Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. *Think* 2: 12-78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.).

GROUNDING SYMBOLS IN THE ANALOG WORLD WITH NEURAL NETS
A Hybrid Model

Stevan Harnad
Department of Psychology
Princeton University
Princeton NJ 08542
harnad@princeton.edu

1.0 COMPUTATIONALISM VS. CONNECTIONISM IN COGNITIVE MODELLING

1.1 The predominant approach to cognitive modeling is still what has come to be called "computationalism" (Dietrich 1990, Harnad 1990b), the hypothesis that cognition is computation. The more recent rival approach is "connectionism" (Hanson & Burr 1990, McClelland & Rumelhart 1986), the hypothesis that cognition is a dynamic pattern of connections and activations in a "neural net." Are computationalism and connectionism really deeply different from one another, and if so, should they compete for cognitive hegemony, or should they collaborate? These questions will be addressed here, in the context of an obstacle that is faced by computationalism (as well as by connectionism if it is either computational or seeks cognitive hegemony on its own): The symbol grounding problem (Harnad 1990).

1.2 First, a little precision, so we make sure we are all talking about the same thing. By "computation" I mean symbolic computation: the manipulation of physical "symbol tokens" on the basis of syntactic rules that operate only on the "shapes" of the symbols (which are arbitrary in relation to what the symbols can be interpreted as meaning), as in a digital computer or its idealization, a Turing Machine manipulating, say, "0's" and "1's." What I will say about symbolic computation does not apply to analog "computation" or to analog systems in general, whose activity is best described as obeying differential equations rather than implementing symbol manipulations (Newell 1980; Pylyshyn 1984; but see McLennan 1987, 1988, in press-a, in press-b, for another view of analog computation).

1.3 Another important feature of computation is that it is implementation independent: Whatever properties or capabilities a system has purely in virtue of its computational properties will be shared by every implementation of that symbol system (computer program), no matter how radically the implementations may differ physically: All the specifics of the implementation are irrelevant except the fact that it implements that particular symbol system.

1.4 The last and most important property of computation (or of nontrivial computation, if we include uninterpretable formal gibberish as a trivial form of computation) is that the symbols and symbol manipulations in a symbol system are *systematically interpretable* (Fodor & Pylyshyn 1988): they can be assigned a semantics, they mean something (e.g., numbers, words, sentences, chess moves, planetary motions, etc.).

1.5 Another point to bear in mind in this discussion is that, as a cognitive psychologist, I am interested in machine intelligence only insofar as it can be taken as a substantive model for human (and other organisms') intelligence. There is a great utility in getting machines to do clever and useful things for their own sake, but that is not my field. So some of the arguments I will raise against certain systems as cognitive models will not necessarily be arguments against their utility as machines. The word "intelligence" is equivocal (describing, as it does, both human intelligence and clever and useful performance by machines, performance that ordinarily requires human intelligence); since my interest is only in homologies, not analogies, I will avoid the word "intelligence" altogether (cf. Turing 1964). However, I believe that some of the points I will be making below about the need to ground symbols may be relevant not only for cognitive modeling, but also for machine intelligence in general: Certain forms of performance may be unattainable from "ungrounded" symbol systems.

2.0 ARE NEURAL NETS JUST COMPUTATIONAL MODELS?

2.1 Having defined computation as implemented but implementation-independent syntactic symbol systems, we must ask first whether neural nets fall under this definition. The answer is a little complicated, but I hope it sorts the alternatives fairly clearly; it turns out to depend both on how the net is used and on how it is implemented:

2.1.1 (SIM) Most neural nets in the literature are actually computational simulations of neural nets: they are not real, parallel, distributed activation states in a set of physically interconnected nodes. They are simply serial, symbolic simulations of the properties of such nets; in other words, they are just implemented symbol systems that are *interpretable* as neural nets: "virtual" nets. As such, since whatever the nets can do, the net simulations can do, the capabilities of nets are really just the capabilities of symbol systems (computation). It is noteworthy that the learning algorithms of some neural nets don't even need to be thought of as nets: A parallel, interconnected system just happens to be one way of implementing them. Construed thus, it is clear that neural nets would just be a family of symbolic (including numerical) learning algorithms. We will return to this point.

2.1.2 (IMP) Another way of construing neural nets that would make comparing nets' capabilities with those of symbol systems like the comparison between "abstract" and "concave" would be to view them merely as as the hardware for implementing symbol systems: Computation is implementation-independent, so whether the hardware it's run on is a Sparc or a neural net does not matter if only its computational properties matter. (I think that the proofs that neural nets can be used as Turing Machines may fall in this category, e.g., Touretzky 19XX, although there is an important difference between using a net as hardware to do symbolic computation and training nets, as nets, to become symbol systems; Harnad 1990d.)

2.1.3 (PAR) In SIM, a symbol system simulates a neural net; in IMP, a neural net implements a symbol system. There remains the third possibility that some of the properties and capabilities of neural nets depend *essentially* on their being implemented in a physically parallel, interconnected and distributed form. To the extent that this is true, and to the extent that nets make *essential* use of analog input and throughput in being able to do what they can do, neural nets would be true alternatives to symbol systems, rather than just special cases of symbol systems or mere (irrelevant) media for implementing symbol systems.

2.2 Ironically, although the distinctions among SIM, IMP and PAR are important in sorting out the polemics about what symbols can do that nets can't and vice versa (e.g., Fodor & Pylyshyn 1988; Harnad 1990d; Minsky & Papert 1969; Pinker & Prince 1988), they will not turn out to be especially important for the synthesis that will be proposed here: I will suggest a way that symbols and nets can be used collaboratively in modeling cognition. In this hybrid model (Harnad 1992), I lean toward construing nets in their analog form, PAR, because of the role I assign them in processing analog sensory input, but if they could fulfill this role symbolically, that would be fine too.

3.0 WHAT NETS CAN AND CANNOT DO

3.1 Some of themes of the polemics between computationalists and connectionists should perhaps still be mentioned here, for the purposes of this Symposium, if only to show what minor issues they are, compared to the real task at hand:

3.1.1 Computationalism has been criticized by connectionists as neurally implausible, because digital computers are neurally implausible (see McClelland & Rumelhart 1986). But if computationalism is otherwise successful, its implementation independence leaves the door open for a neurally plausible implementation too (which would remain just as irrelevant to the real workings of cognition, because according to computationalism those would still remain syntactic; Fodor & Pylyshyn 1988); besides, the "neurosimilitude" of nets is very superficial.

3.1.2 Computationalism has also been criticized as unconcerned and unsuccessful with learning, preferring to build in symbolic "knowledge" in advance rather than acquiring it in real-time from data (Hanson & Burr 1990). But computationalism could in principle create or help itself to any symbolic algorithm, including learning algorithms (of which neural nets are merely one form of implementation), so the difference here is just in the tools and tasks each approach happens to favor: Symbols have excelled on language-like tasks and tasks that draw on background knowledge; nets have excelled on pattern recognition tasks and tasks that draw on bottom-up learning. Neither has been systematically deployed on the tasks favored by the other (nor is it even clear why, if they were, they could not help themselves to one another's methods and results).

3.1.3 Connectionism has been been criticized for being unable to solve, or for solving only with difficulty, problems that are easy to solve with symbols. Minsky & Papert (1969) singled out "exclusive-or" problems, Pinker & Prince (1988) focused on English past-tense formation rules, but neither critique has turned out to be based on a limitation of connectionism in principle -- just the short-comings of a particular connectionist model (a two-layered Perceptron in both cases, as it happens, Rosenblatt 1962). Theorems have been proved about the generalized curve-fitting capacities of multi-layered nets (Hanson & Burr 1990; and even their ability to implement Turing Machines, Touretzky 1990, 1991; Touretzky & Hinton 1988), but no one yet knows how what either symbols or nets can and cannot do in principle squares with what people's minds can and cannot do, so until we know, both nets and symbols (and other candidates, if any) are still in the empirical race (Harnad 1990d).

3.1.4 A potentially more serious objection against nets has been raised by Fodor & Pylyshyn (1988), who pointed out that in general nets lack the the systematic compositional properties conferred automatically by language-like symbol systems, and that these properties are surely essential for the "language of thought" (Fodor 1975). This may be a valid point, but proposing symbols systems as models for the language of thought runs straight into the "symbol grounding problem" (Harnad 1990), which is that although the symbols in a symbol system can be systematically interpreted as if they meant what thoughts mean, those meanings are no more intrinsic to the symbol system than they are intrinsic to a book. They are merely projected onto them by thinking systems, such as ourselves, when we interpret them; hence, on pain of infinite regress, a symbol system cannot be the right model for what is going on in our heads. We turn now to a closer look at the symbol grounding problem.

4.0 THE SYMBOL GROUNDING PROBLEM AND SEARLE'S CHINESE ROOM ARGUMENT

4.1 Searle (1980) was pointing to a manifestation of the symbol grounding problem when he pointed out that a computer running a program that could pass the Turing Test (TT) (Turing 1964) in Chinese (i.e., could correspond for a lifetime indistinguishably from a real Chinese pen-pal) could not really be understanding

Chinese because Searle himself (who knows no Chinese) could implement the very same program by memorizing and executing all the rules and symbol manipulations -- but without understanding Chinese at all. Searle thereby showed that one of the essential properties of computation (the very one that had made some theorists, e.g., Pylyshyn 1984, think that computationalism may even have solved the mind/body problem) namely, its implementation-independence, actually provided a means of *disconfirming* the inference that the system understands Chinese, and hence the hypothesis that cognition is just computation. For although the pen-pal's symbolic inputs and outputs (and even its internal states) are systematically interpretable as meaning what Chinese thoughts mean, there is in fact no meaning, and hence no thinking, in the system; it was all just a projection (Harnad 1989).

4.2 Note that normally we are free to infer that even stones have minds, for the other-minds problem ensures that even if we are wrong, no one other than the stone can ever be the wiser (Harnad 1984, 1991). But in the special case of computationalism, Searle showed that we *could* be the wiser, for we could ourselves become implementations of the very same symbol system that had passed the Chinese TT and had made us infer that it therefore understood Chinese, yet we would not understand Chinese in so doing. So (unless one is prepared to believe, as Dyer [1990; cf Harnad 1990c] does, that memorizing a lot of meaningless symbols could generate multiple personality, or, as Hayes does, that a human implementation does not count as a real implementation, even though it passes the same TT, by performing exactly the same symbol manipulations, for years and years -- see Hayes et al 1992) computationalism is simply wrong: Cognition is *not* just (implemented) computation.

4.3 Why is computationalism wrong, and what might be right in its stead? I think it's wrong because of the symbol grounding problem: Consider why one could never learn Chinese as a first language from a Chinese-Chinese dictionary alone: Except to someone who already knows Chinese, or at least another language into which Chinese can be translated, a Chinese-Chinese dictionary is just a bunch of meaningless symbols that happen to be systematically interpretable (by someone who already knows Chinese) as meaning what Chinese words and sentences (and thoughts) mean. So it is with any symbol system. The symbols, despite their systematic interpretability, are ungrounded; their meanings are parasitic on the mind of an interpreter. So the symbol grounding problem concerns how the meanings of the symbols in a system can be grounded (in something other than just more ungrounded symbols) so they can have meaning independently of any external interpreter.

5.0 ARE NEURAL NETS SUBJECT TO SEARLE'S ARGUMENT AND THE SYMBOL GROUNDING PROBLEM?

5.1 Now, before we move on to my own candidate solution to the symbol grounding problem, let us first ask whether or not neural nets have the problem. Prima facie, they could not, because, as Fodor & Pylyshyn (1988) point out, lacking the compositional properties of a symbol system in the first place, they could not even bear the systematic weight of any kind of semantic interpretation, grounded or ungrounded! But if we were free to suppose a symbol system could pass the TT without our producing one that could actually do so, we can, I suppose, do the same favor to neural nets. So let us suppose a neural net could pass the TT: Would it understand Chinese?

5.2 Here Searle's (1990) own writing on the subject is not helpful, unfortunately, for he has proposed the "Chinese Gym Argument," in which the TT-passing neural net is implemented as a bunch of boys passing around messages in a gymnasium. Searle points to the system and says it is obvious there's no meaning in there. In my opinion, not only is this not at all obvious, but if such mere hand-waving were all that the original Chinese Room Argument had been based on, then that argument would have been wrong too, and the "System Reply" -- to the effect that Searle is just part of a system, and that it is the system as a whole, not Searle, that would understand Chinese -- the reply favored by most of Searle's critics, would have been

correct (Searle 1980, Harnad 1989). For, whether it be a stone, a gymnasium full of boys, a neural net, or a head full of neurons and neurotransmitters, there is in general no way (because of the other-minds problem) of confirming or disconfirming that the system does or does not have a mind except by *being* the system. "How could it possibly be understanding?" is not an argument. Even in the case of the symbol system, Searle could not say for sure that that computer over there, passing the Chinese TT for years, does not understand Chinese: It might, but then only because of special properties of its silicon. For if the hypothesis is that that computer understands Chinese *only because* it implements the TT-passing symbol system, and that every other implementation of that symbol system would likewise understand, because all implementational differences are irrelevant, then Searle can implement the same symbol system, not understand Chinese, and thereby show *that* particular hypothesis to be false.

5.3 Can Searle do the same for the neural net? The Chinese Gym Argument fails, for it is quite possible that the system *does* understand Chinese. There is, however, a variant of Searle's original Chinese Room Argument (related to the alternative construals of neural nets, SIM, IMP and PAR that I outlined above) that one might call the "Three Room Argument," that will accomplish almost the same result:

5.4 We can lay alternative IMP to rest right away: If neural nets are just used as the hardware to implement symbol systems, then they can only make an (arbitrary) claim to being special (insofar as implementing minds is concerned) on pain of violating the implementation-independence of computation, just as special claims about silicon implementations would.

5.5 So let's forget about alternative IMP and consider only SIM and PAR. Suppose we have three rooms, the system in each one successfully passing the TT. In the first room is a real parallel-distributed processing system (PAR), in the second, a complete computational simulation of it (SIM) and in the third room, Searle, also implementing SIM. It is clear that SIM would not be understanding Chinese, for exactly the same reason as in the original Chinese Room Argument (it is just a TT-passing symbol system). What about PAR? Well, SIM and PAR are not only Turing Equivalent (because they both pass the TT) but they are even "strongly equivalent," in that there is a state-for-state computational equivalence between them, just as there is between SIM and Searle, except that the parallelism is simulated rather than real. Hence, if someone had an independent reason for believing that parallelism was an *essential* implementational feature of the mind (even given both Turing equivalence and computational equivalence), he could argue that (for that reason) the first room (PAR) understands whereas the second (SIM) and third (Searle) do not; but in the absence of such an argument, the parallel implementation in the first room fails to understand for exactly the same reason that the computationally equivalent serial implementation in the second room fails, namely, because the computationally equivalent serial implementation in the third room (Searle) fails to understand: Computational equivalence (and hence anything that allegedly "supervenes" on it) is a transitive relation.

5.6 So so much for the hope that neural nets, even if they were systematically interpretable and TT-capable, would be exempt from the symbol grounding problem. Is anything exempt? Essential parallelism would at least be immune to Searle's Argument, but without an argument as to why parallelism should be essential to cognition, invoking parallelism would be as arbitrary as claiming special status for a silicon implementation. I have proposed a variant of the TT, however, called the Total Turing Test (TTT), that also turns out to be immune to Searle's Chinese Room Argument (Harnad 1989, 1991), but for principled reasons that even suggest one possible solution to the symbol grounding problem (Harnad 1990).

6.0 THE TOTAL TURING TEST (TTT) AND THE TRANSDUCER COUNTERARGUMENT

6.1 The force of Turing's (1964) original Test was both (1) intuitive and (2) empirical. Intuitively, if we could not distinguish a candidate from ourselves by exactly the same criteria we use in judging one another in our

ordinary, everyday solutions to the other-minds problem, then it would be arbitrary to invoke new criteria when we were told that the candidate, indistinguishable from ourselves by our normal intuitive criteria, happened to be a machine. That would amount to special pleading, for not only do we not normally invoke extra engineering or biological criteria in making such judgments, but we don't even know what a machine or a person is, from a bioengineering standpoint: No one does yet. That's the rationale for the intuitive aspect of the TT; the empirical side is that if the candidate's behavioral capacities are identical to our own, what more can we ask, empirically?

6.2 Well, in the case of the TT, there was more we could ask for empirically, for human behavioral capacity includes a lot more than just pen-pal (symbolic) interactions. There is all of our sensorimotor capacity to discriminate, recognize, identify, manipulate and describe the objects, events and states of affairs in the world we live in (the same objects, events and states of affairs, by the way, that our thoughts happen to be about). Let us call this further behavioral capacity our *robotic* capacity. Passing the TTT would then require indistinguishability in both symbolic and robotic capacity.

6.3 Now back to the Chinese room, but this time TTT-scale rather than just TT-scale. This time, instead of asking whether the TT-passing candidate really understands Chinese or is merely systematically interpretable as if he were understanding it, we will ask whether the TTT-passing candidate (a robot now) really sees the Buddha statue before him or is merely systematically interpretable as if he were seeing it. The robot, in order even to be interpretable as seeing, must have optical transducers. What about Searle, who is attempting to implement the TTT robot without seeing, as he implemented the TT robot without understanding? There are two possibilities, either Searle receives only the *output* of an optical transducer -- in which case it is no wonder that he reports he is not seeing, because he is not implementing the whole system, only part of it, and hence, as in the Chinese Gym, the System Reply would be correct; or Searle actually looks at the Buddha, in which case he would indeed be implementing the transduction, but then, unfortunately, he *would* be seeing.

6.4 The fact that it was sufficient to block Searle's Argument should suggest that sensory transduction, normally thought of by computationalists as a trivial peripheral function, may not be that trivial -- at least not TTT-scale transduction. Real transduction is in fact *essential* to TTT capacity. A computational simulation of transduction cannot get from real objects to either robotic performance or symbolic performance (not to mention that motor interaction with real objects also requires the output counterparts of transducers: effectors). This is the requisite nonarbitrary argument for the special status of transduction that we did *not* have in the case of parallelism (or silicon). In addition, there are other things to recommend transduction as an essential component in implementing cognition. First, most of the real brain is either doing sensory transduction or analog extensions of it: As one moves in from the sensory surfaces to their multiple analogs deeper and deeper in the brain, one eventually reaches the motor analogs, until finally one finds oneself out at the motor periphery. If one removed all this sensorimotor equipment, very little of the brain would be left, and certainly not some homuncular computational core-in-a-vat that all this transduction was input *to*. No, to a great extent we *are* our sensorimotor transducers and their activities, rather than being their ghostly computational executives.

7.0 A HYBRID ANALOG/SYMBOLIC ROBOT GROUNDED BOTTOM-UP IN SENSORY CATEGORIES BY NEURAL NETS

7.1 So if transduction is special enough to block Searle's Argument, might it play some role in symbol grounding? According to my hypothesis, it plays a central role: A grounded system is one that has the robotic and the symbolic capacity to pass the TTT in such a way that its symbols and symbolic activity cohere systematically with its robotic transactions with the objects, events and states of affairs that its symbols are interpretable as being about. In other words, its symbolic capacity is grounded in its robotic capacity rather than being mediated by an outside interpretation projected onto it.

7.2 Grounding, by its very nature, is something that is better done bottom-up (rather than, say, top-down from a symbolic skyhook). Hence grounding means sensorimotor grounding: Symbols must be grounded in the capacity to discriminate and identify the objects, events and states of affairs that they stand for, from their sensory projections. A robot could perform discrimination (same-different judgments and similarity judgments, which are relative judgments on pairs of objects) using analog processing and comparators alone (superimposing analog projections of objects), but identification requires a mechanism for recognizing *categories* of objects. Here is where neural nets can play a role for which they seem especially well suited: Learning the invariants in the analog sensory projection that will allow objects to be reliably identified and assigned a name (a ground-level symbol). Symbols thus grounded in our capacity to discriminate and identify their referents (e.g., "horse" and "stripes") can then be combined into descriptions of new symbols that inherit their grounding (e.g., "zebra" = "horse" & "stripes") without the need of any direct sensory learning.

7.3 There are objections to this kind of approach, which I will be happy to attempt to answer in my Response to the commentaries. Here I will reproduce only one pre-emptive passage from Harnad (1992).

7.3.1 The anti-empiricist objections can be summarized as follows: For most categories, necessary and sufficient conditions for category membership, and especially sensory ones, simply do not exist. The evidence for this is that we are not aware of using any, and when we think about what they might be, we can't think of any. In addition, categories are often graded or fuzzy, membership being either a matter of degree or even uncertain or arbitrary in some cases. Sensory invariants are even less likely to exist: The intersection of all the properties of the sensory projections of the members of the category "good" is surely empty. Moreover, sensory appearances are often deceiving, and rarely if ever decisive: A painted horse that looks just like a zebra is still not a zebra.

7.3.2 The roboticist's reply is that introspection is unlikely to reveal the mechanisms underlying our robotic and cognitive capacities, otherwise the empirical task would be much easier. Disjunctive, negative, conditional, relational, polyadic, and even constructive invariants (in which the input must undergo considerable processing to extract the information inherent in it) are just as viable, and sensory-based, as the simple, monadic, conjunctive ones that introspection usually looks for. There are graded categories like "big," in which membership is relative and a matter of degree, but there are also all-or-none categories like "bird," for which invariants exist. There may be cases of "bird" we're not sure about, but we're not answerable to God's omniscience about what's what, only to the consequences of miscategorization insofar as they exist and matter to us. And it's our successful categorization performance that a robotic model must be able to capture -- including our capacity to revise our provisional, approximate category invariants in the face of error. As to goodness, truth and beauty: There is no reason to doubt that -- insofar as they are objective rather than subjective categories -- they too are up there somewhere, firmly grounded in the zebra hierarchy, just as the "peekaboo unicorn" is: The peekaboo unicorn is "a horse with a horn that vanishes without a trace whenever senses or measuring instruments are trained on it." Unverifiable in principle, this category is nevertheless as firmly grounded (and meaningful) as "zebra" -- as long as "horse," "horn," "vanish," "trace," "senses" and "measuring instrument" are grounded. And we could identify its members on first encounter -- if we ever could encounter them -- as surely as we could identify a zebra. The case of the painted horse and of goodness, truth and beauty is left to the reader as an exercise in exploring the recursive possibilities of grounded symbols.

7.4 So, assuming that these prima facie objections have some prima facie answers, the question becomes an empirical one: Can a bottom-up approach to symbol grounding along these lines work, and does it have any hope of scaling up to the TTT? I can offer only the bare beginnings of evidence. Neural nets can indeed find the invariants in simple one-dimensional tasks so as to allow the dimension to be partitioned into categories (Harnad et al. 1991 and in prep.; Lubin et al. in prep). No one knows yet whether these nets have the power to accomplish human scale categorization (Hanson & Burr 1990) but they have already shown that they share with human categorization an interesting feature called "categorical perception" (CP), whereby within-category similarities are compressed and between-category differences are enhanced in the service of

categorization (Andrews et al., in prep; Harnad 1987; Lawrence 1950). This feature, which is not readily explained from the cognitive standpoint (why should things in the same category come to look more alike?), is explained by the way certain kinds of nets accomplish successful categorization by "warping" the analog similarity space enough to be able to partition it as dictated by the feedback from the consequences of miscategorization (Harnad et al. 1991).

7.5 The grounding model proposed here is easily summarized: Analog sensory projections are the inputs to neural nets that must learn to connect some of the projections with some symbols (their category names) and some of them with other symbols (the names of other interconfusable categories) by finding and using the invariant features in them that will subserve correct categorization performance. The grounded symbols are then strung into higher order combinations (grounded symbolic descriptions) by a second, combinatory process that differs from classical symbol manipulation in a critical respect. In standard (ungrounded) symbol manipulation, the only constraint on the symbol combinations is the syntax, which operates on the (arbitrary) shapes of the symbols. In a grounded symbol system there is a second constraint, that of the nonarbitrary "shape" of the sensory invariants that connect the symbol to the analog sensory projection of the object to which it refers. I cannot say much about the nature of these grounded "doubly constrained" symbol systems, except that human categorical perception may give some hints about the nature of this interaction between analog and syntactic constraints (Harnad 1992).

8.0 POTENTIAL OBJECTIONS AND ALTERNATIVES

8.1 Let me close by counting the ways my proposal could be wrong:

8.1.1 Nets could fail to have the inductive power to accomplish human-scale categorization, whereas some other, non-connectionistic pattern-learning algorithm could succeed. In that case, neural nets would play no essential role in this particular grounding model, but the basic architecture would be the same.

8.1.2 If those nonconnectionist learning algorithms failed to warp similarity space, then categorical perception would merely be an epiphenomenon, rather than a clue about the constraints exerted by symbolic representations on analog projections or vice versa.

8.1.3 Or my grounding model could fail altogether, because bottom-up grounding of abstract categories in sensory ones turns out to be unlearnable, either by the child or through evolution. This would leave us with a (to me) rather mysterious "Big Bang" Theory of the innate origin and grounding of symbols, but some thinkers have not found such a possibility to be either uncongenial or improbable (e.g., Chomsky 1980; Fodor 1975).

8.1.4 Alternatively, neural nets could conceivably do it all, without having to resort to a higher-level symbol system, successfully generating TTT capacity without symbols (hence no need to ground them); this would leave certain logical and linguistic properties of thought much more holistic than they seem to be (Fodor & Pylyshyn 1988), but who knows? The symbol grounding problem would then be just an epiphenomenon and connectionism would have cognitive hegemony.

8.1.5 Or transduction could somehow turn out to be trivial after all, and a symbol system may turn out to be so powerful that all it needs is to have a few transducers hooked up to it and it can zip through the TTT. In that case, either Dyer (1990) would be right that Searle in the Chinese Room had a second, Chinese mind without knowing it, as a consequence of memorizing and executing all those symbols and rules, or Hayes would be right that Searle's implementation somehow did not count (Hayes et al 1992). In either case, the symbol grounding problem would be a red herring, nets a minor interloper, and computationalism would have cognitive hegemony.

8.2 My own guess is that none of these alternative is very likely to be the ultimate one, including the one I

proposed, and that as yet unthought of forms of analog computation (perhaps putting more weight on the motor side of the sensorimotor transaction) will play a big role in creating a robot with grounded symbols and the capacity to pass the Total Turing Test.

REFERENCES

Andrews, J., Livingston, K., Harnad, S. & Fischer, U. (in prep.) Learned Categorical Perception in Human Subjects: Implications for Symbol Grounding.

Chomsky, N. (1980) Rules and representations. Behavioral and Brain Sciences 3 : 1-61.

Dietrich, E. (1990) Computationalism. Social Epistemology 4: 135 - 154.

Dyer, M. G. Intentionality and Computationalism: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence, Vol. 2, No. 4, 1990.

Fodor, J. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3 - 71.

Fodor, J. A. (1975) The language of thought New York Thomas Y. Crowell

Hanson & Burr (1990) What connectionist models learn: Learning and Representation in connectionist networks. Behavioral and Brain Sciences 13: 471-518.

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

Harnad, S. (1987) (ed.) Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

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

Harnad, S. (1990a) The Symbol Grounding Problem. Physica D 42: 335-346.

Harnad, S. (1990b) Against Computational Hermeneutics. (Invited commentary on Eric Dietrich's Computationalism) Social Epistemology 4: 167-172.

Harnad, S. (1990c) Lost in the hermeneutic hall of mirrors. Invited Commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2: 321 - 327.

Harnad, S. (1990d) Symbols and Nets: Cooperation vs. Competition. Review of: S. Pinker and J. Mehler (Eds.) (1988) Connections and Symbols Connection Science 2: 257-260.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Harnad, S., Hanson, S.J. & Lubin, J. (1991) Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In: Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology (DW Powers & L Reeker, Eds.) pp. 65-74. Presented at Symposium on Symbol Grounding: Problems and Practice, Stanford University, March 1991; also reprinted as Document D91-09,

Deutsches Forschungszentrum fur Kuenstliche Intelligenz GmbH Kaiserslautern FRG.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on the Virtual Mind. Minds and Machines (in press)

Harnad, S. Hanson, S.J. & Lubin, J. (in prep.) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding.

Lawrence, D. H. (1950) Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. Journal of Experimental Psychology 40: 175 - 188.

Lubin, J., Hanson, S. & Harnad, S. (in prep.) Categorical Perception in ARTMAP Neural Networks.

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.

MacLennan, B. J. (1987) Technology independent design of neurocomputers: The universal field computer. In M. Caudill & C. Butler (Eds.), Proceedings, IEEE First International Conference on Neural Networks (Vol. 3, pp. 39-49). New York, NY: Institute of Electrical and Electronic Engineers.

MacLennan, B. J. (1988) Logic for the new AI. In J. H. Fetzer (Ed.), Aspects of Artificial Intelligence (pp. 163-192). Dordrecht: Kluwer.

MacLennan, B. J. (in press-a) Continuous symbol systems: The logic of connectionism. In Daniel S. Levine and Manuel Aparicio IV (Eds.), Neural Networks for Knowledge Representation and Inference. Hillsdale, NJ: Lawrence Erlbaum.

MacLennan, B. J. (in press-b) Characteristics of connectionist knowledge representation. Information Sciences, to appear.

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

Newell, A. (1980) Physical Symbol Systems. Cognitive Science 4: 135 - 83.

Pinker, S & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28(1-2): 73-193.

Pylyshyn, Z. W. (1984) Computation and cognition. Cambridge MA: Bradford Books

Rosenblatt, F. (1962) Principles of neurodynamics. NY: Spartan

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

Searle, J. (1990) Is the brain's mind a computer program?. Scientific American 262: 26-31.

Touretzky, D. S. (ed.) (1991) Machine Learning, vol. 7, nos. 2 and 3, special double issue on ``Connectionist Approaches to Language Learning."

Touretzky, D. S. (1990) BoltzCONS: Dynamic symbol structures in a connectionist network. Artificial Intelligence vol. 46, pp. 5-46.

Touretzky, D. S. and Hinton, G. E. (1988) A distributed connectionist production system. Cognitive Science, vol. 12, number 3, pp. 423-466.

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