

# Do we need symbols for designing complex AI systems?

Achim Hoffmann

School of Computer Science and Engineering

University of New South Wales, Sydney 2052 NSW, Australia

Email: achim@cse.unsw.edu.au

## Abstract

While the symbolic paradigm of AI takes it for granted that symbols are sufficient for the successful development of AI systems, the connectionist paradigm denies the appropriateness of symbols for both, the design of AI systems as well as for modelling cognition. This paper will discuss the different views in regards to the fundamental notion of symbols in the two opposing paradigms. The strong connectionist view claims that even in human cognition there are no symbols of a fixed meaning, hence the idea to build intelligent systems as physical symbol systems is misleading and must fail, when a comprehensive intelligence should be created. This paper acknowledges the connectionist criticism of the role of symbols in cognition. Despite that criticism, the paper argues that symbols are nevertheless essential for any approach of successfully building complex AI systems.

## 1 Introduction

The early years of AI began with two different approaches. Some researchers investigated models of neurons starting as early as 1943 [7] and developed mechanisms on how these simple neural models could show adaptive behaviour. On the other hand, there were mathematically and psychologically minded researchers who set out for the development of algorithms which would correspond rather to the thought processes of man than to the neural processes in human brains. The most influential early work of this group was probably the framing of *physical symbol systems* as a paradigm for developing general intelligent agents by Newell and Simon [13, 9, 10]. Both approaches were competing until in 1969 Minsky and Papert published a book which showed that, although the early neural learning mechanisms could learn anything a neuron could represent, a neuron could represent only very little. After its publication, funding for neural network re-

search was largely cut by government agencies, resulting in a virtual shutdown of all neural network research.

Since then, almost all research in AI was conducted in the *symbolic paradigm* of Newell and Simon's physical symbol systems. In the following years, however, most AI researchers reduced their aspirations: no longer was a general intelligent agent sought. People were rather content with building application specific intelligent systems. The era of expert systems began. It took until the mid-1980s, before neural networks, under the title *connectionism* were rediscovered as a promising alternative to the symbolic paradigm. The reasons for its re-emergence can be attributed to a variety of developments in physics, in psychology, in AI itself,<sup>1</sup> as well as to the increasing acceptance of Dreyfus philosophical critique at the enterprise of AI. Dreyfus criticised AI since the middle of the 1960s, as an ill-conceived endeavour [3].

After neural networks returned to the stage around 1986, a heated debate began arguing that the symbolic paradigm of AI is to be replaced by the newly emerged *paradigm of connectionism*, the idea of massively parallel and highly connected computer systems of a brain-like architecture. In the following years, connectionism became increasingly popular as an alternative approach to the classical symbolic approach to artificial intelligence. In fact, it was claimed that general intelligent agents have to be built as connectionist systems. On the other hand, even within the traditional symbolic paradigm, Nilsson announced the aspiration to build again general intelligent systems in 1995 [11].

The question still remains open, whether we need symbols in order to develop complex AI systems, even if these systems are not general intelligent systems. they encompass the scope of a general intelligence. This paper revisits the basic ideas of the two paradigms and investigates the notion of *symbols* in

---

<sup>1</sup>The achievements were rather disappointing and far behind the earlier predictions.

more detail. While it acknowledges the criticism at the symbolic approach, it is argued in this paper that, for a different reason, symbols are nevertheless necessary for a successful approach to build complex intelligent systems.

The paper is organised as follows: Section 2 briefly reviews the basic ideas of the classical symbolic paradigm of AI. Section 3 summarises the critiques of the symbolic approach to AI. In the following section 4, the central ideas of connectionism are presented. Section 5 discusses the consequences of the foregoing for the design of complex intelligent systems.

## 2 The Symbolic Level

According to Newell and Simon, a physical symbol system is built from a set of elements, called symbols, which may be formed into symbol structures by means of a set of relations.

A symbol system has a memory capable of storing, retaining and retrieving symbols and symbol structures. It has a set of information processes that form symbol structures as a function of sensory stimuli. And it has a set of information processes which produce symbol structures that cause motor actions and modify symbol structures in memory in a variety of ways. A symbol system interacts with its environment in two different ways:

- It receives sensory stimuli from the environment which it converts into internal symbol structures.
- It acts upon the environment in ways determined by symbol structures that it produces by internal information processes.

Thus, its behaviour can be influenced by both: its current environment through its sensory inputs, and by previous environments through the information it has stored in memory from its experiences.

In general, both symbols and symbol structures are called ‘symbols’. Here, symbols are *signs* or tokens which have a reference. Although it is not quite clear as to what kind of domain the symbols may refer, the idea is roughly that symbols correspond to the content of conscious thoughts. For example, symbols may refer to physical objects, like ‘this house’, ‘the river over there’, etc. Certainly, symbols may also refer to concepts which have no physical manifestation, like the general concept ‘house’, which refers rather to a *class* of objects (existing and non-existing in the physical world) than to a

concrete physical object or group of objects. Further, symbols may even refer to completely abstract concepts, like mathematical concepts, e.g. ‘number three’ etc.

Symbols are thought as patterns which may be physically implemented in various ways. In today’s computers they are usually patterns of electromagnetism, though their physical nature may differ radically in integrated circuits or vacuum tubes in the computers of the 1940s. Although it is unknown how patterns are represented in the human mind, it is generally assumed that they are represented as neuronal arrangements of some kind.

Newell and Simon’s notion of symbol systems is extremely general, so general that symbol systems appear to be equivalent to the notion of the Turing machine, i.e. equivalent to the notion of computation. Newell formulated the so-called *physical symbol system hypothesis* claiming that physical symbol systems may include symbols which correspond to those concepts, humans normally use. ... *It becomes a hypothesis that this notion of symbols includes symbols that we humans use everyday in our lives.*<sup>2</sup>

The symbol level describes cognition by using symbols as the elementary units and by using rules to manipulate symbols for describing the behaviour of the system. The application of a rule to symbols depends entirely on symbols which are currently stored in the system.

A particularly strong version of a symbolic level of description allows to compose symbols to new symbol structures where the meaning, i.e. the reference, of the composed symbol structures can be derived from the specific syntactical rules being used for composing the symbol structure. This requires a strong connection between the rules for composing symbol structures and the domain of reference.

### Symbols in the Compositional Sense

The idea of symbols in the compositional sense has essentially been developed by modern logic, when the need for formal semantics of a logical language was recognised. The essential idea is that there is a set of primitive symbols with a fixed meaning, i.e. a fixed reference to certain objects. For a very simple version, see Figure 1. From such elementary symbols it is possible to build more complex symbol structures according to a given set of syntactical rules. Compositionality means that the meaning of a composed complex symbol structure can be derived by following semantic rules which correspond one-to-one to the syntactic rules which have

---

<sup>2</sup>In [8].

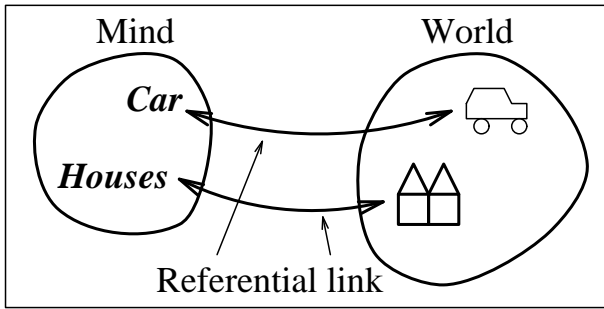


Figure 1: The basic idea of representation. Symbols in the mind represent objects in the world.

been used to build the compound symbol structure. The typical example for symbol systems of this kind are logical languages with model-theoretic semantics. The standard model-theoretic semantics are given by an interpretation function of the atomic syntactical symbols of a given language  $L$  into a domain of individuals  $D$ . Further, there may be functions among individuals which are interpreted as functions among the domain  $D$ . Finally, a predicate  $P$  is interpreted in an extensional way by the set of individuals to which the predicate  $P$  applies (in the case of  $P$  being a single-argument predicate). For multi-argument predicates the extension is given by the set of tuples of individuals to which the predicate applies. Based on the interpretations of these elementary syntactical units, the interpretation of complex syntactical expressions can be constructed from the interpretations of the involved elementary units.

### 3 Critiques of Symbolism

Arguments which indicate that human intelligence does not essentially rest on symbolic representations of an external world, have been put against the physical symbol system hypothesis by e.g. Dreyfus [3, 4] or Winograd & Flores [16]. Based on Heidegger's phenomenology [5], it was argued that human cognition and understanding is basically connected with human practice and social interaction. We engage in certain activities and reflect only on our knowledge as far as it is necessary for the task at hand. For instance, if we are in the situation to justify our decisions, we will only think about those reasons which we can use to convince other humans. We will not reflect about those parts which we deem to be common ground between ourselves and the other parties of the discussion. Heidegger's phenomenology argues that human values and goals, and similarly on-

tological considerations, are *implicitly* contained in the way people act within their environment as well as in the way people look at the world, i.e. which objects exist for them in which situation and which properties the objects have. In contrast to that, it was argued, computers *have* to represent 'values' and 'goals' as well as their 'view' of the world *explicitly* by symbols and therefore computers cannot behave like humans. Another characteristic of the human mind is that it only starts reflecting about objects and their relations in an outer world in situations of *breakdowns*. That is, in situations where the usual course of action cannot proceed as expected. For instance, only when I cannot open the door in the way I am used to, I become aware of the door in its physical appearance. If the front door to my office building is unexpectedly locked, only then I become aware of the door and its locking mechanism. Only then, I may remember where to find a key for the door and how to operate its locking mechanism in order to open the door.

It is interesting to note that the general applicability of the model-theoretic approach has been rejected not only by Heidegger's phenomenology, but also by the late Wittgenstein [17] himself; the philosopher who was once a strong advocate of the model-theoretic idea [18]. Reason being the fact that there is knowledge beyond what can be expressed in model-theoretic terms. It is not enough to say all what is the case and all what is not the case. A static world description fails to give an account on how the world will change. It also fails to give an account on how a human being's knowledge or view of the world will change. The particular way, humans conduct non-deductive inferences cannot be described in terms referring to objects in the world or concepts of those objects. Hence, to describe such ways of reasoning, symbols are needed which do not refer merely to the physical world. Those symbols have to refer to abstract structures, like preference relations among a class of possible inferences which are logically equally justified.

Not only a general intelligence needs to deal with descriptions beyond the model-theoretic approach. Also, many important practical AI applications require non-deductive reasoning. At least some non-monotonic, inductive or analogical reasoning or reasoning under uncertainty and vagueness is required. For example in inductive reasoning, there is usually an ambiguity among different hypotheses, when generalising inferences should be performed. Thus, there is the need for expressing some *preference relation* among competing hypotheses. However, any preference relation among a finite number of hy-

potheses can be represented by symbols, although these symbols do not refer to some ‘natural’ ontological entities, i.e. such symbols do not refer to the ontological entities of the domain of inductive reasoning. The same problem exists for reasoning under uncertainty: human reasoning incorporates a vast number of interdependencies of subjective probabilities, when probability estimates of combined uncertain events are made.

## 4 Connectionism

Connectionist models are usually considered as large networks of simple computing units which act in parallel. Each computing unit carries a numerical *activation value* which is computed from the activation values of connected units in the network. The network’s elements influence each other’s values by connections of specific strengths. The strength of a connection may change over time. The change of a connection may depend on particular *training patterns* that are to be reflected in some sense by the overall network. The network’s overall function is strongly dependent on the currently present connection strengths or *weights*. Hence, the weights of a network are also encoding the system’s knowledge. Many of the connectionist models *program themselves*, i.e. the models have learning procedures for tuning their weights to implement a specific I/O-function for the overall network. At the heart of the following exposition of the connectionist claim lies the idea of *subsymbols*. According to Smolensky [14] are subsymbols, roughly speaking, numerical values which cannot be interpreted in any sensible way individually. Only a large set of such subsymbols and only if they show certain patterns can be said to correspond to symbols. In other words, only certain patterns among a large number of subsymbolic values can be assigned any meaning. Hence, subsymbols differ in their interpretability substantially from the symbols of the symbolic paradigm, although numerical values may also appear in the symbolic paradigm.

### 4.1 The Connectionist Claim

The advocates of connectionism typically claim that a connectionist approach is *necessary* for an appropriate modelling of cognition and intelligent activities.<sup>3</sup>

<sup>3</sup>See e.g. [2] for a discussion of why a connectionist-style framework is needed for cognitive science.

To prove this seems a difficult task because of the fact that virtual machines can usually be implemented on another virtual machine. That is, it has to be shown that the subsymbolic level *cannot be reduced* to the symbolic level of description. One might say that a digital computer is some sort of dynamical system which simulates a von Neumann automaton. Furthermore that digital computers are used for simulating connectionist models. So, it seems plausible that both, the *symbolic* as well as the *subsymbolic* paradigm are correct - that they are two sides of the same coin.

Smolensky points out, that subsymbolic models are not equivalent to symbolic models, even if subsymbolic models are implemented on digital computers using a symbolic programming language. The crucial difference lies in the fact, that the symbols used to implement the subsymbolic models are merely some numbers used in the highly parallel and dynamical connectionist model. But the semantics of these numbers are very different from the semantics of the symbols in the symbolic paradigm. In the symbolic paradigm, the symbols refer to certain concepts which at least are known to the individual programmer, if not even known to a community of system developers or scientists. The known fact that von Neumann machines and certain connectionist models can simulate each other has *no* impact on Smolensky’s point. The crucial issue is not the possible reduction from one level to the other. The crucial issue is rather whether symbols can be used that refer to conceptual entities (as it is assumed in the symbolic paradigm) or whether the employed symbols cannot refer to anything more enlightening than a set of numbers of incomprehensible significance, which essentially makes up the subsymbolic paradigm?

The very debate on an appropriate paradigm for cognitive science and AI cannot be discussed on a purely syntactical level. The debate rather takes place at a semantical level. Essentially the debate addresses the following question:

*What are the conceptual entities to which those syntactical entities refer, which are manipulated according to simple syntactical rules.*

### 4.2 The Connectionist Research Program

As a consequence, substantial progress in subsymbolic cognitive science requires systematic commitments to vectorial representations for individual cognitive domains.

Smolensky believes that: ... *powerful mathemati-*

cal tools are needed for relating the overall behaviour of the network to the choice of representational vectors; ideally, these tools should allow us to invert the mapping from representations to behaviour so that by starting with a mass of data on human performance we can turn a mathematical crank and have representational vectors pop out. ... The subsymbolic paradigm needs tools such as a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments.<sup>4</sup>

Further he states that: ... systematic principles must be developed for adapting to the connectionist context, the featural analyses of domains that have emerged from traditional, non-connectionist paradigms. These principles must reflect fundamental properties of connectionist computation, for otherwise, the hypothesis of connectionist computation is doing no work in the study of mental representation.<sup>5</sup>

Along with this semantic distinction comes a syntactic distinction. Subsymbols are not operated upon by symbol manipulation: they participate in numerical, as opposed to symbolic, computation. Operations in the symbolic paradigm that consist of a single discrete operation (e.g. a memory fetch) are often achieved in the subsymbolic paradigm as the result of a large number of much finer-grained (numerical) operations.

The knowledge in a connectionist architecture is encoded in patterns of its connection strengths. Processing of that knowledge happens highly parallel.

Similar considerations from a more technical point of view can, for example, be found in an article by Werbos [15], who discovered the backpropagation algorithm in 1974 (although in a different context than neural networks - in the latter it was presented in by Rumelhart et al. in 1986 [12]):

... that new developments in neurocontrol would permit a Newtonian revolution in our understanding of the human brain and the human mind, if we pursue these developments in the right way. ... to understand intelligence as it exists in the human brain, in the same way that we understand physics (since Newton), as a real science. If the brain itself - as a whole system - is a neurocontroller, then the mathematics which we need to use is neurocontrol.

People trained in conventional control theory are often skeptical of this idea. They are used to thinking of controllers as systems like glorified thermostats, designed to stabilise a chemical plant at some fixed equilibrium, or to make a robot arm follow a fixed path in space specified in advance, or the like.

<sup>4</sup>Smolensky [14]:8.

<sup>5</sup>Ibid.

In summary, connectionism claims that by the massive parallel and distributed processing of complex activity patterns in a large network, a system can exhibit emergent behaviour which models human cognition more appropriately than a symbolic approach to AI is capable of.

## 5 Discussion and Conclusions

Connectionism represents an approach for building intelligent systems and understanding cognition on the basis of a metaphorical resemblance to the human brain. Connectionism at the subsymbolic level does not attempt to reflect the neuro-physiological details of the human brain, but rather focuses on a particular aspect of what is suspected or even known about the biological functioning of the human brain: some massively distributed processing is taking place, such that the overall behaviour of the system depends on the entirety of the neural network. Furthermore, the elementary processing units are working on rather continuous operands than discrete ones.

While these features of the biological function of the human brain are very appealing and reflect a number of aspects which have been found in cognition, the essential ideas of connectionism are merely small parts of a system description, where a very essential part is missing:<sup>6</sup> namely, *what* is the particular knowledge that has to be encoded in subsymbolic vector representations in order to make the overall system work.

While the symbolic approach attempts to describe cognition at the level of symbols which refer to human concepts, there have been a number of criticisms put forward, arguing that describing cognition at the symbolic level is bound to fail as there are no simple rules operating directly on symbols which could produce a simulation of (at least introspectively) observable cognitive behaviour. The subconscious process, which causes the sudden emergence of a conscious thought, cannot be described in terms of consciously accessible concepts.

Opposed to that, connectionism attempts to describe cognition by rules which do not operate directly on symbols, but rather on syntactical units (also called *subsymbols*) which are only constituents of symbols referring to conceptual entities. Only certain patterns among a large number of such subsymbols constitute symbols, which in turn represent conceptual entities.

<sup>6</sup>One can even argue that *the* essential part is missing.

The crucial point in the discussion is a *semantical* one, not a syntactical one. That is, while symbol systems, such as Turing machines, are computational universal, this does not imply that they really suffice for describing cognitive processes. It is not what paradigm allows to describe the functionality of a system. It is rather, what are the semantics of the units involved in the description of that functionality?

What are the consequences of that point for the controversy between symbolism and connectionism? What are the consequences for the role of symbols in designing complex AI systems, which may be restricted in their scope of intelligent behaviour they exhibit?

To discuss these questions further, a distinction as proposed in Clancey [1] is suitable: *First-person symbols* are those symbols which are used by a person in their conscious thought processes. Opposed to that are *third-person symbols* used by an outside observer who describes the cognitive, neuronal or perhaps subsymbolic processes, which may be going on in the mind of the first person. Those third-person symbols have a much wider domain of concepts to which they may refer. Essentially, every conceptual structure that is useful in explaining or designing an intelligent system can be referred to by third-person symbols.

Engineers as well as scientists are communicating through symbols; i.e. a design needs to be written down and need to be understood by other engineers, who participate in the project. Similarly, scientists attempt to formalise their knowledge, i.e. they try to write down a theory which should unambiguously explain the phenomena in question.

In other words, the *human subject*, being the engineer or the scientist who tries to understand intelligence and cognition, requires that symbols are used for describing intelligent systems. It is not the nature of intelligence or cognition itself, which dictates the paradigm. It is rather the nature and the resulting needs of those who want to understand intelligence or cognition within a certain paradigm! Hence, what is required in descriptions, are *third-person symbols*. This implies that the *first-person symbols*, i.e. those symbols whose role in cognition is to be explained, have to be explained by using any sort of symbols - first-person or third-person symbols - as long as the description is understandable and allows an engineer or a scientist to do their job.

The claim of connectionism was essentially that there are simple manipulation rules only at the subsymbolic level, which govern the cognitive processes

at the symbolic level.<sup>7</sup> Complex patterns of subsymbols constitute symbols in a dynamical way, where the mapping of patterns of subsymbols to the symbols being constituted is everything else than straightforward. The 'knowledge' of such a subsymbolic system is encoded in a distributed way in those numerical vectors, each number representing a subsymbol.

As a consequence, the possibility of an understanding of the knowledge which is encoded in the system is largely abandoned. Instead, it is hoped that 'mathematical cranks' can be developed, which allow a mapping from (desired) behaviour of the overall system to representational vectors at the subsymbolic level. That is, learning mechanisms are required which automatically encode the knowledge of the system, as it would be impossible for an engineer to do it 'manually'. This knowledge can only be derived from samples of desired behaviour, as the connectionist paradigm presumes that it is impossible to express the required knowledge in symbols.

Unfortunately, as proven e.g. in [6], such learning approaches are inherently limited: Complex functions can only be learned from a vast number of training examples. Such a number of training examples is in practical scenarios not available. These theoretical results have been confirmed by the practical experiences with various types of learning approaches in both, the symbolic as well as in the connectionist paradigm. In order to avoid the need for excessive training data, the learning system has to have part of the knowledge beforehand. Hence, a successful learning system would require substantial structure built into it a priori, in order to learn effectively any complex functions. Encoding this a priori knowledge faces the same practical problems as discussed above; i.e. owing to its claimed 'unrelatedness' to symbols, there is no way for a system engineer to encode such knowledge other than by some sort of blind trial-and-error.

While the arguments used by connectionist proponents *against* the symbolic paradigm seem largely valid, they are *not at all* conclusive arguments *for* the connectionist paradigm.

---

<sup>7</sup>The existence of the Universal Turing machine, whose rules of operation are rather simple, shows that indeed such simple rules may exist. However, the program which is executed by the simple Universal Turing machine has to contain the required complexity, i.e. the 'knowledge' which is represented by a large collection of subsymbols has to make up for the simplicity of the manipulation rules of those subsymbols. Generally speaking, there is always the trade-off between the complexity in the program and the complexity in the data being processed. The existence of the Universal Turing machine shows that both are interchangeable.

In any case, the design of a complex system has to be done to a large extent by human system developers; even if the system has some learning capabilities. As a consequence, the design has to take place in way that is comprehensible for those human system developers. Hence, symbols are inevitable for the successful design of complex AI systems. Those symbols, however, may largely be third-person symbols.

## References

- [1] W. Clancey. Situated action: A neuropsychological interpretation. *Cognitive Science*, 17:87–116, 1993.
- [2] A. Clark. *Associative Engines*. MIT Press/Bradford, Cambridge, MA, 1993.
- [3] H. L. Dreyfus. Alchemy and artificial intelligence. Technical Report P3244, RAND Corporation, Santa Monica, CA, December 1965.
- [4] H. L. Dreyfus. *What Computers Still Can't do*. MIT Press, 1992.
- [5] M. Heidegger. *Sein und Zeit*. 1927.
- [6] A. Hoffmann. *Paradigms of Artificial Intelligence - A Methodological and Computational Analysis*. Springer-Verlag, 1998.
- [7] W. S. McCulloch and W. Pitts. A logical calculus for the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–143, 1943.
- [8] A. Newell. Physical symbol systems. *Cognitive Science*, 4:135–183, 1980.
- [9] A. Newell and H. Simon. GPS, a program that simulates human thought. In E. Feigenbaum and J. Feldman, editors, *Computers and Thought*, pages 279–293. McGraw Hill, New York, 1963.
- [10] A. Newell and H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery*, 19:113–126, 1976.
- [11] N. Nilsson. Eye on the prize. *AI Magazine*, (Summer), 1995.
- [12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, I: Foundations*. MIT Press, Cambridge, MA, 1986.
- [13] H. A. Simon and A. Newell. Heuristic Problem Solver: The Next Advance in Operations Research. *Operations Research*, 6, January/February 1958.
- [14] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988. (Includes peer commentaries).
- [15] P. J. Werbos. Brain-like intelligence in artificial models: How can we really get there? *INNS Above Threshold*, June 1993. A publication of the International Neural Network Society.
- [16] T. Winograd and F. Flores. *Understanding Computers and Cognition: A new Foundation for Design*. Norwood Publisher, 1986.
- [17] L. Wittgenstein. *Philosophical Investigations*. Macmillan, Oxford, 1953.
- [18] L. Wittgenstein. *Tractatus logico-philosophicus*. Routledge & K. Paul, London, UK, 1961. First German edition in *Annalen der Naturphilosophie*, 1921.