Harnad, S. (1995) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents. New Haven: Lawrence Erlbaum. Pp. 276-286.

GROUNDING SYMBOLIC CAPACITY IN ROBOTIC CAPACITY

Stevan Harnad
Laboratoire Cognition et Mouvement
URA CNRS 1166 I.B.H.O.P.
Universite d'Aix Marseille II
13388 Marseille cedex 13, France
harnad@riluminy.univ-mrs.fr
33-91-66-00-69

According to "computationalism" (Newell, 1980; Pylyshyn 1984; Dietrich 1990), mental states are computational states, so if one wishes to build a mind, one is actually looking for the right program to run on a digital computer. A computer program is a semantically interpretable formal symbol system consisting of rules for manipulating symbols on the basis of their shapes, which are arbitrary in relation to what they can be systematically interpreted as meaning. According to computationalism, every physical implementation of the right symbol system will have mental states.

Artificial intelligence (AI) is the branch of computer science that is concerned with designing symbol systems that have performance capacities that are useful to human beings. Cognitive science includes AI as well as mind-modelling (MM), which is concerned with building systems that are not only useful to people with minds, but that *have*
minds of their own. According to computationalism, AI can do both these things, and for several decades it was hoped that it would. AI's advantages in this regard were the following:

AI could indeed (1) generate performance that ordinarily requires human intelligence and, unlike, say, behavioral psychology (Harnad 1982, 1984; Catania & Harnad 1988), AI could explain the functional and causal basis of that performance.

There was also (2) reason to be optimistic about scaling up the performance of AI's initial "toy" models to human-scale performance because of formal results on the power and generality of computation; according to one construal of the Church-Turing Thesis, computation captures everything we mean by being able to "do" just about *anything*, whether formally or physically (Dietrich 1993). Hence a computer can do anything any physical system can do -- or, conversely, every physical system is really a computer.

The last of the initial reasons for optimism about AI for MM was (3) the apparent capacity of the software/hardware distinction to solve the mind/body problem: If computationalism is correct, and mental states are just implementations of certain symbolic states, then the persistent difficulty that philosophers have kept pointing out with equating the mental and the physical is resolved by the independence of a physical symbol system's formal, symbolic level (the software level) from its physical implementation (the hardware level). A symbol system is implementation-independent, and so is the mind.

Unfortunately, problems arose for AI, and not just when it tried to do MM. AI systems have so far not proved to scale up readily, not only for the human-scale performance necessary for MM, but even for the kinds of performance that were merely intended to be useful to people, such as pattern recognition and robotics. Rival approaches began to appear, among them (a) robots that made minimal use (or none at all) of internal symbol systems (Brooks 1993); (b) "neural nets," which were systems of interconnected units whose parallel distributed activity likewise did not seem to have a structured symbolic level (Hanson & Burr 1990); and (c) nonlinear dynamical systems in general, including continuous and chaotic ones that were not

readily covered by the Church-Turing Thesis (Kentridge 1993).

In addition, conceptual challenges were posed to the computationalist thesis, the one that had made it seem that AI would be capable of doing MM in the first place. Two such challenges were Searle's (1980) "Chinese Room Argument" and my own "Symbol Grounding Problem" (Harnad 1990a). Searle pointed out that the tenets of computationalism ("Strong AI") amounted to three hypotheses: (i) mental states are implementations of symbolic states; (ii) all physical implementational details are irrelevant (because any and all implementations of the right symbol system will have the same mental states, hence the differences between them are all inessential to having a mind); (iii) performance capacity is decisive (and hence the crucial test for the presence of a mind is the Turing Test (T2), which amounts to the capacity to interact with a person as a life-long pen-pal, indistinguishable in any way from a real person; note that this test is purely symbolic). Searle then pointed out that if there were a computer that could pass T2 in Chinese, he (Searle) could become another implementation of the same symbol system it was implementing (by memorizing all the symbol manipulation rules and then performing them on all the symbols received from the Chinese pen-pal), yet he would not thereby be understanding Chinese, hence neither would the computer that was doing the same thing. In other words, there was something wrong with hypotheses (i) - (iii): They couldn't all be correct, yet computationalism depended on the validity of all three of them.

Searle's (1990, 1993) recommended alternative to computationalism and AI for those whose real interest was MM was to study the real brain, for only systems that had its "causal powers" could have minds. The only problem was that this left no way of sorting out which of the brain's causal powers were and were not *relevant*
to having a mind (Harnad 1993a). We now knew (thanks in part to Searle) that the relevant causal powers were not exclusively the symbolic ones, but we did not know what the rest of them amounted to; and to assume that every physical detail of the real brain -- right down to its specific gravity -- was relevant and indeed essential to MM was surely to take on too much. A form of functionalism that sought to abstract the relevant causal powers of the brain already motivated AI: The relevant level was the symbolic one, and once that was specified, every physical implementation would have the relevant causal powers. Searle showed that this particular abstraction was wrong, but he did not thereby show that there could be no way to abstract away from the totality of brain function and causal power.

The symbol grounding problem pointed out another functional direction in which the causal powers relevant to having a mind may lie: The symbols in a symbol system are systematically interpretable as meaning something; however, that interpretation is always mediated by an external interpreter. It would lead to an infinite regress if we supposed the same thing to be true of the mind of the interpreter: that all there is in his head is the implementation of a symbol system that is systematically interpretable by yet *another* external interpreter. My thoughts mean what they mean intrinsically, not because someone else can or does interpret them (e.g., Searle understands English and fails to understand Chinese independently of whether the English or Chinese symbols he emits are systematically interpretable to someone else). The infinite regress is a symptom of the fact that the interpretation of a pure symbol system is ungrounded. I think the "frame problem," which keeps arising in pure AI -- what changes and what stays the same after an "action"? (Pylyshyn 1991, Harnad 1993f) -- is another symptom of ungroundedness.

Another way to appreciate the symbol grounding problem is to see it as analogous to trying to learn Chinese as a second language from a Chinese/Chinese dictionary alone: All the definientes and definienda in such a dictionary are systematically interpretable to someone who *already* knows Chinese, but they are of no use to someone who does not, for such a person could only get on a merry-go-round passing from meaningless symbols to still more meaningless symbols in cycling through such a dictionary. Perhaps with the aid of cryptography there is a way to escape from this merry-go-round (Harnad 1993c, 1994), but that clearly depends on being able to find a way to decode the dictionary in terms of a first language one already knows. Unfortunately, however, what computationalism is really imagining is that the substrate for this *first* language (whether English or the language of thought, Fodor 1975) would likewise be just more of the same: ungrounded symbols that are systematically interpretable by someone who already knows what at least some of them mean.

So the problem is that the connection between the symbols and what they are interpretable as being about

must not be allowed to depend on the mediation of an external interpreter -- if the system is intended as a model of of what is going on in the external interpreter's head too, as MM requires. One natural way to make this connection direct is to ground the symbols in the system's own capacity to interact robotically with what its symbols are about: it should be able to discriminate, manipulate, categorize, name, describe and discourse about the real-world objects, events and states of affairs that its symbols are about -- and it should be able to do so Turing indistinguishably from the way we do (I have called this the Total Turing Test or T3, Harnad 1989, 1992b, 1993a.) In other words, T2 and computationalism ("symbolic functionalism") are ungrounded and hence cannot do MM, whereas T3 and "robotic functionalism," grounded in sensorimotor interaction capacity, can.

In my own approach to symbol grounding I have focussed on the all-important capacity to categorize (sort and name) objects (Harnad 1987, 1992a) -- initially concrete categories, based on invariants (learned and innate) in their sensory projections, and then abstract objects, described by symbol strings whose terms are grounded bottom-up in concrete categories (e.g., if the categories "horse" and "striped" are grounded directly in the capacity to sort and name their members based on their sensorimotor projections, then "zebra" can be grounded purely symbolically by binding it to the grounded symbol string "striped horse": a robot that could only sort and name horses and striped objects before could then sort and name zebras too). What is important to keep in mind in evaluating this approach is that although all the examples given are just arbitrary fragments of our total capacity (and the initial models, e.g. Harnad et al. [1991, 1994], are just toys), the explicit goal of the approach is T3-scale capacity, not just circumscribed local "toy" capacity. In my own modelling I use neural nets to learn the sensorimotor invariants that allow the system to categorize, but it is quite conceivable that neural nets will fail to scale up to T3-scale categorization capacity, in which case other category-invariance learning models will have to be found and tried. On the other hand, rejecting this approach on the grounds that it is already known that bottom-up grounding in sensorimotor invariants is not possible (e.g. Christiansen & Chater 1992, 1993) is, I think, premature (and empirically ungrounded; Harnad 1993e).

The symbol grounding approach to MM can be contrasted with other approaches that prefer to dispense with symbols altogether. I will consider two such approaches here. One is pure connectionism (PC), which replaces the computationalist hypothesis that mental states are computational states with the connectionist hypothesis that mental states are dynamical states in a neural net (e.g., Hanson & Burr 1990). The crucial question for connectionists, I think, is whether the critical test of the PC hypothesis is to be T2 or T3 (I think it is a foregone conclusion that mere toy performance demonstrates nothing insofar as MM is concerned). If it is to be T2, then I think PC is up against the same objections as AI, if for no other reason than because connectionist systems can be simulated by symbol systems without any real parallelism or distributedness, and if those too can pass T3, then they are open to Searle's Argument and the symbol grounding problem (Harnad 1993a, Searle 1993). On the other hand, if the target is to be T3, and PC can manage to do it completely nonsymbolically, I, for one, would be happy to accept the verdict that it was not necessary to worry about the problem of grounding symbols, because symbols are not necessary for MM. On the other hand, there do exist prima facie reasons to believe that a PC approach would fail to capture the systematicity that is needed to pass the T2 (a subset of T3) in the first place (Fodor & Pylyshyn 1988, Harnad 1990b), so perhaps it is best to wait and see whether or not PC can indeed go it alone.

There is a counterpart to PC in robotics -- let's call it "pure nonsymbolic robotics" PNSR (Brooks 1993) -- which likewise aspires to go the distance without symbols, but this time largely by means of internal sensorimotor mechanisms -- sometimes neurally inspired ones, but mostly data-driven ones: driven by the contingencies a robot faces in trying to get around in the real world. Such roboticists tend to stress "situatedness" and "embeddedness" in the world of objects (which they take to be "grounding" without symbols) rather than symbol grounding. PNSR places great hope in internal structures that will "emerge" to meet the bottom-up challenges of navigating and manipulating its world; much has been made, for example, of a wall-following "rule" that emerged spontaneously in a locomoting robot that had been given no such explicit rule (Steels REF). As with PC, however, it remains to be seen whether such "emergent" internal structures and rules, driven only by bottom-up contingencies, can scale up to the systematicity of natural language and human reasoning (Fodor & Pylyshyn 1988) without recourse to internal symbols.

My own work on categorical perception (Harnad 1987), which is pretty low in the concrete/abstract scale

leading from sensorimotor categories to language and reasoning, already casts some doubt on the possibility of scaling up to T3 without internal symbols, as PNSR hopes to do. A category name, after all, is a symbol, and we all use them. Categorical perception occurs when the analog space of interstimulus similarities is "warped" by sorting and naming objects in a particular way, with the result that within-category distances (the pairwise perceptual similarities between members of the same category, bearing the same symbolic category name) are compressed and between-category distances (the similarities between members of different categories, bearing different symbolic category names) are expanded. This seems to occur because after category learning, the sensorimotor projections of objects are "filtered" by invariance detectors that have learned which features of the sensory projection will serve as a reliable basis for sorting and labelling them correctly (and the warping of similarity space seems to be part of how backpropagation, at least, manages to accomplish successful categorization; Harnad et al. 1991, 1994). The next stage is to combine these grounded symbols into propositions about more abstract categories (e.g., "zebra" = "striped horse"). It is hard to imagine how this could be accomplished by a data-driven "emergent" property such as wall-following. It seems more likely that explicit internal symbolization is involved.

Such internal symbols, unlike those of AI's pure symbol systems, inherit the constraints from their grounding. In a pure symbol system the only constraints are formal, syntactic ones, operating rulefully on the arbitrary shapes of the symbols. In a grounded symbol system, symbol "shapes" are no longer arbitrary, for they are constrained (grounded) by the structures that gave the system the capacity to sort and name the members of the category the symbol refers to, based on their sensorimotor projections: the shape of "horse" is arbitrary, to be sure, but not that of the analog sensory projections [see Chamberlain & Barlow 1982, Shepard & Cooper 1982, Harnad 19993f, Jeannerod, 1994] of horses nor of the invariants in those sensory projections that the nets have detected and that connect the "horse" symbol to the projections of the objects it refers to. All further symbolic combinations that "horse" enters into (e.g., "zebra" = "striped horse") inherit this grounding. Think of it as the "warping" of similarity space as a consequence of which things no longer look the same (from color sorting [Bornstein 1987], where "warping" is innate, to chicken-sexing [Biederman & Shiffrar 1987; Harnad et al., in prep.], where it is learned) after you have learned to sort and name them in a certain way. All further symbol combinations continue to be constrained by the invariance detectors and the changed in "appearance" that they mediate.

So I am still betting on internal symbols, but grounded ones. In my view, robotic constraints play two three in MM: (1) They ease the burden of trying to second-guess T3 constraints a priori, with a purely symbolic "oracle": Instead of just simulating the robot's world, it makes more sense to let the real world exert its influence directly (Harnad 1993b). More important than that, (2) the robotic version of the Turing Test, T3, is just the right constraint for the branch of reverse engineering that MM really is. T2 clearly is not (because of Searle's argument and the symbol grounding problem), whereas Searle's preferred candidate, "T4" (total neurobehavioral indistinguishability from ourselves) is overconstraining, because it includes potentially irrelevant constraints. Finally, (3) robotic capacity is precisely looks like just what one would want to ground symbolic capacity IN, given that symbols cannot generate a mind on their own.

Depite these considerations in favor of *symbol*
grounding, neither PC nor PNSR can be counted out yet, in the path to T3. So far only computationalism and pure AI have fallen by the wayside. If it turns out that no internal symbols at all underlie our symbolic (T2) capacity, if dynamic states of neural nets alone or sensorimotor mechanisms subserving robotic capacities alone can successfully generate T3 performance capacity without symbols, that is still the decisive test for the presence of mind as far as I'm concerned and I'd be ready to accept the verdict. For even if we should happen to be wrong about such a robot, it seems clear that no one (not even an advocate of T4, or even the Blind Watchmaker who designed us, being no more a mind-reader than we are) can ever hope to be the wiser (Harnad 1982b, 1984b, 1991, 1992b).

REFERENCES

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

Biederman, I. & Shiffrar, M. M. (1987) Sexing day-old chicks: A case study and expert systems analysis of

a difficult perceptual-learning task. Journal of Experimental Psychology: Learning, Memory, & Cognition 13: 640 - 645.

Bornstein, M. H. (1987) Perceptual Categories in Vision and Audition. In S. Harnad (Ed.) Categorical perception: The groundwork of Cognition. New York: Cambridge Univerity Press

Catania, A.C. & Harnad, S. (eds.) (1988) The Selection of Behavior. The Operant Behaviorism of BF Skinner: Comments and Consequences. New York: Cambridge University Press.

Chamberlain, S.C. & Barlow, R.B. (1982) Retinotopic organization of lateral eye input to Limulus brain. Journal of Neurophysiology 48: 505-520.

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

Dietrich, E. (1993) The Ubiquity of Computation. Think 2: 27-30.

Brooks, R.A. (1993) The Engineering of Physical Grounding. Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. NJ: Erlbaum

Christiansen, M. & Chater, N. (1992) Connectionism, Learning and Meaning. Connectionism 4: 227 - 252.

Christiansen, M.H. & Chater, N. (1993) Symbol Grounding - the Emperor's New Theory of Meaning? Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. NJ: Erlbaum

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

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

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

Harnad, S. (1982a) Neoconstructivism: A unifying theme for the cognitive sciences. In: Language, mind and brain (T. Simon & R. Scholes, eds., Hillsdale NJ: Erlbaum), 1 - 11.

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

Harnad, S. (1984a) What are the scope and limits of radical behaviorist theory? The Behavioral and Brain Sciences 7: 720 -721.

Harnad, S. (1984b) Verifying machines' minds. (Review of J. T. Culbertson, Consciousness: Natural and artificial, NY: Libra 1982.) Contemporary Psychology 29: 389 - 391.

Harnad, S. (1987) The induction and representation of categories. In: Harnad, S. (ed.) (1987) 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) 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. (1992a) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Harnad, S. (1992b) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October) 9 - 10.

Harnad, S. (1993a) 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.).

Harnad, S. (1993b) Artificial Life: Synthetic Versus Virtual. Artificial Life III. Proceedings, Santa Fe Institute Studies in the Sciences of Complexity. Volume XVI.

Harnad, S. (1993c) The Origin of Words: A Psychophysical Hypothesis In Durham, W & Velichkovsky B (Eds.) Muenster: Nodus Pub. [Presented at Zif Conference on Biological and Cultural Aspects of Language Development. January 20 - 22, 1992 University of Bielefeld]

Harnad, S. (1993d) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. PSYCOLOQUY 4(34) frame-problem.11.

Harnad, S. (1993e) Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component. Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. NJ: Erlbaum

Harnad, S. (1993f) Exorcizing the Ghost of Mental Imagery. Commentary on: JI Glasgow: "The Imagery Debate Revisited." Computational Intelligence (in press)

Harnad, S. (1994, in press) Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. Special Issue on "What Is Computation" Minds and Machines

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.

Harnad, S. Hanson, S.J. & Lubin, J. (1994) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. In: V. Honavar & L. Uhr (eds) Symbol Processing and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration. (in press)

Jeannerod, M. (1994) The representing brain: neural correlates of motor intention and imagery. Behavioral and Brain Sciences 17(2) in press.

Kentridge, R.W. (1993) Cognition, Chaos and Non-Deterministic Symbolic Computation: The Chinese Room Problem Solved? Think 2: 44-47.

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

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

Pylyshyn, Z. W. (Ed.) (1987) The robot's dilemma: The frame problem in artificial intelligence. Norwood NJ: Ablex

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

Searle, J.R. (1993) The Failures of Computationalism. Think 2: 68-73.