Reconstructing Physical Symbol Systems

David S. Touretzky and **Dean A. Pomerleau**

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213-3891
email: dst@cs.cmu.edu

Cognitive Science **18**(2):345-353, 1994.

INTRODUCTION

In attempting to force ALVINN[1] into their already bulging symbolist tent, Vera and Simon (1993) have burst the seams of the Physical Symbol System Hypothesis (Newell and Simon, 1976; Newell, 1980a). We would like to present another view of representation in connectionist networks and respond to statements about ALVINN by both Vera and Simon and Greeno and Moore (1993). In so doing, we will reconstruct the PSSH.

The primary error in Vera and Simon's argument is to mistake *signal* for *symbol*: "We call patterns symbols when they can designate or denote" (p. 9). This leaves no possibility of non-symbolic representations. On this view, patterns of neuronal activity in the retina (p. 125), and presumably even the elementary signalling conventions of doorbells and vending machines (Clancey, 1993, p. 89) are elevated to symbol systems, rendering the concept of "symbol" meaningless. Vera and Simon also refer to ALVINN in particular as a "connectionist symbol-processing system" (p. 79), and note that it is simulated by a von Neumann symbol processor. This conflates what should be two distinct levels of representation, and leads to the mistaken conclusion that any implementation of a Turing-computable function is a symbol processor. We believe that symbol processors are a very special type of device; this is what gives the PSSH significance as a theory of intelligence.

PHYSICAL SYMBOL SYSTEMS

A concise and elegant formulation of symbol systems can be found in (Harnad, 1990), drawing on earlier work by Newell, Simon, Fodor, Pylyshyn, and others. Two key clauses in Harnad's definition distinguish symbols from signals:

- Symbols have arbitrary shapes unrelated to their meanings.
- Symbol structures are recursively composable by rule, forming a combinatorial representation.

Why should the shapes of symbols be arbitrary? So that symbols can designate anything at all:

[1] ALVINN is the neural net road follower component of CMU's autonomous vehicle, Navlab (Pomerleau et al., 1991; Pomerleau, 1993).

> A symbol may be used to designate any expression. That is, given a symbol, it is not prescribed a priori what expressions it can designate. This arbitrariness pertains only to symbols; the symbol tokens and their mutual relations determine what object is designated by a complex expression. (Newell and Simon, 1976)

Another reason to make symbols arbitrary is so that the problem of inference cannot be finessed by stipulating special causal properties of symbols with just the right shapes. Arbitrariness insures that inference rules can only rely on the *identities* of symbols; there can be no magical properties hidden in their physical realization.

Why must symbol structures be composable? A central claim of the Physical Symbol System Hypothesis is that combinatorial expressive power is a prerequisite for general intelligent action.[2] Recursively composable structures constitute a *representation language*, and we believe this is what Clancey (1993) meant by his use of the term "linguistic representation." Some specialized information processing tasks do not require a combinatorial representation language, but the PSSH tells us that such a language is unavoidable in a generally intelligent agent.

Analog numerical representations violate both of the above requirements for a physical symbol system. First, they lack the capability of arbitrary designation, because they are constrained to maintain an analogical relationship to the thing they represent. Consider, for example, the homomorphism between voltage levels in a thermostat and temperature values in a room. Each voltage level is a distinct analog pattern, but not an arbitrary one. And relationships *between* analog patterns (e.g., that voltage level x denotes a warmer room state than voltage level y) are not defined explicitly by symbol structures or symbol manipulation rules. Rather, they are predetermined by the causal structure of the thermostat sensor—the source of the analogy between voltage and temperature.

Analog representations also appear to lack combinatorial power. One may speculate about uses for fractal structures or chaotic attractors in knowledge representation schemes, but it is not evident to us how to achieve useful compositionality in a low dimensional, analog numerical system.

We do not mean to suggest that all connectionist networks are necessarily nonsymbolic. Some models do address the requirements for being a symbol system. Arbitrary designation is certainly achievable, and the issue of compositionality has been explored by Touretzky (1990), Smolensky (1990), Pollack (1990), and Elman (1991). The latter two have demonstrated that recurrent backpropagation networks can construct representations with at least limited recursive combinatorial power. But unlike the case in symbolic systems, compositionality in connnectionist networks need not be discrete and concatenative (van Gelder, 1990).

LEVELS OF PROCESSING

We agree with Vera and Simon that the scope of the PSSH is by no means limited to use of symbols in conscious reasoning. It seems quite plausible to assume, for example, that human linguistic behavior is a result of extensive symbol processing at a level impenetrable to introspection. However, we differ with Vera and Simon on the issue of whether low-level processing of meaningful signals should be deemed symbol processing. In particular, processing schemes that do not employ a combinatorial representation should not be regarded as symbol systems.

[2] This idea is widespread in cognitive science. See (Fodor and Pylyshyn, 1988) for its relevance to connectionist models.

> If the domain is narrow enough, considerable intellectual power may be possible from systems that are not physical symbol systems. (Newell, 1980a, p. 171)

Rather than acknowledge this possibility, Vera and Simon have mistakenly tried to define all processing as symbolic. In describing Brooks' creatures, they say that "sensory information is converted to symbols, which are then processed and evaluated in order to determine the appropriate motor symbols that lead to behavior" (p. 34). Yet elsewhere they confer symbolic status on retinal images themselves, noting that:

> As one moves upstream in the visual system, increasing levels of abstraction and combination are found in the representation of an image. It is wholly arbitrary to call one step in this sequence symbolic, but deny that label to its predecessor. (Vera and Simon, 1993, p. 125)

We disagree. A representation becomes a symbol system at the point where it acquires combinatorial power. As one moves upstream in the visual system, processing splits into several pathways (Kandel, Schwartz, and Jessell, 1991, ch. 29). One of these leads, via the lateral geniculate and primary visual cortex, to higher cortical areas, where information is believed to be recoded in symbolic form. It is here, in the cortex, that object recognition takes place. Another path leads from the retina to the pretectal area of the midbrain, where pupillary responses are generated. And a third set of pathways involves the frontal eye fields, superior colliculus, cerebellum, and pons. This signal processing system, with additional cortical input, is responsible for saccadic eye movements and the smooth tracking response to moving objects. If everything from the retina on up were a symbol system, i.e., if reflexive object tracking were symbol processing, then "symbol" would mean nothing more than "signal."

Elsewhere in their reply to Clancey, Vera and Simon state (p. 126): "We are aware of no evidence ... that research at the neuropsychological level is in conflict with the symbol-system hypothesis." There can be no such evidence. The PSSH makes what is today a very modest claim:

> *The Physical Symbol System Hypothesis.* A physical symbol system has the necessary and sufficient means for general intelligent action. (Newell and Simon, 1976)

Sufficiency is a given, unless one is a dualist or some other sort of mystic, because physical symbol systems are Turing-universal. Thus, the PSSH is so weak that the only way it could conflict with "research at the neuropsychological level" is by a violation of the necessity clause. If a a generally intelligent organism were found to have no symbol system hidden anywhere inside it, the PSSH would be refuted. But this is extremely unlikely given our present understanding of humans and animals. In fact, the necessity of having the capabilities of a symbol system is tautological when one examines Newell and Simon's definition of "general intelligent action." This does not make organisms *nothing but* symbol systems; in biology there are practical reasons for performing information processing functions using "direct", non-symbolic means where possible.

THE SIGNIFICANCE OF ALVINN

ALVINN, the neural net road follower, employs analog representations for both steering direction and hidden layer activity patterns. Its patterns are not arbitrary-shaped symbols, and they are not combinatorial. Its hidden unit feature detectors are tuned filters. Their activation denotes the presence of visual features in the view ahead of the vehicle, and analysis of their input weights and response properties reveals that they are responding to objects such as road edges and lane stripes (Pomerleau and Touretzky, 1993). Greeno and Moore (1993, p. 54) claim there is no semantic interpretation of these hidden unit activity patterns. But their reponse, albeit analog in nature, is just as meaningful as the discrete output LANE-STRIPE-DETECTED.

What do Greeno and Moore mean when they say of ALVINN's representations that "there is no process of semantic interpretation by which these patterns are given referential meaning"? They understand "referential meaning" as a three-way predicate involving an arbitrary pattern, the thing it designates, and the agent for whom this designation holds. They view NAVLAB's symbolic map component as semantically interpretable in this sense. Presumably this property does not depend on having a human-like agent doing the interpretation, so the crucial distinction between NAVLAB's neural net and symbolic map components is that in the latter, the relationships between patterns and referents are arbitrary; the map component *requires* an interpretation to connect its arbitrary formal symbols to the environmental features they designate. But ALVINN has no capacity for arbitrary designation; its activation patterns are *analogically* related to specific states of the world. ALVINN is thus, we agree, an example of direct perception with no interpretation required. Its symbols have intrinsic, not referential meaning.

What is the relevance of ALVINN to the situated action debate? SA theorists suggest that significant levels of intelligent action can be achieved nonsymbolically, i..e, by "direct" means. But we would not take ALVINN's success as evidence for this claim. Its perceptual abilities are modest, and its reasoning abilities nonexistent. Unlike Vera and Simon, our objection to the SA position is not that ALVINN is really a symbol system, but rather that direct, nonsymbolic systems are not powerful enough to succeed outside highly restricted domains. This is the essential claim of the PSSH.

NON-ARBITRARY SYMBOLS

The requirement that symbols have purely arbitrary shapes unrelated to their meanings is probably not fully met in humans, not because a pure physical symbol system would be inadequate to the task of cognition, but because of the process by which humans were constructed. Cognitive science has long emphasized the importance of selecting good representations. Why then should evolution not incorporate properties of the world into our realizations of symbols, despite the fact that the Turing-universality of symbol systems makes it theoretically unnecessary to do so?

Implicit or analog properties of a symbol might be viewed as affordances. They could be exploited by nonsymbolic, pre-conscious processors, and perhaps also by symbolic processors seeking computational shortcuts. For example, the symbols LION and TIGER[3] might well have similar neural realizations if Nature finds that useful in implementing real-time intelligent behavior.

Some writers would go further than this. Harnad (1992) argues that the grounding of symbols in analog sensory predicates is an essential feature of human intelligence, affecting even our most abstract conceptualizations. He says of grounded symbols:

> [I]n 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… [I]t 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. (Harnad, personal communication)

[3]If such symbols exist truly exist. See the discussion of symbols vs. concepts that follows.

Situated action, to us, implies exploiting properties of the world in constructing representations, reducing the need to rely on formal symbol processing for all of reasoning. By "formal symbol processing" we mean what physical symbol systems do: rule-based manipulation of arbitrary-shaped tokens recursively composed into combinatorial structures.

SYMBOLS VS. CONCEPTS

The distinction between primitive symbol tokens and composite expressions, or "symbol structures", reflects Newell's conception of a PSS as a Turing machine (see Newell 1980a), and the common view of compositionality as concatenation (van Gelder, 1990) . However, it is important to remember that cognitive theories deal in concepts, not symbols (Lakoff, 1993). The atoms we call symbols in today's AI programs are placeholders (names) for concepts. The concepts in our heads, such as "boy" or "give", are hardly atomic. They are complex webs of reference to other concepts and sensory impressions, with shapes that are far from arbitrary. According to the classical formulation of symbol systems, there must then be a lower level of representation, the primitive symbol level, from which such concepts are composed.

We cannot assign concise meanings to elements at this lower level. Whatever phrase we choose would correspond to a concept structure, not a symbol. Except for those symbols that serve as names for familiar concepts, primitive symbol tokens would appear to be indescribable. A comparable notion exists in the connectionist literature, where "subsymbolic representations" are postulated whose components are "microfeatures." The subsymbolic level is really a *subconceptual* level, except in the most radical connectionist proposals where symbols themselves are hypothesized to be an emergent property of complex dynamical systems (Smolensky, 1988).

Rather than accept the classical dichotomy between primitive symbol tokens and composite structures, we could instead work solely with "patterns", as in the patterns of activation in a connectionist net. We could stipulate that some of these patterns play the role of symbols when they function as names for things, as in Hinton's notion of *reduced descriptions*: patterns that simultaneously point to more complex patterns and summarize their properties (Hinton, 1990). In order to retain the power of arbitrary designation all we really require is that at least some arbitrary mappings be possible. This connectionist reformulation of symbol processing is not yet as fully developed as the classical concatenative view, but it is an intriguing alternative.

AUTOMATICITY

A surprising result of recent connectionist work is the wide variety of behaviors obtainable from simple nonlinear mapping functions, such as three-layer backprop nets. ALVINN is such a network. It learns to drive a vehicle, in traffic, at speeds of up to 55 miles per hour. And it forms internal representations of the road and the steering direction. But they are analog representations from start to finish; there are no symbolic intermediaries. This does not violate the PSSH, as we make no claim that a non-symbolic theory could account for *general* intelligent action. Nor do we claim that physical symbol systems can't simulate the operations ALVINN performs. Clearly they can, just as a Turing machine can simulate a finite state machine.

The success of connectionist networks in limited domains leads us to reexamine Vera and Simon's notion of automaticity as chunking. Chunking constructs new symbolic rules from combinations (or instantiations) of old ones. The new rules are able to perform the same task in fewer steps, or perhaps with fewer resources if variable binding is a bottleneck (Newell, 1980b). It seems quite reasonable to propose that repeated conscious behavior would lead to the generation of chunks that could later execute subconsciously, resulting in automaticity for the task. And repeated execution of these subconscious rules could produce further chunking, yielding additional improvements in performance with practice. For inherently symbolic domains such as language use or chess playing, chunking is an attractive theory of learning, though not the only theory.

Where we part company with Vera and Simon is the application of chunking to non-symbolic domains, such as motor skills. Why assume that bicycle riding, for example, is a symbol processing task? We see nothing in the structure of bicycle riding that requires arbitrary designation ability or combinatorial expressive power. Rather, effective bicycle riding is likely to involve a high-dimensional mapping from perceptual, kinesthetic, and vestibular space to motor space, mediated perhaps by the cerebellum. One must of course augment this mapping with higher level cognitive functions such as the ability to read street signs and plan a route, but the "automatic" component of bicycle riding need not be in any way symbolic. Fine tuning this high-dimensional mapping would require a large number of training instances, which could potentially give rise to a power law of practice without generating an exponential number of discrete symbolic chunks.

We are led to a model of cognition in which conscious, deliberate symbol manipulation is the top level, *sub*-conscious symbol processing the intermediate level, and specialized non-symbolic modules appear at the lowest level.[4] Some of our neural pathways proceed from sensors to effectors by way of the symbolic level, but others bypass the higher cortical areas entirely. These latter may proceed directly from receptors to motor neurons in the case of a simple reflex arc, or via processing modules that are quite sophisticated but not symbolic, as in the pathways that generate automatic eye movements. These direct pathways are unavailable to introspection, but they *are* subject to overriding by higher level brain areas when a person makes a conscious effort to control what would normally be unconscious or reflexive behavior.

The seamless integration of these three processing levels is a hallmark of human phenomenological experience. That is why it is so difficult for us to say what we are doing when we automatically drive a familiar route while preoccupied with other matters, or walk down the hall with our nose buried in the latest issue of *Cognitive Science* yet still manage to avoid obstacles.

CONCLUSION

We agree with much of Vera and Simon's critique of situated action theory, but reject their characterization of ALVINN as a symbolic system. In this article we have tried to make clear what "non-symbolic" means, thereby restoring the force of the Physical Symbol System Hypothesis.

How much of intelligent behavior can be implemented by non-symbolic means? The PSSH tells us that *not all* the processing in a generally intelligent agent can be non-symbolic. SA theorists argue—unconvincingly, we feel—that "not all" will turn out to be "quite a lot." The success of systems such as ALVINN justifies replying "at least some."

[4] Greeno and Moore (1993, pp. 55-56) seem drawn to a similar conclusion.

Acknowledgements

We thank Stevan Harnad, Herb Simon, and James Greeno for helpful discussions. This work was supported in part by a grant from the Fujitsu Corporation.

References

Clancey, W. J. (1993). Situated action: a neuropsychological interpretation. *Cognitive Science, 17*, 87-116.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-225.

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

Greeno, J. G., & Moore, J. L. (1993). Situativity and symbols. *Cognitive Science, 17*, 49-59.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*, 335-346.

Harnad, S. (1992). Connecting object to symbol in modeling cognition. In A. Clarke and R. Lutz (Eds.), *Connectionism in context*. Springer Verlag.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence, 46*, 47-75.

Kandel, E. R., Schawrtz, J. H., & Jessell, T. M. (1991). *Principles of neural science*. Third edition. New York: Elsevier.

Lakoff, G. (1993). Grounded concepts without symbols. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 161-164. Hillsdale, NJ: Erlbaum

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM, 19*, 113-126.

Newell, A. (1980a). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Newell, A. (1980b). Harpy, production systems, and human cognition. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence, 46*, 77-105.

Pomerleau, D. A., Gowdy, J., & Thorpe, C. E. (1991). Combining artificial neural networks and symbolic processing for autonomous robot guidance. *Engineering Applications of Artificial Intelligence, 4,* 961-967.

Pomerleau, D. A. (1993). *Neural network perception for mobile robot guidance*. Boston: Kluwer.

Pomerleau, D. A., & Touretzky, D. S. (1993) Understanding neural network internal representations through hidden unit sensitivity analysis. In C. E. Thorpe (Ed.), *Proceedings of the international conference on intelligent autonomous systems*. Amsterdam: IOS Publishers.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159-216.

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

van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science, 14*, 355-384.

Vera, A. H., & Simon, H. A. (1993). Situated action: a symbolic interpretation. *Cognitive Science, 17,* 7-48, and replies in 77-86 and 117-133.