On the Connectionist Reduction of Conscious Rule Interpretation

Paul Smolensky

Department of Computer Science &
Institute of Cognitive Science
University of Colorado at Boulder

Abstract

Connectionist models have traditionally ignored conscious rule application in learning and performance. Conceptual problems arise in treating rule application in a connectionist framework because the level of analysis of connectionist models is lower than that which is natural for describing conscious rules. An analysis is offered of the relation between these two levels of description, and of the kind of reduction involved in connectionist modeling. From this vantagepoint an approach is formulated to the treatment of conscious rule application within a connectionist framework. The approach crucially involves connectionist language processing, and leads to a distinction between two types of knowledge that can be stored in connectionist systems.

Introduction

Connectionist models have traditionally ignored the role played by conscious application of rules in human cognition. Many tasks are learned through rules, and initially performed by consciously applying those rules; connectionist models have by and large been unable to address learning and performance in which such rule application occurs. One of the most striking phenomena in cognitive science is the shift during the acquisition of expertise from conscious processing to processing that I will simply refer to as *intuitive*. While the connectionist approach has provided successful models of purely intuitive processes, since the approach cannot now address rule application, it cannot shed light on this aspect of the transition to expertise.

Incorporating conscious rule application into a connectionist paradigm poses serious conceptual problems that must be resolved, at least provisionally, before the associated technical problems can even be recognized, let alone solved. Many of these conceptual problems stem from the fact that the *level of analysis* adopted by connectionist modeling is lower than the level at which rule interpretation is most naturally described. Conscious rules involve consciously accessible concepts, and are therefore naturally described at the level of these concepts: what I will call the *conceptual level*. In the kind of connectionist system I consider here, concepts are represented by patterns of activity over large numbers of processing units; the semantic interpretation of the individual units is considerably finer-grained than that of the consciously accessible concepts. Such a connectionist system uses *distributed representations* (eg., Anderson & Hinton, 1981; Hinton, McClelland & Rumelhart, 1986; Smolensky, 1986b). The semantics of the individual connectionist processors resides at a level lower than the conceptual level: what I will call the *subconceptual level*.

The cognitive modeling paradigm employing connectionist models with distributed representations, i.e., employing connectionist networks with subconceptual semantics, will be called the *subsymbolic paradigm*; this contrasts with the traditional *symbolic paradigm* that employs symbol manipulating models in which the symbols have conceptual semantics. The incorporation of conscious rule interpretation into the subsymbolic paradigm involves a *reduction* of the symbolic account of rule interpretation to the subconceptual level. Thus the first question to address is: *How do the symbolic and subsymbolic paradigms relate at the conceptual and subconceptual levels of analysis, and what kind of reduction is involved?*

Reduction of cognition to the subconceptual level

Imagine three physical systems: a brain that is executing some cognitive process, a massively parallel connectionist computer running a subsymbolic model of that process, and a von Neumann computer running a symbolic model of the same process. The cognitive process may involve conscious rule application, intuition, or a combination of the two. In Smolensky (1987b) I have characterized the subsymbolic paradigm as positing the following relationships between descriptions of these three physical systems at the neural, subconceptual, and conceptual levels:

(1) a. Describing the brain at the neural level gives a neural model.
b. Describing the brain approximately, at a higher level—the subconceptual level—yields, to a good approximation, the model running on the connectionist computer, when it too is described at the subconceptual level. (At this point, this is a goal for future research. It could turn out that the degree of approximation here is only rough; this would still be consistent with the subsymbolic paradigm.)
c. We can try to describe the connectionist computer at a higher level—the conceptual level—by using the patterns of activity that have conceptual semantics. If the cognitive process being executed is conscious rule application, we will be able to carry out this conceptual level analysis with reasonable precision, and will end up with a description that closely matches the symbolic computer program running on the von Neumann machine.
d. If the process being executed is an intuitive process, we will be unable to carry out the conceptual-level description of the connectionist machine precisely. Nonetheless, we will be able to produce various approximate conceptual-level descriptions that correspond in various ways to the symbolic computer program running on the von Neumann machine.

For a cognitive process involving both intuition and conscious rule application, (1c) and (1d) will each apply to certain aspects of the process.

The relationships (1a) and (1b), which are discussed at some length in Smolensky (1987b), are not relevant for the present considerations; they are mentioned here only to link the physical instantiations of the subsymbolic and symbolic models to the physical system they are in some sense models of. A number of the relations (1d) between subsymbolic and symbolic accounts of intuitive processing have been addressed elsewhere (Rumelhart, Smolensky, McClelland & Hinton, 1986; Smolensky, 1986a, 1986b; see Smolensky, 1987a, 1987b for summaries). The relationship (1c) between a subsymbolic implementation of conscious rule interpretation and a symbolic implementation is the subject of the next section.

The relationships in (1) can be more clearly understood by introducing the concept of "virtual machine." If we take one of the three physical systems and describe its processing at a certain level of analysis, we get a virtual machine that I will denote "system_{level}". Then (1) can be written:

(2) a. brain_{neural} = neural model
b. $\text{brain}_{subconceptual} \approx \text{connectionist}_{subconceptual}$
c. $\text{connectionist}_{conceptual} \approx \text{von Neumann}_{conceptual}$ (conscious rule application)
d. $\text{connectionist}_{conceptual} \sim \text{von Neumann}_{conceptual}$ (intuition)

Here, the symbol "$\approx$" means "equals to a good approximation" and "$\sim$" means "equals to a crude approximation." The two nearly equal virtual machines in (2c) both describe what I will call the *conscious rule interpreter*. The two roughly similar virtual machines in (2d) provide the two paradigms' descriptions of the *intuitive processor* at the conceptual level.

Table 1 indicates these relationships and also the degree of exactness to which each system can be described at each level—the degree of precision to which each virtual machine is defined. The levels included in Table 1 are those relevant to predicting high-level behavior. Of course each system can also be described at lower levels, all the way down to elementary particles. However, levels below an exactly describable level are ignorable from the point of view of predicting high-level behavior, since it is possible (in principle) to do the prediction at the highest level that can be exactly described (and it is presumably much harder to do the same at lower levels). This is why in the symbolic paradigm any descriptions below the symbolic level are not viewed as significant. For modeling high-level behavior, how the symbol manipulation happens to be implemented can be ignored—it is not a relevant part of the cognitive model. In a subsymbolic model, exact behavioral prediction must be performed at the subconceptual level—but how the units happen to be implemented is not relevant.

Table 1: Three cognitive systems and three levels of description

level *(process)*	cognitive system		
	brain	**subsymbolic**	**symbolic**
conceptual *(intuition)*	?	rough approximation ~	exact
conceptual *(conscious rule application)*	?	good approximation ≈	exact
subconceptual	good approximation ≈	exact	
neural	exact		

The relation between the conceptual level and lower levels is fundamentally different in the subsymbolic and symbolic paradigms. This leads to important differences in the kind of explanations the paradigms offer of conceptual-level behavior, and the kind of reduction used in these explanations. A symbolic model is a *system* of interacting processes, all with the same conceptual-level semantics as the task behavior being explained. Adopting the terminology of Haugeland (1978), this *systematic explanation* relies on a *systematic reduction* of the behavior that involves no shift of semantic domain or *dimension*. Thus a game-playing program is composed of subprograms that generate possible moves, evaluate them, and so on. In the symbolic paradigm, these systematic reductions play the major role in explanation. The lowest-level processes in the systematic reduction, still with the original semantics of the task domain, are then themselves reduced by *intentional instantiation:* they are implemented exactly by other processes with different semantics but the same form. Thus a move-generation subprogram with game semantics is instantiated in a system of programs with list-manipulating semantics. This intentional instantiation typically plays a minor role in the overall explanation, if indeed it is regarded as a cognitively relevant part of the model at all.

Thus cognitive explanations in the symbolic paradigm rely primarily on reductions involving no dimension shift. This feature is not shared by the subsymbolic paradigm, where accurate explanations of intuitive behavior require descending to the subconceptual level. The elements in this explanation, the units, do *not* have the semantics of the original behavior. Thus unlike symbolic explanations, subsymbolic explanations rely crucially on a semantic ("dimension") shift that accompanies the shift from the conceptual to the subconceptual levels.

The overall dispositions of cognitive systems are explained in the subsymbolic paradigm as approximate higher level regularities that emerge from quantitative laws operating at a more fundamental level with different semantics. This is the kind of reduction familiar in natural science, exemplified by the explanation of the laws of thermodynamics through a reduction to mechanics that involves shifting dimension from thermal semantics to molecular semantics. (Section discusses some explicit subsymbolic reductions of symbolic explanatory constructs.)

Indeed the subsymbolic paradigm repeals the other features that Haugeland identified as newly introduced into scientific explanation by the symbolic paradigm. The inputs and outputs of the system are not "quasilinguistic representations" but good old-fashioned numerical vectors. These inputs and outputs have semantic interpretations, but these are not constructed recursively from interpretations of imbedded constituents. And the fundamental laws are good old-fashioned numerical equations.

Haugeland went to considerable effort to legitimize the form of explanation and reduction used in the symbolic paradigm. The explanations and reductions of the subsymbolic paradigm, by contrast, are of a type well-established in natural science.

In summary, let me emphasize that in the subsymbolic paradigm the conceptual and subconceptual levels are not related as the levels of a von Neumann computer (high-level-language program, compiled low-level program, etc.). The relationship between subsymbolic and symbolic models is more like that between quantum and classical mechanics. Subsymbolic models accurately describe the microstructure of cognition, while symbolic models provide an approximate description of the macrostructure. An important job of subsymbolic theory is to delineate

the situations and respects in which the symbolic approximation is valid, and to explain why.

Conscious rule application in the subsymbolic paradigm

In the symbolic paradigm, both conscious rule application and intuition are described at the conceptual level: as conscious and unconscious rule interpretation, respectively. In the subsymbolic paradigm, conscious rule application can be formalized at the conceptual level but intuition must be formalized at the subconceptual level. This suggests that a subsymbolic model of a cognitive process involving both intuition and conscious rule interpretation would consist of two components employing quite different formalisms. While this hybrid formalism might have considerable practical value, there are some theoretical problems with it. How would the two formalisms communicate? How would the hybrid system evolve with experience, reflecting the development of intuition and the subsequent remission of conscious rule application? How would the hybrid system elucidate the fallibility of actual human rule application (eg. logic)? How would the hybrid system get us closer to understanding how conscious rule application is achieved neurally?

All these problems can be addressed by adopting a unified subconceptual-level analysis of both intuition and conscious rule interpretation. The virtual machine that is the conscious rule interpreter is to be implemented in a lower-level virtual machine: the same connectionist system that models the intuitive processor. How this can, in principle, be achieved is the subject of this section. The relative advantages and disadvantages of implementing the rule interpreter in a connectionist system rather than a von Neumann machine will also be considered.

The observation is this.

> The competence to represent and process linguistic structures in a native language is a competence of the human intuitive processor, so the subsymbolic paradigm assumes that this competence can be modeled in a subconceptual connectionist system. By combining such linguistic competence with existing memory capabilities of connectionist systems, sequential rule interpretation can be implemented.

Assuming that sentences of natural language can be represented in a subconceptual connectionist system means that such sentences correspond to certain patterns of activity. Assuming that sentences can be processed means in particular that a pattern of activity representing a verbal instruction can be used to carry out that instruction. Once sentences are represented as patterns of activity, the well-known procedures of associative memories can be used to store them. These are content-addressable memories in which reinstantiation of a part of the stored item causes reinstantiation of the complete item. A collection of such memories can then be used to drive sequential behavior, as follows.

First, a set of linguistically expressed rules is presented to the connectionist system and thereby stored. For concreteness we can imagine the rules to be productions: "if *condition* holds, then do *action*." In a particular situation when the condition of a rule holds, the pattern of activity representing the condition will be instantiated in the network. This will cause the entire pattern of activity representing that rule to be reinstantiated by the memory retrieval mechanism. Now it is as if the rule had been linguistically presented to the system from an external instructor. The language processing mechanism can interpret the sentence, generating the appropriate action and leading to a new pattern of activity in the network representing the new situation. This new pattern leads to the reinstantiation of another stored rule, and the cycle repeats.

Using the stored rules the network can perform the task. The standard learning procedures of connectionist models turn this experience performing the task into a set of weights for going from inputs to outputs. Eventually, after enough experience, the task can be performed directly by these weights. The input activity generates the output activity so quickly that before the relatively slow interpretation process has a chance to reinstantiate the first rule and carry it out, the task is done. With intermediate amounts of experience, some of the weights are well enough in place to prevent some of the rules from having the chance to instantiate, while others are not, enabling other rules to be retrieved and interpreted.

Rule interpretation, consciousness, and seriality

What about the conscious aspect of rule interpretation? Since consciousness seems to be a quite high-level description of mental activity, it is reasonable to suspect that it reflects the very coarse structure of the cognitive system. Considering coarseness on the time dimension, we are lead to hypothesize:

> Patterns of activity that are stable for relatively long periods of time (on the order of 100 msec) determine the contents of consciousness.

(See Rumelhart, Smolensky, McClelland & Hinton 1986.) The rule interpretation process requires the maintenance of the retrieved linguistically coded rule while it is being carried out. Thus the pattern of activity representing the rule is stable for a relatively long time. By contrast, after connections have been developed to perform the task directly, there is no correspondingly stable pattern formed during the performance of the task. Thus the loss of conscious phenomenology with expertise can be understood naturally.

On this account, the sequentiality of the rule interpretation process is not built into the architecture; rather it is a consequence of the fact that we can follow only one instruction at a time. Connectionist memories have the capability to retrieve a single stored item, and here this is necessary to avoid asking the linguistic interpreter to simultaneously interpret more than one instruction.

It is interesting to note that the preceding analysis does not require that the "rules" be linguistic; any notational system that can be appropriately interpreted would do. Another type of "rule" is a series of musical pitches; a memorized collection of such rules allows a musician to play a tune by "conscious rule interpretation." With practice the need for conscious control goes away. Since pianists learn to interpret several notes simultaneously, the present account suggests that pianists might be able to apply more than one musical rule at a time (provided their memory for the rules can simultaneously recall more than one rule). A symbolic account of such conscious rule interpretation would involve something like a production system capable of firing multiple productions simultaneously.

Finally it should be noted that even if the memorized rules are assumed to be linguistically coded, the preceding analysis is uncommitted about the form the rules take in memory: phonological, orthographic, semantic, or whatever.

Symbolic vs. subsymbolic implementation of rule interpretation

The (approximate) implementation of the conscious rule interpreter in a subsymbolic system—a reduction of the traditional kind, as I have argued above—has both advantages and disadvantages relative to an (exact) implementation in a von Neumann machine—a reduction by intentional instantiation.

The main disadvantage is that subconceptual representation and interpretation of linguistic instructions is very difficult and we can't now actually do it. Most existing subsymbolic systems simply don't use rule interpretation.[1] They can't take advantage of rules to check the results produced by the intuitive processor. They can't bootstrap their way into a new domain using rules to generate their own experience: they must have a teacher generate it for them.[2]

There are several advantages of a subconceptually implemented rule interpreter. The intuitive processor and rule interpreter are highly integrated, with broad-band communication between them. Understanding how this communication works should allow design of efficient hybrid symbolic/subsymbolic systems with effective

---

1. A notable exception is Touretzky & Hinton 1985.

2. And when a network makes a mistake, it can be told the correct answer but it can't be told which rule it violated. Thus it must assign blame for its error in a very undirected way. It is quite plausible that the large amount of training currently required by subsymbolic systems could be significantly reduced if blame could be focussed by citing violated rules.

communication between the processors. A principled basis is provided for studying how rule-based knowledge leads to intuitive knowledge. Perhaps most interesting, in a subsymbolic rule interpreter, the process of rule selection is intuitive! Which rule is reinstantiated in memory at a given time is the result of the associative retrieval process, which has many nice properties. The "best match" to the productions' conditions is quickly computed, and even if no match is very good, a rule can be retrieved. The selection process can be quite context-sensitive.

An integrated subsymbolic rule interpreter/intuitive processor in principle offers the advantages of both kinds of processing. Imagine such a system creating a mathematical proof. The intuitive processor would suggest goals and steps, and the rule interpreter would verify the validity of proposals. The serial search through the space of possible steps that is necessary in a purely symbolic approach is replaced by intuitive generation of possibilities. Yet the precise adherence to strict inference rules that is demanded by the task can be enforced by the rule interpreter; the creativity of intuition can be exploited while its unreliability can be controlled.

Two kinds of knowledge; one medium

Most existing subsymbolic systems perform tasks without serial rule interpretation: patterns of activity representing inputs are directly transformed (possibly through multiple layers of units) to patterns of activity representing outputs. The connections that mediate this transformation represent a form of task knowledge that can be applied with massive parallelism: I will call it ***P-knowledge***. For example, the P-knowledge in a native speaker encodes lexical, morphological, syntactic, semantic, and pragmatic constraints in such form that all these constraints can be satisfied in parallel during comprehension and generation.

The connectionist implementation of sequential rule interpretation described above displays a second form that knowledge can take in a subsymbolic system. The stored activity patterns that represent rules also constitute task knowledge: call it ***S-knowledge***. Like P-knowledge, S-knowledge is imbedded in connections: the connections that enable part of a rule to reinstantiate the entire rule. Unlike P-knowledge, S-knowledge cannot be used massively in parallel. For example, a novice speaker of some language cannot satisfy the constraints contained in two memorized rules simultaneously; they must be serially reinstantiated as patterns of activity and separately interpreted. Of course the connections responsible for reinstantiating these memories operate in parallel, and indeed these connections contain within them the potential to reinstantiate either of the two memorized rules. But these connections are so arranged that *only one rule at a time* can be reinstantiated. The retrieval of a single rule is a parallel process, but the satisfaction of the constraints contained in the two rules is a serial process. After considerable experience, P-knowledge is created: connections that can *simultaneously satisfy* the constraints represented by the two rules.

P-knowledge is considerably more difficult to create than S-knowledge. To encode a constraint in connections so that it can be satisfied in parallel with thousands of others is no easy task. Such an encoding can only be learned through considerable experience in which that constraint has appeared in many different contexts, so that the connections enforcing the constraint can be tuned to operate in parallel with those enforcing a wide variety of other constraints. S-knowledge can be much more rapidly acquired (once the linguistic skills on which it depends have been encoded into P-knowledge, of course). Simply reciting a verbal rule over and over will usually suffice to store it in memory (at least temporarily).

That P-knowledge is so highly context-dependent while the rules of S-knowledge are essentially context-free is an important computational fact underlying many of the psychological explanations offered by subsymbolic models. Consider, for example, Rumelhart and McClelland's (1986) model of the U-shaped curve for past-tense production in children. The phenomenon is striking: a child is observed using *goed* and *wented* when at a much younger age *went* was reliably used. This is surprising because we are prone to think that such linguistic abilities rest on knowledge that is encoded in some context-free form such as "the past tense of *go* is *went*." Why should a child *lose* such a rule once acquired? A traditional answer invokes the acquisition of a different context-free rule, like "the past tense of *x* is *x+ed*" which, for one reason or another, takes precedence. The point here, however, is that *there is nothing at all surprising about the phenomenon when the underlying knowledge is assumed to be context-dependent and not context-free.* The young child has a small vocabulary of largely irregular verbs. The connections that implement this P-knowledge are capable of reliably producing the large pattern of activity representing *went*, as well

as those representing a small number of other past-tense forms. Informally we can say that the connections producing *went* do so *in the context of the other vocabulary items* that are also stored in the same connections. There is no guarantee that these connections will produce *went* in the context of a different vocabulary. As the child acquires additional vocabulary items, most of which are regular, the context radically changes. Connections that were, so to speak, perfectly adequate for creating *went* in the old context now have to work in a context where very strong connections are trying to create forms ending in *–ed*; these "old connections" are not up to the new task. Only through extensive experience trying to produce *went* in the new context of many regular verbs can the "old" connections be modified to work in the new context. (In particular, strong new connections must be added that, when the input pattern is that for *go*, cancel the *–ed*; these were not needed before.)

These observations about context-dependence can also be framed in terms of inference. If we choose to regard the child as using knowledge to in some sense "infer" the correct answer *went*, then we can say that after the child has added more knowledge (about new verbs), the ability to make the (correct) inference is lost. In this sense the child's inference process is *non-monotonic*—perhaps this is why we find the phenomenon surprising. Non-monotonicity is a fundamental property of subsymbolic inference (see Smolensky, 1987b).

To summarize:

- Knowledge in subsymbolic systems can take two forms, both resident in the connections.
- The knowledge used by the conscious rule interpreter lies in connections that reinstantiate patterns encoding rules; task constraints are coded in context-free rules and satisfied serially.
- The knowledge used in intuitive processing lies in connections that constitute highly context-sensitive encodings of task constraints that can be satisfied with massive parallelism.
- Learning such encodings requires much experience.

Conclusion

The approach described above sets out a rather clear program for developing subsymbolic models of conscious rule application. The crucial technical problems to be solved involve subsymbolic models of natural language processing: the representation of linguistic structures—including procedural descriptions such as productions—and their interpretation—effecting the designated procedures as a result of activating their representation. These are hard problems, but they are problems that need to be solved anyway by a connectionist approach to language processing. With even very limited solutions to these problems, we can begin to seriously explore the interaction of rule application and intuitive processing, in both learning and performance, within a subsymbolic connectionist framework.

Acknowledgements

I am indebted to a number of people for very helpful conversations on these issues: Jerry Fodor, Zenon Pylyshyn, and especially Dave Rumelhart, Rob Cummins, and Denise Dellarosa. This research has been supported by NSF grant IST-8609599 and by the Department of Computer Science and Institute of Cognitive Science at the University of Colorado at Boulder. This paper is based on a small portion of a paper to appear in *The Behavioral and Brain Sciences*.

References

Anderson, J.A. & Hinton, G.E. (1981). Models of information processing in the brain. In G. E. Hinton and J. A. Anderson, Eds., *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1: 215–226.

Hinton, G.E., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. Cambridge, MA: MIT Press/Bradford Books.

Rumelhart, D.E. & McClelland, J.L. (1986). On learning the past tenses of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*, J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. Cambridge, MA: MIT Press/Bradford Books.

Rumelhart, D.E., Smolensky, P., McClelland, J.L., and Hinton, G.E. (1986). Schemata and sequential thought processes in parallel distributed processing models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*, J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. Cambridge, MA: MIT Press/Bradford Books.

Smolensky, P. (1986a). Information processing in dynamical systems: Foundations of harmony theory. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. Cambridge, MA: MIT Press/Bradford Books.

Smolensky, P. (1986b). Neural and conceptual interpretations of parallel distributed processing models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*, J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. Cambridge, MA: MIT Press/Bradford Books.

Smolensky, P. (1987a). Connectionist AI, symbolic AI, and the brain. *AI Review*, special issue on the foundations of AI.

Smolensky, P. (1987b). On the proper treatment of connectionism. Technical Report CU-CS-359-87, Department of Computer Science, University of Colorado at Boulder.

Touretzky, D.S. & Hinton, G.E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the International Joint Conference on Artificial Intelligence*.