Is Distributed Connectionism Compatible with the Physical Symbol System Hypothesis?

Mark Derthick and David C. Plaut

Computer Science Department
Carnegie-Mellon University

In *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (pp. 639-644). Hillsdale, NJ: Erlbaum, 1986.

1. Introduction

All existing intelligent systems share a similar biological and evolutionary heritage. Based on the conviction that cognition is computation, artificial intelligence researchers are investigating computational models as a means of discovering properties shared by all intelligent systems.

One property that has been proposed as central to intelligence is the ability to construct and manipulate symbol structures. If intelligence may be described completely in terms of symbol processing, then cognitive science need not be concerned with the particular physical implementation details of either artificial or biological examples; neuroscience would no longer be part of cognitive science. On the other hand, if important aspects of intelligence evade symbolic explanation, it may prove necessary to consider phenomena below the symbol level. The connectionist approach to artificial intelligence is founded on the conviction that the structure of the brain critically constrains the nature of the computations it performs. However, if the symbolic position is correct and neural networks only implement symbol systems, then connectionism contributes little to cognitive science.

The notion of intelligence as symbol processing was made explicit by Newell and Simon with the Physical Symbol System Hypothesis (PSSH) [Newell & Simon 76, Newell 80] and the Knowledge Level Hypothesis (KLH) [Newell 82]. Taken together, these hypotheses have significant implications for the nature of any system capable of general intelligence. We examine a number of connectionist systems in light of the hypotheses and distinguish three kinds: (1) *rule-based* systems, which are symbol systems; (2) *rule-following* systems, which are symbol systems only under a weakened version of the PSSH; and (3) systems which are not rule-following, and thus are not symbol systems even in a weak sense.

According to the PSSH, non-symbolic connectionist systems must be incapable of general intelligence. There are strong arguments both for and against this conclusion. On the one hand, such connectionist systems may provide more parsimonious accounts of certain cognitive phenomena than do symbolic approaches. On the other hand, these connectionist systems have significant limitations, relating to universality, not shared by symbol systems. We conclude that a comprehensive theory of intelligence may require a hybrid model that combines the strengths of both approaches.

2. Physical Symbol Systems

The PSSH states that a physical symbol system, defined as "a machine that produces through time an evolving collection of symbol structures" [Newell & Simon 76, p. 116], has the necessary and sufficient means for general intelligent action. Newell and Simon explain,

> By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. [Newell & Simon 76, p. 116]

Of course, since symbol systems are universal they are sufficient for carrying out *any* behavior. To accept the sufficiency condition on this basis, however, would be to fall prey to the "Turing tarpit" [Newell 80], in which significant structural differences are blurred under the notion that all universal systems are equivalent. In order for the sufficiency claim to be substantive, the further organization required to exhibit general intelligence must not resort to simulating a non-symbolic system.

Physical symbol systems[1] are composed of

1. a collection of *symbols*, each a discrete, identifiable physical pattern in a machine;
2. *symbol structures*, or expressions, composed of symbols related in some physical way;
3. *processes* operating on expressions to produce other expressions.

A symbol structure *designates* another symbol structure or process if having the first structure allows behavior that affects or depends on the second. The system can *interpret* an expression if the expression designates a process and if, given the expression, the system can carry out the process. This formulation has a number of consequences for the nature of any physical symbol system [Newell & Simon 76], the most important for our purposes being:

1. A symbol may be used to designate any expression whatsoever.
2. There exist expressions that designate every process of which the machine is capable.
3. The number of expressions that the system can hold is essentially unbounded.

We will argue in the next section that these characteristics exclude certain connectionist systems from the class of physical symbol systems.

The semantics of a physical symbol system is developed in the context of the relation between the symbol level and a higher, *knowledge level*, in which the behavior of the system is described in terms of knowledge, goals, and actions [Newell 82]. The Knowledge Level Hypothesis states that the knowledge level is implemented directly by the symbol level. Knowledge level entities are represented by particular symbol level structures, and each symbol structure has a coherent interpretation at the knowledge level. In other words, symbols and symbol structures are the formal entities of a physical symbol system that are given a semantic interpretation.

With these characteristics of physical symbol systems in mind, we turn to an analysis of the relationship between symbol systems and connectionist systems.

3. Connectionist Systems

We take the essential characteristic of a connectionist system to be the existence of a physical level description in terms of the operation of a large number of very simple computing devices (units) locally interacting across very low bandwidth channels (connections). Such an architecture is quite different from that underlying standard physical symbol systems. However, the fact that connectionist systems are built out of units and connections does not bear on the question of whether they are symbol systems. The critical issue here is the relation between formal operations of the system and what they represent. In this regard, certain types of "localist" connectionist systems [Feldman & Ballard 82, Cottrell 84, Shastri & Feldman 84], in which the activity of individual units may be given a coherent knowledge level semantics, meet the requirement of physical symbol systems that the formal level map directly to the semantic level.

On the other hand, "distributed" connectionist systems [Hinton & Anderson 81, McClelland & Rumelhart 86], in which a knowledge level semantics is ascribed only to *patterns* of activity of a large number of units, present difficulties for any attempt to assign to the system the type of formal semantics required of physical symbol systems. In particular, the formal level of the system (i.e. the interaction of units via connections) and the semantic level (i.e. the interactions of patterns of activity) do not correspond, nor do they operate according to the same principles. While units obey formal input/output rules (specified in terms of unit activities and connection strengths), the interaction of patterns of activity *as patterns* need not be formal. There are various ways for the interaction of patterns in such systems to generate knowledge level behavior. In the following sections we make three distinctions among distributed connectionist systems: *rule-based* systems, with explicitly encoded rules; *rule-following* systems, with implicitly encoded rules; and systems whose behavior is not strictly rule-following.

---

[1] Our analysis is based solely on the formulation of symbol systems developed by Newell and Simon, and does not exclude the possibility of alternative formulations which might encompass the connectionist systems we discuss.

3.1. Explicit Rules

We first examine a distributed connectionist system in which the interaction of the patterns of activity *is* governed by explicit rules and is therefore formal. Touretzky and Hinton (1985) have developed a connectionist implementation of a production system. Production memory consists of sets of units, each dedicated to a particular production. The units in each set are wired to units which are active in the representation of working memory elements that are matched, added, or deleted by the production. Although the behavior of the system can be explained in terms of interactions between individual units, a higher level explanation making reference to the production rules can be used. Furthermore, this explanation is not just a way of speaking; it corresponds directly to physical structure. This is just what Newell and Simon expect will be the case for any intelligent system: the symbol level may be implemented in various technologies, but it is a real, necessary system level. The "Touretzky tarpit" (in contrast to the Turing tarpit) traps those who gratuitously distinguish this system from symbolic ones by attributing psychological importance to its underlying connectionist basis. It *is* symbolic, and any theory based on it could as well be non-connectionist.

3.2. Implicit Rules

In contrast, a connectionist system developed by Rumelhart and McClelland (1986) exhibits rule-following behavior without containing explicit representations of the rules. The task is to form a phonological representation of the past tense of English verbs from a phonological representation of the present tense form. Linguists typically model this task with a large number of rules, which form a hierarchy of exceptions, exceptions to exceptions, and so on. The rule for regular verbs consists of adding /ed/. Irregular verbs may be grouped according to other rules, such as changing /ing/ to /ang/ (*sing*/*sang*), changing /d/ to /t/ (*build*/*built*), etc.

Instead of using explicit rules, Rumelhart and McClelland's system captures the rule following nature of forming past tenses in terms of regularities between the substructures of the phonological codes of the present and past tense forms. This substructure is represented in terms of the activity of a set of units, each representing a context-sensitive triple of phoneme features (called "Wickelfeatures" in the spirit of Wickelphones [Wickelgren 69]). The important characteristics of this code are that it can sufficiently discriminate between any two English verbs, and that it provides a natural basis for generalizations to emerge about what aspects of a present tense form correspond to what aspects of the past tense form [Rumelhart & McClelland 86].

In the model, a fixed encoding network converts the actual phonological representation of the present tense form into a slightly blurred pattern of activity over a large set of *input* units, each representing a particular Wickelfeature. Each input unit is connected to each member in a similar set of *output* units for representing the phonological substructure of the past tense form. The activity of the output units is then decoded by a second fixed network into its corresponding phonological representation. The goal of the network is to produce in the output units the pattern of activity representing the past tense form given the pattern of activity over the input units for the present tense form. The network is presented with the codes for a large number of present/past tense pairs, and a learning algorithm modifies the strengths of the connections between the input and output units to reduce for each pair the difference between the correct phonological representation and the one produced by the network. As it learns the task, the performance of the network passes through three important stages (which we describe below), eventually producing the appropriate rule-following and exception behavior, as demonstrated by proper generalization to verb pairs not seen in the training.

The way in which Rumelhart and McClelland's system produces rule-following behavior violates an important constraint on the structure of a physical symbol system; processes in the system do not have symbol structures which designate them. Regularities between the substructure of present and past tense forms are encoded (in connection strengths) in terms of the interaction of "microfeatures" (in this case, Wickelfeatures). The presence of each microfeature in the input representation provides support for some microfeatures in the output representation while inhibiting others. These "microinferences" allow the substructure of the input to be combined in very complex and subtle ways to produce the appropriate substructure for the output. The actual semantic rules (which the input/output activity patterns can be described as following) are nowhere stated explicitly, but emerge from complex interactions among the

microinferences [Hinton 81].

The lack of explicit rules excludes this system from the class of physical symbol systems. However, rule-following connectionist systems are compatible with a weaker version of the PSSH. It could still be maintained that intelligence can be explained with a rule-based system, regardless of the fact that it is also possible to explain intelligence in terms of a system which is only rule-following. This corresponds to interpreting the symbol level primarily as a means of *explaining* knowledge level behavior, and not necessarily as the means of *implementing* it.

3.3. Not Rule-following

A major contribution of Rumelhart and McClelland's system is modeling the stages that children pass through in acquiring the ability to form past tenses: (1) an initial stage in which all past tenses are learned as separate words; (2) an intermediate stage in which an insufficient number of rules are used to form all past tenses, resulting in overregularization; and (3) a final stage in which more and more rules are learned, so all known verbs are handled correctly and novel verbs generalize appropriately. While the stages are relatively well defined, the transition from one stage to another is gradual, so that at times a child may use several past tense forms of the same verb in the same conversation. Such behavior is difficult to account for using rules, but is explained quite elegantly (in terms of competing microinferences) in the connectionist model. While the system can be described as following rules once the ability to form past tenses has been fully learned, the system viewed as evolving over time cannot be given an adequate formal symbol level *explanation*, and hence is not a symbol system even in the weaker sense described above.

Of course, a rule-based approach with a very large set of highly interacting, fine-grained rules which fire in parallel might succeed in reproducing such graded behavior.[2] In general, rule-based systems embody symbol systems, and so by virtue of their universality they are, in principle, capable of reproducing *any* behavior (the Turing tarpit argument). Yet as the complexity of such systems increases, it becomes more and more difficult to give a clear account, in terms of the task or environment, of what a single rule is doing. More and more of the structure of a rule depends on the entire set of rules with which it interacts. The system becomes a less and less parsimonious *rule-based* explanation, and more and more similar to a microinference-based *connectionist* explanation.

4. Limitations of Connectionist Systems

The PSSH states that only symbol systems are capable of general intelligence. As we have shown, there are connectionist systems that are not symbol systems even in a weak sense. Therefore, either the PSSH is false, or connectionist systems of this type cannot be extended from limited domains to general intelligence. We contend that non-symbolic connectionist systems *are* limited with respect to symbol systems: they are not universal, and hence lack the flexibility to potentially carry out any possible behavior.

Any computational system is implemented by what Pylyshyn (1980) calls the *functional architecture*; "the fixed functional capacities provided by the biological substrate...out of which cognitive processes are composed." Pylyshyn (1984) argues that the essential difference between connectionist systems and symbol systems hinges on significant differences in their functional architectures. In a symbol system, an unbounded number of symbols may be manipulated by a finitely specified control. In a connectionist system, *all* of the knowledge used to carry out a process is contained within a finite structure; control and data are not separate. The unbounded storage capacity that underlies the universality of symbol systems is, in principle, not available in the specification of connectionist systems.

It could be argued that if an unbounded pool of uncommitted units were available, and if learning could take place during the execution of an operation, then it might be possible to store a potentially unlimited amount of knowledge. However, allowing learning to be a primitive operation of the functional architecture would be allowing universality in by the back door. By definition, the capabilities of the

---

[2] In fact, the evolution of large expert systems, such as R1 [McDermott 82, Bachant & McDermott 84], seems to lead in this direction.

functional architecture are fixed. In a connectionist system learning is a method for *modifying* the functional architecture and cannot be controlled at the semantic level.

Yet Touretzky and Hinton's system *can* add a potentially unbounded amount of knowledge without the use of learning. Productions match clauses represented in *clause spaces*, which are groups of units that can represent one clause at a time. The clauses which correspond to stable patterns are not determined by connection strengths, but by the states of units in working memory space. Connectionist systems require learning to modify weights, but they modify states during ordinary operation. Thus it is possible in this system to store an arbitrary amount of knowledge in working memory and have it appropriately affect clause retrieval. Changing the states of working memory units is done by the production memory units, so universality appears exactly when the division between control and data does.

5. A Hybrid System

We have seen how generalization arises naturally in non-symbolic connectionist system, but that such systems are limited with respect to the universality of symbol systems. In order to determine the relative significance of these two approaches, it is necessary to characterize the processes that must be accounted for by a comprehensive theory of intelligence.

Pylyshyn (1980) argues that the class of processes that psychology is committed to explain are those that are "cognitively penetrable"; processes whose function depends on the agent's beliefs and goals in meaningful ways. Any function that is cognitively *impenetrable* represents a fixed primitive operation of the functional architecture. By definition, these primitive operations are not symbolic. Therefore, it is possible to suppose that, for example, past tense formation is primitive. In general, any process found to require a non-symbolic explanation may be relegated to the functional architecture. However, if too much of cognition is swept under the primitive rug, there is nothing left for a theory of intelligence to do: psychology becomes trivial.

Pylyshyn claims that *all* cognitively penetrable processes are symbolic. We believe that processes like past tense formation are both cognitively penetrable and best carried out non-symbolically.

Therefore, we claim that a comprehensive theory of intelligence will require both symbolic and non-symbolic processes. The relative strengths and weaknesses of connectionist and symbolic approaches suggests a natural division of labor within such a hybrid system. A large collection of task-specific connectionist modules would carry out overlearned processes under the executive control of a flexible symbol system. The fact that this division is aligned with a number of classic dichotomies used to characterize various aspects of cognition (e.g. "controlled vs. automatic" [Shiffrin & Schneider 77], "central vs. peripheral" [Fodor 83]) suggests that it may reflect a real structural property of the cognitive architecture.

6. Conclusion

We have shown that there are connectionist systems which generate knowledge level behavior but fall outside Newell and Simon's characterization of physical symbol systems. The claim of the Physical Symbol System Hypothesis that symbol processing is necessary for intelligence implies that such systems cannot be extended from simple domains to general intelligence. The lack of universality of non-symbolic connectionist systems supports the claim that some symbolic processing is necessary for intelligence.

On the other hand, non-symbolic connectionist systems, like Rumelhart and McClelland's, provide a more satisfactory explanation of certain aspects of cognition than do symbol systems. This supports the claim of the connectionist approach that implementation details are important to understanding cognition, and is incompatible with the claim of the PSSH that symbol processing is *sufficient* for general intelligence. It is striking and perhaps quite significant that what is natural for each approach is quite difficult and unwieldy in the other. A theory of intelligence incorporating both connectionist and symbolic components would be more capable than either approach alone of integrating the various aspects of cognition into a comprehensive explanation of intelligent behavior.

Acknowledgements

We wish to thank the Boltzmann research group at Carnegie-Mellon University for useful discussions on the relationship between symbol systems and connectionist systems, and David Ackley, Ron Brachman, Hector Levesque, Jay McClelland, Steven Nowlan, Benjamin Pierce, and David Touretzky for providing useful comments. This research was supported by an ONR Graduate Fellowship to the first author, and an R.K. Mellon Fellowship to the second author.

References

[Bachant & McDermott 84]
Bachant J. and McDermott J.
R1 revisited: Four years in the trenches.
AI Magazine 5:21-32, 1984.

[Cottrell 84] Cottrell G.W.
A model of lexical access of ambiguous words.
In *Proceedings of the National Conference on Artificial Intelligence*. Austin, TX, August, 1984.

[Feldman & Ballard 82]
Feldman J.A. and Ballard D.H.
Connectionist models and their properties.
Cognitive Science 6:205-254, 1982.

[Fodor 83] Fodor J.A.
The modularity of mind.
M.I.T. Press, Cambridge, MA, 1983.

[Hinton 81] Hinton G.E.
Implementing semantic networks in parallel hardware.
Parallel models of associative memory.
In G.E. Hinton and J.A. Anderson,
Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[Hinton & Anderson 81]
Hinton G.E. and Anderson J.A. (Eds.).
Parallel models of associative memory.
Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[McClelland & Rumelhart 86]
McClelland J.L. and Rumelhart D.E. (Eds.).
Parallel distributed processing: Explorations in the microstructure of cognition.
Bradford Books, Cambridge, MA, 1986.

[McDermott 82] McDermott J.
R1: A rule based configurer of computer systems.
Artificial Intelligence 19:39-88, 1982.

[Newell 80] Newell A.
Physical symbol systems.
Cognitive Science 4:135-183, 1980.

[Newell 82] Newell A.
The knowledge level.
Artificial Intelligence 18:87-127, 1982.

[Newell & Simon 76]
Newell A. and Simon H.A.
Computer science as empirical inquiry: Symbols and search.
Communications of the ACM 19:113-126, 1976.

[Pylyshyn 80] Pylyshyn Z.W.
Computation and cognition: Issues in the foundations of cognitive science.
Behavioral and Brain Sciences 3:111-169, 1980.

[Pylyshyn 84] Pylyshyn Z.W.
Why "computing" requires symbols.
In *Proceedings, 6th Conference of the Cognitive Science Society*, pages 71-73. Boulder, CO, 1984.

[Rumelhart & McClelland 86]
Rumelhart D.E. and McClelland J.L.
On learning the past tenses of English verbs.
Parallel distributed processing: Explorations in the microstructure of cognition.
In J.L. McClelland and D.E. Rumelhart,
Bradford Books, Cambridge, MA, 1986.

[Shastri & Feldman 84]
Shastri L. and Feldman J.A.
Semantic networks and neural nets.
Technical Report 131, Computer Science Department, University of Rochester, January, 1984.

[Shiffrin & Schneider 77]
Shiffrin R.M. and Schneider W.
Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory.
Psychological Review 84:127-190, 1977.

[Touretzky & Hinton 85]
Touretzky D.S. and Hinton G.E.
Symbols among the neurons: Details of a connectionist inference architecture.
In *Proceedings, 9th International Joint Conference on Artificial Intelligence*. Los Angeles, August, 1985.

[Wickelgren 69] Wickelgren W.A.
Context-sensitive coding, associative memory and serial order in (speech) behavior.
Psychological Review 76:1-15, 1969.