Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component

Stevan Harnad
Laboratoire Cognition et Mouvement
CNRS URA 1166 Universite d'Aix Marseille II
13388 Marseille cedex 13 FRANCE
email: harnad@princeton.edu

ABSTRACT: *"Symbol Grounding" is beginning to mean too many things to too many people. My own construal has always been simple: Cognition cannot be just computation, because computation is just the systematically interpretable manipulation of meaningless symbols, whereas the meanings of my thoughts don't depend on their interpretability or interpretation by someone else. On pain of infinite regress, then, symbol meanings must be grounded in something other than just their interpretability if they are to be candidates for what is going on in our heads. Neural nets may be one way to ground the names of concrete objects and events in the capacity to categorize them (by learning the invariants in their sensorimotor projections). These grounded elementary symbols could then be combined into symbol strings expressing propositions about more abstract categories. Grounding does not equal meaning, however, and does not solve any philosophical problems.*

Christansen & Chater. Although their critique of connectionist approaches to meaning is valid and well-taken, Christiansen & Chater (this volume, and 1992, henceforth C & C) seem to miss the mark when they apply it to hybrid (nonsymbolic/symbolic) approaches of the kind I advocate (Harnad 1990a, 1992a), for the difference between a hybrid model and a purely connectionistic one is as radical as the difference between a hybrid model and a purely symbolic one. Moreover, even the hybridism is not just 2-way symbolic/connectionist in my approach; rather, it is 3-way (analog - connectionist - symbolic), with the "connectionist" component just a place-holder for *any* mechanism able to learn the invariants in the analog sensorimotor projection that allow the system to do categorization (Harnad 1987). If neural nets turn out to be unable to do this job, other pattern learning mechanisms might still succeed (Harnad 1990b, 1993a).

And even my "symbolic" component is not etched in stone, for it could turn out that a "symbol system" that is grounded in my sense is *no longer symbolic or computational in the formal sense at all* (because, formally, the only constraints on symbols, whose shapes are arbitrary, ought to be syntactic constraints, operating rulefully on those arbitrary shapes, whereas in a grounded system there is a second set of constraints on the "symbols" that is exerted by the nonarbitrary shapes of the sensorimotor projections of the objects that the symbols are *about*, and the invariants in those projections that allow the objects to be assigned symbolic names -- whether those invariants are found by nets or by something else). Such a hybrid, doubly-constrained system may no longer fit the technical definition of a formal symbol system at all; at the very least, it would be a "dedicated" system (like a dedicated computer), with its computations highly constrained by its grounding, over and above (or, rather, under and below) its syntax.

Most of C & C's valid criticism of the limitations of connectionist approaches to meaning is applicable only to a pure connectionism that aspires to do all of cognition alone, replacing the computationalist dogma that mental states are just symbolic states with the connectionist dogma that mental states are just neural-net states. I subscribe to neither of these dogmas. In addition, I have always been very careful neither to state nor to imply that grounding equals meaning. On the contrary, it is (and will always remain) a logical possibility that even the kind of grounded system that is the ultimate goal of my approach -- a system capable of passing the "Total Turing Test," (T3) i.e., one whose linguistic (T2) *and* robotic capacity is totally indistinguishable from our own -- could fail to have any intrinsic internal meanings (Harnad 1989, 1991).

I have also consistently stressed that not only is grounding not *equivalent* to meaning, but there is no way to prove that it is either necessary or sufficient for it: It is logically possible that an ungrounded symbol system has intrinsic meanings or that a grounded symbol system fails to have them. I'm merely betting (probabilistically, but with reasons) that T3-capacity *is* sufficient for having a mind and meaning (Harnad 1992b, 1993a).

Unlike computationalists (e.g., Dietrich 1993), who hold that cognition is a form of implementation-independent symbol manipulation, I reject ungrounded symbol manipulations of *any* kind, even T2-scale ones, in favor of a system with full T3 capacity (T2 is our full, Turing-indistinguishable

pen-pal capacity, i.e., our symbolic capacity, and T3 is our full, Turing-indistinguishable robotic capacity, i.e., full sensorimotor + symbolic capacity, with the latter grounded in the former; Harnad 1992b, 1993a,b, 1994). As a logical matter, such a system will have to be hybrid, because although pure computation (symbol manipulation) is implementation-independent, the performance requirements of such a T3-scale robot depend *essentially* on analog and other nonsymbolic forms of internal structure and function (Harnad 1993a).

C&C seem to think that "the symbol grounding problem" is the "possibility of an externally imposed arbitrary re-interpretation of the representational primitives" (p. 232) in a symbol system. This is *not* the symbol grounding problem. For symbol ungroundedness would continue to be the problem even if only *one, unique* interpretation of a symbol system were possible, indeed, even if its uniqueness were *provable*. The real problem of symbol grounding is that the interpretation of the symbols, whether or not it is unique, is not intrinsic to the symbol system: It is projected onto it by the mind of the interpreter, whereas that is *not* true of the meanings of the thoughts in *my* mind.

The goal of symbol grounding is *not* to guarantee uniqueness but to ensure that the connection between the symbols and the objects they are systematically interpretable as being about does not depend exclusively on an interpretation projected onto the symbols by an interpreter outside the system. This is the role for which I proposed T3-grounding, for the T3-scale robot not only has internal states that are systematically interpretable as being about the objects in the world, but its own causal interactions with the world *cohere totally* (T3-scale) with that interpretability.

To put it simply, a grounded T3-scale symbol system, when it tokens "THE CAT IS ON THE MAT," is not merely systematically interpretable by you and me as meaning the cat is on the mat: it also T3-interacts with the referent of that proposition (and all other systematically related ones) in a way that coheres with the interpretation. Of course, this is still just interpretation (doubly-constrained now, however, not just symbolically [syntactically], but also robotically [causally]); so it still leaves open the possibility that even grounded symbols do not have intrinsic meaning. But at least their grounding is no longer just a matter of *symbol* interpretability or interpretation. In addition, the robot's causal interactions -- with the objects that its symbols are interpretable (by us) as being about -- are *autonomous*: they do *not* depend on our interpretations of the symbols. The T3 robot will (like us) continue to pick out cats when it tokens CAT (etc.) whether or not anyone else interprets CAT as cat.

Thoughts are not dynamic states in a neural network (that "correlate" with input in some way) in my model. The nets connect symbols (arbitrary category names) to objects categories on the basis of their sensorimotor projections. These grounded symbols are then combined into symbol strings that are systematically interpretable as propositions about the world. A "thought" would consist of the activity in (S) the symbol system, (A) analogs of the invariants in the sensorimotor projection, and (N) the nets that have learned to detect them. It is not only, or even mainly, the activity of the net, as in a pure connectionistic model.

In my hybrid model, *no semantic content is assigned to distributed representations* They're just sensorimotor invariance detectors. It's a mistake to assign semantics to a feature-detector. It's also important to stress that the input to the nets in the hypothetical T3-scale model that is the ultimate target of my approach (as opposed to the isolated toy nets we actually test; Harnad et al. 1991) is supposed to be the sensorimotor projection itself, *not* symbols that are interpretable as the descriptions of objects or their sensory projections (cf. Lakoff).

C&C go on to write (correctly) of such purely connectionistic models that "the internal states of present day connectionistic networks appear to be no more "grounded" than their symbolic counterparts ... the distributed representations in question are only non-arbitrary in relation to the structure of the given input representations, not in relation to what the latter are representations *of*, i.e., the entities they refer to in the outside world" (p. 233).

I would go even further: All we really have here is a relation (correlation) between a symbolic input and internal states of a net. The (ungrounded) input symbols are of course interpretable as being about something, but that's neither here nor there. The net activity correlates with the symbols, but that doesn't help either. There's no grounding here, first, because it's symbols that need to be grounded, and it's not clear that the net itself has symbols in the first place. But even if the net does have symbols, the "correlation" is just with yet another set of symbols -- the input symbols. As C&C ask, rightly: what about the connection with what the symbols are *about*?

In contrast, if the inputs, instead of being symbols strings that were interpretable as being about something, were simply sensory projections from real objects onto the system's transducer surfaces, and the net learned the invariant features in the projections

that allowed the objects to be reliably assigned an *(arbitrary)* symbolic category *name*, then that name would indeed be connected to what it was (interpretable as being) about. Yet even that would still not yet be semantics; it would just be a static sensory classifying device. But now suppose that the system scaled up to T3-scale categorization capacity, and that category names could be combined into strings of propositions about more abstract objects, events and states of affairs. The symbol strings would now not only "correlate" with what they were (interpretable as being) about: they would also be causally connected with them. Now, however, we are no longer talking about distributed representations in a mere feature-detector, but about the system as a whole, or at least much larger systematically interacting chunks of it, including its (A) analog, (N) connectionist network, and (S) all-important combinatory symbolic components.

The "Problem of Error" (how can a purely correlational "concept" be wrong?) does not arise as long as one is careful to separate (irrelevant) *ontic* questions (about what the things referred to by our words *really* are) from the *empirical* questions that are proper to psychology and cognitive science (what do we call what?): Our only responsibility is to explain how people use words in the world, what objects, events, and states of affairs people can and cannot categorize and name as they do, and how a system can manage to do that (Harnad 1993b).

So what about the problem of error? Human beings have a certain categorizing capacity. There are some things they can sort and label reliably and consistently (and, perhaps even by some ontic criterion, "correctly") and some things they cannot. Our mission is to find the mechanism that can generate (and hence explains) what they can and do do.

In microcosm, suppose the world consisted of nothing but mushrooms, and our only subsistence activity consisted of finding and eating them. And suppose the mushrooms came in two varieties that were very similar and interconfusable: an edible and a poisonous variety (and, for the sake of argument, let us say the poisonousness was a matter of degree, so if you only tasted a little bit of a poisonous mushroom you would not die, but would simply become a little sick).

Is there any "problem of error" here? You sample mushrooms. At first they all look alike, but some of them make you sick, so you start calling some "mushrooms" and some "toadstools" and you try to avoid the latter. But whenever you categorize wrongly (getting sick from eating what you took to be a mushroom, or going hungry from abstaining while seeing a friend get nourished from what you took to be a toadstool). *The feedback from the consequences of miscategorization* "trains" your internal (supervised) networks so that you eventually learn to categorize the mushrooms correctly.

Now, the features that your successful internal nets find, and that do the trick for you (keep you from getting poisoned or going hungry) may only be provisional and approximate features: If you deplete your foraging territory and move on, features that have served you well may turn out to be false friends, and your inner nets may have to revise their provisional invariants (perhaps even radically, if things change so much that all prior bets are off) to get you safely nourished again. But it seems quite clear that there is no problem with the *empirical* sense of error: You're wrong whenever you eat the wrong stuff, or fail to eat the right stuff. What mushrooms and toadstools "really" are is not at issue: Just what you've sampled, and what provisional features have managed to get you by. And what you "have in mind" is clearly to eat what you take to be edible and avoid what you take to be poisonous.

Nor is there any problem in principle with generalization for a sufficiently powerful invariance-learning mechanism (though there may, of course, be problems for neural nets, if they do not turn out to have human-scale power in this respect): If the input is underdetermined, features will have to be revised, perhaps even radically revised, in the face of new data (Harnad 1987). We can do it; so we need to find learning devices that can do it too, T3-scale.

The Problem of Underdetermination: Human input is indeed underdetermined, so whatever the winning learning device turns out to be, it will have to have the power to learn invariants from human input (initially mostly sensory) under the same conditions humans face. What's infinitely more underdetermined than human input data, though, is *toy cognitive models* that only do a tiny, arbitrary fragment of what people can do. Scaling up from toys to T3 narrows the degree of underdetermination to the normal degrees of freedom for this branch of reverse bioengineering.

The Problem of Non-Existing Entities: Nonexistent entities are only problems for ontologies, not for T3 grounding:

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 our first zebra [armed with a prior grounded description: "striped horse"]. 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 (Harnad 1992a).

This example suggests why a net alone is not likely to be enough for grounding: Grounded symbols are part of a hybrid system. There has to be a symbolic level at which the higher-order categories formed out of propositional strings are represented. The peekaboo unicorn also shows that there is no problem with "defining" entities that are unobserved, nonexistent, or even unobservable-in-principle.

At the sensorimotor level, category names (elementary symbols) are "bound" to sensory projections via learned invariants ("supervised" by feedback from the consequences of miscategorization). Other symbols are then bound to still other symbols through grounded symbolic propositions ("Zebra" = "Horse" & "Stripes"). For this, as C&C indicate, "structured descriptions" are indeed needed, and that is one of the reasons my model is *hybrid* rather than connectionistic. It is true that "the correlational account *[alone]* cannot fix the meanings of primitives." For that you need the rest of the system too.

Lakoff. Lakoff (this volume) favor the pure connectionist approach, as exemplified by Regier's thesis. A great deal depends, however, on whether the input to Regier's model is *sensory projections*, or merely *descriptions* of sensory scenes, and/or of nervous system activities. With the former, we have a "situated" model (situated in optical projections, and capable, presumably, of performing spatial analyses on them), but it is still not appropriate to describe it as "grounded" (at least not as I use the term, Harnad 1990a), because that term was coined for the grounding of *symbols* -- and on Lakoff's construal, Regier's system has no symbols.

If the input to Regier's model is not sensory projections, but descriptions of them (and/or of their putative neural substrates) then it is a symbolic model after all, but an ungrounded one, with symbolic inputs pseudo-grounded in internal network activities (whereas, of course, they ought to be grounded in the actual scenes they are interpretable as being about). Let us call these two alternatives situated spatial analyzers (SSAs) and ungrounded spatial descriptions (USDs), respectively.

"Concepts" is a theoretical construct; "symbol" is a technical term. The structures and states of Regier's network may or may not meet the criteria for being a systematically interpretable formal symbol system. Let us suppose they do not. Then it is still an open question what those structures and states are. To say they are "concepts" is to jump the gun, somewhat. There are no concepts at the level of neural net activity alone; neural nets are just invariance learners. Concepts involve a much bigger chunk of the cognitive machinery in the hybrid system, including the symbolic component.

The technical matter of whether or not there are formal symbols in Regier's model can be settled, but whether there are *concepts* in it is a much tougher question, and one that certainly cannot be settled by merely making a posit to that effect. On the face of it, the net is whatever it is, and contains whatever it contains, and does whatever it does. If its input is sensory projections and its output is symbol strings that are interpretable as a classification and analysis of the spatial scenes of which the inputs were the sensory projections, then that's what you have: sensory projections in, interpretable symbol strings out -- an SSA. On the other hand, if both the inputs and the outputs to the model are symbolic (and the model itself is, like most neural net models, just a symbolically simulated net, not a real physically parallel and distributed structure) then the chances are good that it will indeed fit the technical definition of a symbol system -- and an ungrounded one (USD), at that (Harnad 1993a).

The question about "concepts," however, is not merely terminological, I think, because for "grounding" you must have an entity that is interpretable as referring to something -- and it is in that something that it is supposed to be grounded. Symbols are the systematically interpretable entities par excellence, and Fodor & Pylyshyn (1988) rightly stress that semantic interpretability is a *systematic* property, not an isolated, punctate one. So if we have systematicity, we have symbols; if we don't have systematicity, it is not clear what we have. But concepts, whatever they are, are surely *stronger* constructs than symbols, grounded or no. So we can't just assume their presence by fiat.

In the hybrid architecture I advocate, there is a second source of constraints (analogous to Regier's) on *grounded* symbol systems, over and above (or rather, under and below) the usual formal syntactic ones. Syntactic constraints operate only on the shapes of the symbol tokens (which are arbitrary in relation to the things they can be interpreted as being about); the second (or rather, the *first*), bottom-up source of constraints in my hybrid system comes from the physical connections formed between the

ground-level symbols (category names) and the sensory projections of the objects (category members) they stand for, as mediated by neural nets that have learned to detect the invariants in the sensory projection that allow the objects to be categorized and named correctly. A Regier net would be a welcome component in such hybrid system -- a component that detects spatial invariants. But I would never speak of activity in that net (N) alone as "conceptual": It's just a spatial analyzer. Concepts involve much more, including not only the raw analog projections themselves (A), but also all the (now doubly-constrained) formal relations between the symbols in the grounded symbolic component (S).

Lakoff writes: "The symbol-grounding approach appears to accept... [the AI-style] symbol-manipulation view of mind, assumes concepts are symbols, and only then seeks to ground the symbols." Well, I don't own the term, but my own symbol grounding approach (Harnad 1990a, 1992a) certainly is not correctly described this way: It is conventional top-down AI that imagines it can do all the substantive work at the symbolic level and can then somehow "ground" the whole system by hooking it up to the world "in the right way" by means of sensory devices. My approach is bottom-up (is there any other way to get off the ground?), starting with analog sensory projections, using nets to find the invariants in those projections that allow object categories to be categorized and named, and *then* those grounded elementary names are combined into propositions that, unlike ordinary symbol strings in ungrounded computation, are constrained *both* by the nonarbitrary shapes that ground them (the net connections to the sensory projections of object categories, via learned invariants) *and* by the boolean rules of symbol composition ("Zebra" = "Horse" & "Stripes").

"Regier's approach suggests that the whole level of symbolic manipulation is unnecessary, since the inferential properties of spatial relations concepts are built into the grounding," writes Lakoff. Unnecessary for sensory spatial analysis, perhaps, but cognitive modelling also needs to scale up to the rest of our cognitive competence -- including *concepts*, which have systematic language-of-thought properties that I cannot discern in Regier's model.

Touretzky. Touretzky (this volume) suggests that "[p]erceptual predicates such as "red" or "striped" can be directly grounded in sensory processes, but conceptual categories such as "horse" cannot." Yet the claim that "horse" is a "conceptual category" that is somehow dissociated from or independent of perceptual categories such as "red" or "striped" is a philosophical one that could use some closer scrutiny: What is the evidence for it? Indeed, what would even *count* as evidence for or against it? Surely not introspections about what horses are and what "horse" means! On the face of it, we people can correctly sort and label red things, striped things, squares, triangles, horses, zebras, unicorns, games, true sentences, good things and beautiful things. A Martian behaviorist could tell you that about us. The question is: *How*? And I don't think it is at all obvious that some of these categorizations are accomplished in a radically different way from the others.

Let us not forget that our categories are interrelated, indeed, to a great extent hierarchical. The "subordinate/superordinate" relation is just an arbitrary entry-point into a vast, systematic category network. The systematicities can be *described* symbolically, to be sure, but according to grounding theory, they are *enforced* another way, namely, either by direct, nonsymbolic, invariance filtering of the sensorimotor projection, or by boolean recombinations of category names that are themselves either grounded directly or grounded in category names that are grounded in category names that are grounded directly (i.e., nonsymbolically). Otherwise, all these names would be trapped in that ungrounded symbolic circle I called the "dictionary-go-round" (in Harnad 1990): all systematically interpretable to *us*, of course, but intrinsically meaningless in themselves. *That*'s the symbol grounding problem.

Touretzky attributes to me "the story... that we have a layer of transducers at the bottom to associate physical phenomena with primitive concepts, and from there we proceed upward via symbolic composition to increasingly abstract concepts... zebra defined as horse plus striped." But for me transduction is *not* direct mapping of sensations onto symbols: It is any form of transformation of sensory energy states. The transformation could be analog: Sensory projections could go into other sensory projections, or into motor projections. Symbols need never intervene. In the case of categorization they *do* intervene, but not through the "simple bottom-up mapping of sensations to symbols," but through the laborious learning of sensory invariants on the basis of feedback from the consequences of miscategorization (or, as Touretzky points out, partly also as a result of already tuned invariance detectors shaped by evolution and now inborn). In my view, our stripe detectors might be largely innately tuned (so their path to the arbitrary label "striped" might be easy), whereas our horse-detectors required some nontrivial

learning to be grounded. Once both are grounded, though, they open a purely symbolic path to "zebra."

Touretzky reminds us that "grounded symbols" are used in a variety of abstract and figurative ways not directly related to their grounding. I agree. But in the bottom-up hybrid system I am advocating, they -- and all the symbol combinations they enter into -- continue to be constrained by their origins in sensory grounding. The grounded symbolic component certainly does need to be analyzed and elaborated beyond the vague notion of boolean recombinations of grounded category names, but it is precisely the question of how their analog grounding continues to exert its special influence on the combinatory possibilities of what would otherwise just be arbitrary-symbol-token manipulations that is the crucial question about this hybrid mechanism.

Touretzky writes of a "perceptual schema" for a heart. I'm not sure what this would be (it sounds like a diagram for a homunculus to look at and use), but what a system with a grounded symbol for "heart" would have to be able to *do* (at the very least) is to discriminate and identify literal hearts (those cardiac biological organs) as we do. Once that was successfully modelled, we could worry about its metaphorical extensions (Harnad 1982, 1993b). Its literal meaning may not figure directly in most everyday adult uses of the symbol "heart," but if my grounding theory is right, its *grounding* still underlies all those uses. Does Touretzky really think you can bootstrap to metaphorical uses without a firm anchor in literal uses?

Touretzky writes: "If external observers assign meanings to the agent's symbols they will find that its computations produce meaningful symbolic results." This is a reasonable goal for a builder of useful machines, but not for someone who wants to model thinking, which is grounded in what it actually means, irrespective of what meanings external observers assign to it. According to my theory, that grounding comes from the robotic capacity to discriminate (categorize, manipulate etc.) the objects, events and states of affairs that the system's symbols are interpretable as being about, and it is embodied (mostly) in the transducer structures and processes that give the robot that capacity. Anything less would just be symbol hermeneutics, hanging from a skyhook.

REFERENCES

Dietrich, E. (1993) The Ubiquity of Computation. *Think* (Special Issue on Machine Learning) (in press)

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

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

Harnad, S. (1982) Metaphor and mental duality. In: *Language, mind and brain*. (T. Simon & R. Scholes, eds., Hillsdale NJ: Erlbaum), 189 - 211.

Harnad, S. (1987) The induction and representation of categories. In: 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) 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* (Special Issue on Machine Learning) (in press)

Harnad, S. (1993b) The Origin of Words: A Psychophysical Hypothesis In Durham, W & Velichkovsky B (Eds.) Muenster: Nodus Pub.

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

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.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on Virtual Mind. *Minds and Machines* 2: 217-238.