "Weak" AI is a questionable one. According to Searle, the proponents of "Strong AI" believe that (i) the mind is a program, (ii) the brain is irrelevant and (iii) the Turing Test is decisive. The proponents of "Weak AI" believe only that the computer is "a very useful tool in the study of mind." (Useful for what? one is inclined to ask, in the face of Searle's sceptical arguments!) I don't know about the beliefs of others, but I certainly can't fit my own readily into either camp. I do indeed believe that the computer is useful, whether it is modeling brain or mind. In neither case do I believe that the data (here, what neurons do; there, what people do) will explain "themselves" . Computer modeling is a very powerful way of testing hypotheses about how mechanisms work. We already know, generally speaking, what humans can do: They can perceive, speak, reason, understand, etc., but we have no idea how they do it -- we don't know the causal mechanism.[14] We also know a little about what neurons do: They produce action potentials and slow potentials; they secrete modulatory chemicals; they have complex connectivity, receptive fields, columnar organization, etc., and for some simple lower organisms we are even getting an idea of the causal mechanism of some modules or subsystems (Selverston 1980; Hoyle 1984).

There is not much cross-talk between these two levels of research (brain and mind), however. And although modelers in both areas would, I'm sure, be more than happy to accept and use whatever insights they can get -- from one another, or from anywhere else, for that matter -- there are not many insights forthcoming,[15] especially between brain and mind. So-called "bottom-up" neuroscientific data about vision and language (e.g., visual feature detectors, Hubel & Wiesel 1965, or linguistic deconnection syndromes, Geschwind 1965) have not turned out to be very helpful to AI workers attempting to model vision and language. Perhaps "top-down" cognitive information will prove more useful to neuroscientists, but it is too early to say. It is also much too early to judge whether the current enthusiasm for connectionistic modeling (McClelland et al. 1986; Smolensky 1988) has any deeper justification. Right now the optimism seems uncomfortably reminiscent of the ill-fated hopes for Perceptrons (Rosenblatt 1962 quashed by Minsky & Papert 1969), both in its projections and in its evidential basis. After all, at this point all we really have is a handful of demonstration experiments (the equivalent of AI's "toy" problems) and some structural and statistical properties (weighted positive and negative connections between units and rules for updating them) that may or may not resemble real neural properties at the right functional level. Inductively speaking, the grounds seem to be just as strong for assuming that the new connectionism will eventually turn out to be stunted by higher-order variants of the same kinds of limitations that prevented Perceptrons from living up to their promise as for assuming that connectionism will go on to succeed where Perceptrons failed. What are needed are (a) dramatic demonstrations of connectionism's performance capacity (preferably on life-size problems that are especially hard for symbolic AI, but feasible by the human mind) and/or (b) formal proofs of connectionism's functional capability. Moreover, those proofs must negotiate between the Scylla of inflating connectionism into just another Turing-powerful architecture for doing symbolic computation and the Charybdis of reducing it to a parochial class of statistical algorithms.

Yet the fact that connectionism is algorithmic and inductive rather than just an architecture for doing symbolic processing is an advantage, for it allows for the possibility of generality across problems, perhaps even general principles of learning. And the fact that the algorithms are not merely symbolic protects connectionism from some of the shortcomings of symbolic functionalism discussed here (and in Harnad 1987b). The ostensible neurosimilitude of connectionism, however, is probably more a liability than an asset at this stage (at least until (a) or (b) are forthcoming), because it allows performance weaknesses to be masked by putative brain-likeness, and brain-unlikenesses to be masked by performance strengths. Hence it will probably turn out to be methodologically sounder and more informative to evaluate the scope and limits of connectionism in cognitive modeling and in neural modeling quite "independently" .[16]

In practice, however, mutual irrelevance between neuroscience and cognitive science does not seem to be so much a matter of dogma as of fact in these two fields. Add to this the fact that the known performance data on the mind are much richer than those on the brain, and that brain-modelers are dealing with the smaller class of possible implementations of brains, whereas mind-modelers are concerned with the larger class of possible implementations of minds, of which real brains are only a subset: The upshot is that there are good reasons for expecting research paths to diverge, at least for the time being.[17]

Some convergence may come as both fields (neuroscience and cognitive science) move from their own respective "toy" problems to something approaching the totality of their respective mechanisms, for at that point the boundary-line between what a whole organism can do and what its brain can do may become fuzzy, as may the distinction between cognitive and noncognitive function. It is likewise only at that advanced stage, approaching convergence, that the (Total) Turing Test becomes pertinent; so for the time being that too fails to distinguish Strong from Weak AI.

2.7 Strong AI versus Weak AI.

As to the belief that the mind is a program: Consider my own beliefs, for example. I happen to believe that all of our cognitive capacities, conscious or otherwise, are the functions of a causal mechanism; I also believe that the best way to get to understand a mechanism is to try to model it. So far, we only have trivial toy models for tiny parts of what the cognitive mechanism can do (cf. Dennett 1978). I believe that with continued effort, more creative talent in cognitive science and many new and powerful cognitive principles, those toy models will grow (perhaps first into models of axolotls and aurochs), until they converge on an all-purpose model for all of our human capacities. It will be possible (in principle) to test that grand model, not only with a computer, but also with an abacus, a Chinese army, or a paper and pencil. The computer simulation will not have a mind. The implementation of that model as an actual mechanism, however, will pass the Total Turing Test, and we will have no better (or worse) grounds for denying that it has a mind than we have for denying that anyone else does (Harnad 1984). Does that make me a believer in Strong AI?

The successful performance of this complete simulation will also involve a lot more than look-up tables. The question of whether someone "simulating" that simulation could step through all of its functions without coming not only to understand its contents, but to understand understanding better than any of us do now, is an interesting, but not a particularly critical question. It is all, however, that seems to be left of Searle's Chinese Room argument.

3. Grounding Symbolic Function in Robotic Function

3.1 Syntax versus Semantics.

There are still some loose ends. Searle speaks of the problem of "syntax" versus "semantics." The semantics problem actually subdivides into two problems:

(1) How is it that symbols mean anything at all? How can they stand for, refer to, or represent objects and states of affairs in the world? This is also called the problem of "intentionality." Searle (1980b) notes that whereas the intentionality of human symbols is "intrinsic," that of machine symbols and other artifacts is "derived" from or parasitic on human intentionality. Nonhuman symbols only have meaning if they are so interpreted by people; otherwise they are just meaningless "squiggles." The problem of intentionality is related to what I have called the "symbol grounding problem" for symbolic functionalism (Harnad 1987b): With symbols defined only in terms of still more symbols, their meanings appear to be "ungrounded" . The problem is rather like trying to learn Chinese from a Chinese dictionary alone, without any prior knowledge of Chinese (or even of English or the sensory world).

(2) How is it that symbols are experienced as meaningful (in the way English is meaningful and Chinese

is not, to someone who does not understand Chinese)? This is related to the problem of how anything is subjectively experienced at all: the problem of "qualia," or consciousness (Nagel 1974; Harnad 1982b). What is the functional difference between ourselves and insentient robots that behave "exactly as if" they were conscious, as we are, but are not?

The answers to these questions will depend in part on future empirical and theoretical findings, but one can already anticipate that even the empirical answers will not be completely satisfying, for embedded in the questions are some philosophical issues that cannot be resolved empirically. My contention is that the undecidable problems can be separated from the decidable ones, and that when this is done, they do not represent an obstacle to any empirical research program. In particular, Searle's Chinese Room Argument should not daunt AI. However, if Searle's argument is recast, as it is in this paper, as an attack on "symbolic functionalism" (the hypothesis that mental function consists only of formal symbol-manipulation) in favor of "robotic functionalism" (the hypothesis that nonsymbolic functions are critically involved in mental states), then the syntax/semantics problem can be seen as a potential obstacle to certain approaches to AI: The "top-down" approaches that assume that symbol-manipulation is an autonomous function that can successfully implement human mental function without being grounded in nonsymbolic function (e.g., Fodor 1981; Pylyshyn 1984).

Ever since the beginning of computer modeling, people have been concerned that, after all, computers only treat uninterpreted formal symbols (0 and 1, for the most part). How can they ever understand what anything "means" ? Answer: Computers don't understand. Computer programs don't understand (though they can in principle model formally all of the relevant causal factors in understanding). Only implemented mechanisms that can pass the Total Turing Test -- i.e., respond to all of our inputs indistinguishably from the way we do -- can understand.

At the very least, to make a computer (which is, after all, a mechanism, and will surely be a part, perhaps several parts, of the ultimate hybrid device that does successfully pass the Total Turing Test) into a candidate for a person you would have to give it something like sense organs: If the computer is to see and hear as we do, there must be some way for it to turn visual and auditory stimulation into code. If you still regard the computer-plus-sensory-transducers as merely a syntactic device, then surely we are only syntactic devices too, since light just produces code at the retina and sound only produces code at the cochlea. But there are codes and there are codes, and not all of them are symbolic. A symbolic code is a set of physical tokens that are manipulated in virtue of their (arbitrary) form according to certain formal rules; the relation between the tokens and what they "stand for" is dependent on an interpretative convention or notational system (rather like encryption and decryption in computation and cryptography), i.e., the symbolic token/object relation is "derived," in Searle's sense. A nonsymbolic code is one in which the relation between the symbol tokens and what they stand for is not arbitrary or conventional, but governed by physics in some way, such as through reliable causal connections between similar physical properties such as shape. The nonsymbolic token/object relation is "intrinsic."[18] The neural code may be partly analog, that is, it may preserve the "shape" of the input fairly faithfully (and, as discussed in Harnad 1987b, there are reasons to expect that this may turn out to be an important feature of successful models), but it's still code: It's never the "thing itself" that participates in an internal state, only its sensory code.[19] So, if there is any real semantics in the mind (and there surely is), it must derive, at least in part, from causal interactions with the outside world.

Perhaps semantics also derives in part from the existence of internal hierarchical levels of information-processing, the uninterpreted elements of one level becoming the elementary patterns of another: Most cognitive processes will of course turn out to be accomplished below the level of consciousness altogether (Harnad 1982b). And even conscious processes will not occur at a level that would allow them to be resolved into elementary, uninterpreted data-structures (cognitive "grain" will be blurred). Some inputs will just be hard-wired for meaning: For primates, (the sensory representation of) a snake innately "means" something you're afraid of and want to get away from. Other inputs will derive their meaning from experience. Note that I have no illusions at all about cognitive science's having provided (or even being likely to provide) any insight into what consciousness and meaning "are".[20] It simply holds out the promise of showing us what mechanistic principles will generate their outward manifestations (cf. Nagel 1986), in forms indistinguishable from those exhibited by other people -- "and brain science can promise no more" . This is the methodology and the logic of the Total Turing Test.

What, after all, gives the sequence of squiggles "2 + 2 = 4" its meaning for us? Is it not (or was it not, before it became automatic) "interpreting" the numbers as objects (two apples), operating on these objects (adding two apples to two other apples), confidence (from experience) in freely repeatable actions (adding still more apples) and, later, confidence in formal proofs, in which the primitive terms and operations likewise get concrete interpretations, and higher-order abstractions take lower ones as primitives? Even (the sensory representation of) an apple is just a class of input data, partly, perhaps, hard-wired for approach as food, but largely something whose place in our classification hierarchies is learned from experience, manipulation, naming, categorizing and describing (Harnad 1987c).

I'm certainly not going to solve the problem of how raw data become interpreted here, but I do want to point out that Searle's claim that the interpretation can only come from a programmer is only correct for an all-purpose computer, not for the kind of "dedicated" mechanism that the all-purpose implementation we are concerned with here demands. Once a computer is hard-wired to sensory transducers and motor effectors or to any other specialized peripheral devices, certain "interpretations" are thereby fixed. The mechanism that successfully passes the Total Turing Test is likely to be an integrated hybrid system consisting of many specialized modules, hardware and peripherals (possibly, though not necessarily, including biochemical components) along with many generalized computers: The notion of just an all-purpose VAX, taking nothing but 0's and 1's, is totally inappropriate for this implementation (but not for simulating it). The functional states underlying mental states in such a device would accordingly not be merely symbolic either.[21]

3.2 Computation and Cognition.

Note that there is a reason for conjecturing that computers will be components in the Utopian mechanism that will pass the Total Turing Test. One thing distinguishing the simulation of flying from the simulation of cognitive activity is that computation, or formal symbol manipulation, which is the medium of the simulation in both cases, is not being proposed as one of the implemented functions in the case of flying (except perhaps in modern, computerized "smart" flying), whereas it is in the case of cognition. There is, in other words, reason to suspect that there may be some formal similarity between thinking and computation, but not between flying and computation. So some of the processes in mind-modeling may indeed be computations -- not just simulated by computation, but computations themselves. However, the actual extent and specifics of computation as an implemented physical process in cognition as opposed to computation as just a theoretical medium for simulating cognition -- that is, the question of how much of the ultimate implemented mechanism will consist of all-purpose computer modules doing symbol-manipulation, rather than specialized and dedicated devices -- is a matter for future research.

3.3 The Teletype Turing Test versus the Robot Turing Test: The Modularity Assumption.

Some words also need to be said about that cornerstone of mind-modeling, the Turing Test. First, it should be made clear that Turing's test has nothing to do with "proof" (cf. Searle 1982a, p. 3). It is only a thesis. If a mechanism fails the Turing Test, it need not be assigned a mind. If it passes, the success is only provisional, partly for the same reason that the success of any scientific hypothesis is provisional (the mechanism may fail later), but also partly because any conjecture that tries to capture our intuitions can only be provisional. (Success on the test may fail to convince us, or a better test of our intuitions may be found.) Right now, the production of behavior that is indistinguishable from our own seems to be our only objective basis for assigning a mind to anyone else but ourselves under any circumstances -- whether we are Turing-Testing in the lab, reality-testing in the world, or just engaging in pure cartesian contemplation.

How much physical appearance is allowed to vary in Turing-Testing while still deserving the benefit of the doubt is a fuzzy matter. Some of us (not this writer) are sceptical about animals' minds, but of course animals differ from us not only in appearance but also in behavior. For one thing, they can't talk, and we seem to set a lot of intuitive store by the medium of language (Harnad et al. 1976). In any case, it seems intuitively clear that it is largely external appearance that matters. Since I know so little about how our bodies work anyway, I hardly think that I would deny consciousness to Hungarians if on the inside they all turned out to be made of some radically different stuff, perhaps even transistors. Even being informed that they were man-made (in the engineering sense) would strike me as an arbitrary reason for suddenly revising my beliefs about their mental lives, particularly if they continued to be completely indistinguishable from other people (except, of course, as Hungarians). It seems that Searle's intuition, on the other hand, is conditional on whether or not the engineering in question happens to turn out to be biochemical/genetic. For my own part, my ignorance and lack of insight into either the biochemical or the electronic basis of mind is enough to keep such considerations from influencing my intuitions.

A more serious intuitive consideration, however, is a rather radical variant of the question of appearance. This concerns the teletype version versus the full robot version of the Turing Test. It seems that we assign so much weight to the capacity to communicate linguistically that we feel just about as confident in assigning minds on the basis of written communication as we do in assigning them on the basis of face-to-face contact. (Consider the people who correspond as pen-pals for years and years and never get to see one another -- sometimes not even a photo -- and yet never feel inclined to doubt one another's mentality.) It may be that language captures the full expressive power of our behavior, at least insofar as our inclinations to assign minds are concerned.

What remains an open question is whether language is an independent enough module to be modeled successfully without the necessity to model our full robotic (sensory/motor) capacities as well. The convergence argument (section 2.5) suggests that this is unlikely. Evolution also seems to suggest that our robotic capacities precede our linguistic ones, because many organisms have the former but not the latter, whereas no organism has the latter but not the former. So although it may not take the full robot version of the Turing Test to convince a person, it may take something closer to it -- say, the short-circuited central components and capacities of the implemented mind-in-a-vat -- in order to get the candidate model to work in the first place.

3.4 Robotics and Causality: The Transducer Counterargument.

A great deal depends on the question of what it would actually take to get a model to perform Turing-indistinguishably from ourselves. Note that one can speak of two kinds of causality: real or physical causality (p-causality) and symbolic or formal "causality" (f-causality). F-causality is coded in symbols tokens: "If Input = 1, make Output = 0" is as close as a formal computer model ever comes to causality. Such commands are implemented by electronic circuits and circuit logic, which are, of course, p-causal, but what they encode ("If the temperature goes below 65, turn on the furnace," or "If you see a snake, run") is just f-causality. Only physical mechanisms (thermostats, robots) can implement this formalism and make it real. It is hence at the interface with the outside world that f-causality becomes p-causality (and causal "powers" become total input/output performance capacity). This is why sensory/motor transducers/effectors and robotic interactions with the environment are so important. It is also why the question of the teletype versus the robot version of the Total Turing Test is such a crucial one: because (objectively speaking) "language too is just formal" (symbolic). So if we allowed all inputs and outputs to a mechanism to be just symbolic, then there would be no reason everything in between couldn't be just symbolic (computational) as well -- in which case Searle's scepticism (and my own) might be justified.

Nor is it at all clear that there are only symbols between transducers and effecters. In real brains there seems to have been a need for multiple, internal analog re-presentations of sensory receptor surfaces (Lieblich & Arbib 1982). This means that environmental interfaces are hard to localize, being partly internal, perhaps partly even indeterminate. How true this will have to be for the ultimate total cognitive mechanism -- or its disembodied "mind-in-a-vat" core -- is a matter for future research. A case can be made that symbols tokens, manipulable only in virtue of their (arbitrary) form, must be grounded in the shape of the real world through sensorimotor processes in order to be meaningful (see Harnad 1987b).

Note that the simulation/implementation distinction already points to the critical status of transduction, since Searle's Chinese Room Argument fails completely for the robot version of the Turing Test, when the corresponding mental property at issue is the perception of objects rather than the understanding of symbols. To see this, note that the terms of the Argument require Searle to show that he can take over all of the robot's functions (thereby blocking the Systems Reply) and yet clearly fail to exhibit the mental property in question, in this case, perceiving objects. Now consider the two possible cases: (1) If Searle simulates only the symbol manipulation between the transducers and effectors, then he is not performing all the functions of the robot (and hence it is not surprising that he does not perceive the objects the robot is supposed to perceive). (2) If, on the other hand, Searle plays homunculus for the robot, himself looking at its scene or screen, then he is being its transducers (and hence, not surprisingly, actually perceiving what the robot is supposed to perceive). A similar argument applies to motor activity.[22] Robotic function, unlike symbolic function, is immune to Searle's Chinese Room Argument.

Perhaps Searle's Chinese Room Argument should be taken as evidence against symbolic functionalism (also known as the "computational theory of mind"; Fodor 1981; Pylyshyn 1984) rather than against "Strong AI." According to symbolic functionalism, only computation -- formal, language-like symbol-manipulation -- counts as cognition.[23] This computational theory, like any theory, may well prove to be wrong, or, more precisely, radically incomplete. It certainly does not have dazzling empirical successes to its credit at this point. Hence, rather than prejudging what counts as cognitive on the strength of a candidate theory, it may be more prudent (as suggested in Harnad 1982a) to insist only that cognitive theories be computable (simulable) rather than necessarily computational (symbol-manipulating). Could that be what Searle means by "Weak AI"?

For my own part, if it did turn out that just a program running on a VAX could indeed pass the teletype version of the Total Turing Test, I think my intuitive loyalties would be almost evenly divided (between my scepticism about computers, no matter how they are configured by their software, and my confidence in the Teletype Total Turing Test) and I might just decide to reject the teletype version in favor of the robot version. -- But I wouldn't necessarily be right in this, of course; so, rather than contemplating hypothetical confrontations between irresistible forces and immovable objects, I think I'd rather just wait to see whether such a simulation is possible before committing myself to any intuitive contingency plans.

4. Summary and Conclusions

Searle's provocative "Chinese Room Argument" attempted to show that the goals of "Strong AI" are unrealizable. Proponents of Strong AI are supposed to believe that (i) the mind is a computer program, (ii) the brain is irrelevant, and (iii) the Turing Test is decisive. Searle's argument is that since the programmed symbol-manipulating instructions of a computer capable of passing the Turing Test for understanding Chinese could always be performed instead by a person who could not understand Chinese, the computer can hardly be said to understand Chinese. Such "simulated" understanding, Searle argues, is not the same as real understanding, which can only be accomplished by something that "duplicates" the "causal powers" of the brain. In the present paper the following points have been made:

1. Simulation versus Implementation:

   Searle fails to distinguish between the simulation of a mechanism, which is only the formal testing of a theory, and the implementation of a mechanism, which does duplicate causal powers. Searle's "simulation" only simulates simulation rather than implementation. It can no more be expected to understand than a simulated airplane can be expected to fly. Nevertheless, a successful simulation must capture formally all the relevant functional properties of a successful implementation.

2. Theory-Testing versus Turing-Testing:

Searle's argument conflates theory-testing and Turing-Testing. Computer simulations formally encode and test models for human perceptuomotor and cognitive performance capacities; they are the medium in which the empirical and theoretical work is done. The Turing Test is an informal and open-ended test of whether or not people can discriminate the performance of the implemented simulation from that of a real human being. In a sense, we are Turing-Testing one another all the time, in our everyday solutions to the "other minds" problem.

3. The Convergence Argument:

   Searle fails to take underdetermination into account. All scientific theories are underdetermined by their data; i.e., the data are compatible with more than one theory. But as the data domain grows, the degrees of freedom for alternative (equiparametric) theories shrink. This "convergence" constraint applies to AI's "toy" linguistic and robotic models too, as they approach the capacity to pass the Total (asymptotic) Turing Test. Toy models are not modules.

4. Brain Modeling versus Mind Modeling:

   Searle also fails to appreciate that the brain itself can be understood only through theoretical modeling, and that the boundary between brain performance and body performance becomes arbitrary as one converges on an asymptotic model of total human performance capacity.

5. The Modularity Assumption:

   Searle implicitly adopts a strong, untested "modularity" assumption to the effect that certain functional parts of human cognitive performance capacity (such as language) can be be successfully modeled independently of the rest (such as perceptuomotor or "robotic" capacity). This assumption may be false for models approaching the power and generality needed to pass the Turing Test.

6. The Teletype Turing Test versus the Robot Turing Test:

   Foundational issues in cognitive science depend critically on the truth or falsity of such modularity assumptions. For example, the "teletype" (linguistic) version of the Turing Test could in principle (though not necessarily in practice) be implemented by formal symbol-manipulation alone (symbols in, symbols out), whereas the robot version necessarily calls for full causal powers of interaction with the outside world (seeing, doing and linguistic competence).

7. The Transducer/Effector Argument:

   Prior "robot" replies to Searle have not been principled ones. They have added on robotic requirements as an arbitrary extra constraint. A principled "transducer/effector" counterargument, however, can be based on the logical fact that transduction is necessarily nonsymbolic, drawing on analog and analog-to-digital functions that can only be simulated, but not implemented, symbolically.

8. Robotics and Causality:

   Searle's argument hence fails logically for the robot version of the Turing Test, for in simulating it he would either have to use its transducers and effectors (in which case he would not be simulating all of its functions) or he would have to be its transducers and effectors, in which case he would indeed be duplicating their causal powers (of seeing and doing).

9. Symbolic Functionalism versus Robotic Functionalism:

   If symbol-manipulation ("symbolic functionalism") cannot in principle accomplish the functions of the transducer and effector surfaces, then there is no reason why every function in between has to be symbolic either. Nonsymbolic function may be essential to implementing minds and may be a crucial constituent of the functional substrate of mental states ("robotic functionalism"): In order to work as hypothesized (i.e., to be able to pass the Turing Test), the functionalist "brain-in-a-vat" may have to be more than just an isolated symbolic "understanding" module -- perhaps even hybrid analog/symbolic all the way through, as the real brain is, with the symbols "grounded" bottom-up in nonsymbolic representations.

10. "Strong" versus "Weak" AI:

    Finally, it is not at all clear that Searle's "Strong AI"/"Weak AI" distinction captures all the possibilities, or is even representative of the views of most cognitive scientists. Much of AI is in any case concerned with making machines do intelligent things rather than with modeling the mind.

Hence, most of Searle's argument turns out to rest on unanswered questions about the modularity of language and the scope and limits of the symbolic approach to modeling cognition. If the modularity assumption turns out to be false, then a top-down symbol-manipulative approach to explaining the mind may be completely misguided because its symbols (and their interpretations) remain ungrounded -- not for Searle's reasons (since Searle's argument shares the cognitive modularity assumption with "Strong AI"), but because of the transdsucer/effector argument (and its ramifications for the kind of hybrid, bottom-up processing that may then turn out to be optimal, or even essential, in between transducers and effectors). What is undeniable is that a successful theory of cognition will have to be computable (simulable), if not exclusively computational (symbol-manipulative). Perhaps this is what Searle means (or ought to mean) by "Weak AI."

FOOTNOTES

[1] "My own notion of what constitutes [the core of cognitive science] could be summed up in the following way: it consists of a careful and detailed explanation of what's really silly about Searle's Chinese room argument" (Hayes, p. 2, in Lucas & Hayes 1982).

[2] Searle (1980a) has dubbed the various prior replies to his argument the "Systems Reply," the "Robot Reply," etc. One is tempted to call this one the "Total Reply," for reasons that will become apparent. Lest this be misunderstood as imperialism, however, perhaps it should rather just be referred to as the "Robotic Functionalist Reply."

[3] Something of this sort is actually being done these days in optimizing advanced airplane and rocket designs prior to actually building anything.

[4] There are other ways; and there are deeper subtleties to "understanding" a mechanism, such as the problem of unexplicated successes in fortuitous simulations, complex analog simulations, or connectionistic nets (McClelland et al. 1986; Rumelhart et al. 1986) that "relax" into successful configurations after a series of trials. In such cases the causality is not fully understood despite the successful performance of the model. Also, unlike in cognitive science, which seems to be a branch of theoretical engineering, causal theories in physics are not exactly recipes for building a universe. These fine points, however, are not critical for the issues under consideration here.

[5] And nontrivial: Searle's observation that "(e)verything, by the way, instantiates some program or other. . . in [a] trivial sense" (1982a, p. 6) has the mark of the armchair theorist who has never had to worry about how to get a simulation to actually work successfully. The fact that any functional state of any mechanism "instantiates" some Turing machine state ("Turing equivalence") is -- like the fact that any state-of-affairs is describable by some set of sentences (Steklis & Harnad 1976; Katz 1976) -- compelling testimony to the power and generality of computation and of natural language, respectively. In both cases, however, the nontrivial trick is to come up with the actual state description: a creative, theoretical feat. Anything else is just nonconstructive hand-waving (Harnad 1982a).

[6] If the furnace too is simulated, then this becomes what is known as a "negative feedback loop."

[7] The states of the circuits -- on or off, "0" or "1" -- can be interpreted in many ways by the programmer.

[8] Causality and "dedicated" computer systems are discussed later, in section 3.1; see also Harnad 1982b.

[9] Among the far-fetched consequences of the hardware fallacy are the irrelevance of neuroscience and the suzerainty of symbol manipulation.

[10] Even the restriction to telecommunication makes the test somewhat unrealistic -- a robot version would be more decisive and convincing, and may well turn out to be the only one that could actually succeed; but that question will be deferred till section 3.

[11] The unreasonableness of Searle's expecting an isolated thirst module alone to feel thirsty (1982a, p. 4) seems too obvious to require an answer.

[12] "Caused-by-and-realized-in" sounds in any case like a portmanteau expression roughly equivalent to claiming that you can have-your-cake-and-eat-it-too, just as long as you say it fast enough to sound like a unitary operation. The mind/body problem can't be solved quite as quickly as that.

[13] I am here using "simulation" and "model" interchangeably.

[14] As suggested earlier, the basic empirical facts about our cognitive capacities are already familiar. The problem now seems to be to resist the temptation to keep gathering more and more performance data on their fine tuning (like the proverbial drunkard who keeps looking for his keys near the light) instead of tackling the daunting problem of finding viable causal explanations of how we (or any mechanism) can do what we already know we can do. (A similar point has apparently been made by Clowes as cited by Johnson-Laird; see interview by Groeger 1987.)

[15] Searle may wonder why, given my realism about the primitiveness of the current state of the art in computer modeling, I am nevertheless so sanguine about its future prospects. The answer is that computation and mechanism are the only game in town; so -- until someone invents another one, or at least offers some evidence that there can even be another one -- that's the way I'm betting. The new kid on the block, connectionism (McClelland et al 1986; Smolensky 1988), does not seem to be a rival game, by the way, but a particular candidate mechanism, in the form of a family of statistical algorithms for incremental pattern learning. Its functional potential, like the potential of any other formal class of algorithms, is being tested by simulation as well as formal analysis.

[16] Abstaining from neuralistic interpretations until connectionism's performance capacities are ascertained (and vice versa) would probably be as helpful in keeping connectionism honest (i.e., minimizing unwarranted claims and projections) as abstaining from mentalistic interpretations altogether would be in the case of symbolic modeling.

[17] It is an oversimplification, however, merely to baptize the two approaches "bottom-up" and "top-down," in the pious hope that convergence will consist of a meeting in between. There are some very different forms of bottom-upism (e.g., psychophysical vs. psychophysiological) as well as top-downism (e.g., symbolic vs. robotic functionalism) with, presumably, differing respective prospects of converging at Utopia (see Harnad 1987a).

[18] This definition is nonstandard, and has some surprising consequences. For example, a program that is running on a digital computer and that depends for its interpretation on a human programmer would be symbolic according to this definition, whereas the very same program running on a dedicated, implemented device, hard-wired to its transducers and effectors, would not be symbolic (or not purely symbolic), because its interpretations would be physically "fixed" by the causal input/output system as a whole. The definition is consistent, however, with more standard definitions of the symbolic/nonsymbolic

distinction (Pylyshyn 1984, pp. 147-191); it merely follows through what the more standard definitions logically entail with respect to transducer/effector, dedicated, and analog functions, namely, that the symbolic/nonsymbolic nature of the symbol/object relation depends on causal scale (e.g., on whether or not one includes human beings in the causal system under consideration) and that the analog/digital distinction depends on the degree of of invertibility of the causal connection between object and symbol-token (see Harnad 1987b).

19. Perhaps any information-processing mechanism (with human-scale capacities) is consequently bound to have a mind/body problem, and to be sceptical about whether there are "really" objects and other minds out there behind its sense data -- always yearning for the "real" semantics behind its phenomenal syntax; cf. Fodor 1980).

20. Note also that the two aspects of the problem of semantics defined earlier -- (a) the connection between the symbol and its intended object (i.e., the intentionality problem) and (b) what it's like to use a symbol to intend an object (i.e., the qualia problem) -- may not be separable. Would there really exist an autonomous "intentionality" problem for devices that had no qualia (i.e., is there anything left of the problem of meaning if there's nothing it's "like" to mean something)? And, conversely, if we could somehow know that we had designed devices that really had qualia, would we still worry independently about whether we had captured the right "aboutness" relation? It would seem that if intentionality is "derived" from anything, it is derived from qualia: Propositional content derives from experiential content.

21. We must carefully distinguish our (perhaps justifiable) intuitive reluctance to equate our mental states with uninterpreted symbolic states from our (irremediable) intuitive inability to equate them with any physical state at all (Harnad 1982b; Nagel 1986). The latter is also known as the mind/body problem.

22. Recall that transducer/effector function can be simulated, but not implemented, computationally (2.2).

23. Symbolic functionalism defines cognitive science on the strength of one particular theoretical approach: sentential representation. It seems to be motivated by (i) the success of generative-transformational models in linguistics (Chomsky 1980); (ii) the success of computational "toy" models in AI; (iii) the power and generality of language and computation (which inclines theorists of this persuasion [iii'] to conflate sentences and the states-of-affairs they describe, [iii''] to assume that because everything is symbolically representable, all representations must be symbolic, and [iii'''] to overlook the fact that even if a token is symbolically interpretable, it may have a [nonsymbolic] causal role that fixes or "dedicates" its interpretation); (iv) arguments for the essentially inferential nature of cognition and for the existence of a language-like code ("mentalese," Fodor 1975) underlying all cognitive processes; (v) arguments against mental imagery and analog "re-presentations" (Pylyshyn 1973); (vi) the apparent uniqueness and naturalness of the computational as opposed to the physical level of explanation in capturing mental concepts such as "belief" and "knowledge"; (vii) emphasis on a (somewhat circular) criterion for cognition called "cognitive penetrability" -- roughly, that beliefs and knowledge affect beliefs and knowledge (and behavior) -- which likewise seems well-captured at the computational level (Pylyshyn 1980).

REFERENCES

Abelson, R.P. (1980) Searle's argument is just a set of Chinese symbols. *Behavioral and Brain Sciences* 3: 424-425.

Block, N. (1980) What intuitions about homunculi don't show. *Behavioral and Brain Sciences* 3: 425-426.

Carleton, L. (1984) Programs, language understanding and Searle. Synthese 59: 219-230.

Chomsky, N. (1980) Rules and representations. *Behavioral and Brain Sciences* 3: 1-62.

Dennett, D. C. (1978) Why not the whole iguana? *Behavioral and Brain Sciences* 1: 103-104.

Dennett D. C. (1981) Where am I? In: Hofstadter, D. R. & Dennett, D. C. *The mind's I: Fantasies and reflections on mind and soul* New York: Basic Books, pp. 217-229.

Dennett, D.C. (1982) The myth of the computer: An exchange. N.Y. Review Books: XXIX (11): 56.

Donchin, E. (Ed.) (forthcoming) Proceedings of the 2nd Carmel Conference on Philosophical Aspects of Event-Related Potentials.

Edelson, T. (1982) Simulating understanding: Making the example fit the question. *Behavioral and Brain Sciences* 5: 338-339.

Fodor, J.A. (1975) *The language of thought* . N.Y.: T.Y. Crowell.

Fodor, J.A. (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3: 63-109.

Fodor, J. A. (1981) RePresentations. Cambridge MA: MIT/Bradford.

Fodor, J. A. (1985) Précis of *The Modularity of Mind. Behavioral and Brain Sciences* 8: 1-42.

Geschwind, N. (1965) Disconnection syndromes in animals and man. Brain: 88: 237-295; 585-644.

Groeger, J.A. (1987) Computation - the final metaphor? An interview with Philip Johnson-Laird. *New Ideas in Psychology* 5: 295-304.

Hanna, P. (1985) Causal powers and cognition. Mind XCIV: 53-63.

Harnad, S. R., Steklis, H.D. & Lancaster, J. B. (Eds.) (1976) Origins and evolution of language and speech. *Annals of the New York Academy of Sciences* 280.

Harnad, S. (1976) Induction, evolution and accountability. *Annals of the N.Y. Academy of Sciences* 280: 58-60.

Harnad, S. (1982a) Neoconstructivism: A unifying theme for the cognitive sciences. In: T.W. Simon & R.J. Scholes (Eds.), *Language, mind and brain* Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 1-11.

Harnad, S. (1982b) Consciousness: An afterthought. *Cognition and Brain Theory* 5: 29-47.

Harnad S. (1984) Verifying machines' minds. *Contemporary Psychology* 29: 389-391.

Harnad, S. (1987a) Psychophysical and cognitive aspects of categorical perception: A critical overview. In: *Categorical perception: The groundwork of cognition* (S. Harnad, Ed.) Cambridge: Cambridge University Press, pp. 1-25.

Harnad, S. (1987b) Category induction and representation. In: *Categorical perception: The groundwork of cognition* (S. Harnad, Ed.) Cambridge: Cambridge University Press, pp. 535-565.

Harnad S. (Ed.) (1987c)

Categorical perception: The groundwork of cognition.

NY: Cambridge University Press.

Harvey, R. J. (1985) On the nature of programs, simulations and organisms. *Behavioral and Brain Sciences* 8: 741-2.

Hoyle, G. (1984) The scope of neuroethology. *Behavioral and Brain Sciences* 7: 367-412.

Hubel, D.H. & Wiesel, T.N. (1965) Receptive fields and functional architecture in two nosntriate areas (18 and 19) of the cat. *Journal of Neurophysiology* 28: 229-289. Katz, J.J. (1976) Effability: A hypothesis about the uniqueness of natural language. *Annals of the New York Academy of Sciences* 280: 33-41.

Lieblich, I. & Arbib, M.A. (1982) Multiple representations of space underlying behavior. *Behavioral and Brain Sciences* 5: 627-659.

Lucas, J. R. (1961) Minds, machines and Gödel. Philosophy 36: 112-117.

Lucas, M. M. & Hayes, P. J. (Eds.) (1982) *Proceedings of the Cognitive Curricula Conference.* University of Rochester: Rochester NY

McClelland, J.L., Rumelhart, D. E., and the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition,* Volume 1. Cambridge MA: MIT/Bradford.

McDermott, D. (1982) Minds, brains, programs and persons. *Behavioral and Brain Sciences* 5: 339-341.

Minsky, M. & Papert, S. (1969) *Perceptrons: An introduction to computational geometry.* Cambridge MA: MIT Press

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83: 435-451.

Nagel, T. (1986) *The view from nowhere.* New York: Oxford University Press.

Putnam, H. (1975) *Mind, language and reality.* New York: Cambridge University Press.

Pylyshyn, Z. W. (1973) What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* 80: 1-24.

Pylyshyn, Z. W. (1980) Computation and cognition: Issues in the foundations of cognitive sciences. *Behavioral and Brain Sciences* 3: 111-169.

Pylyshyn, Z. W. (1984) *Computation and cognition.* Cambridge MA:MIT/Bradford.

Rey, G. (1986) What's really going on in Searle's "Chinese Room"? *Philosophical Studies* 50: 169-185.

Rosenblatt, F. (1962) *Principles of neurodynamics.* Washington DC: Spartan. & Rougeul-Buser (1978, 215 - 232).

Rumelhart, D. E., McClelland, J.L., and the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition,* Volume 2. Cambridge MA: MIT/Bradford.

Russow, L.-M. (1984) *Nature & System* 6: 221-227.

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

Searle, J. R. (1980b) Instrinsic intentionality. *Behavioral and Brain Sciences* 3: 450-457.

Searle, J. R. (1982a) The Chinese room revisited. *Behavioral and Brain Sciences* 5: 345-348.

Searle, J. R. (1982b) The myth of the computer. *New York Review of Books* XXIX(7): 3-7.

Searle, J. R. (1982c) The myth of the computer: An exchange. *New York Review of Books* XXIX(11): 56-57.

Searle, J. R. (1985a) Patterns, symbols and understanding. *Behavioral and Brain Sciences* 8: 742-743.

Searle, J. R. (1985b) *Minds, brains and science.* Cambridge MA: Harvard University Press.

Selverston, A. I. (1980) Are central pattern generators understandable? *Behavioral and Brain Sciences* 3: 535-571.

Slezak, P. (1982) Goedel's theorem and the mind. *British Journal for the Philosophy of Science* 33: 41-52.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11: 1-XX.

Steklis, H. D. & Harnad, S. R. (1976) From hand to mouth: Some critical stages in the evolution of language. *Annals of the New York Academy of Sciences* 280: 445-455.

Turing, A. M. (1964) Computing machinery and intelligence. In: *Minds and machines,* A. R. Anderson (ed.), Engelwood Cliffs NJ: Prentice Hall.

Wilensky, R. (1980) Computers, cognition and philosophy. *Behavioral and Brain Sciences* 3: 449-450.