

In his very comprehensible essay *An Introduction to Mind Design* John Haugeland explains why the vision of creating artificial intelligence could be realized with the help of computers. Haugeland’s main point is that *if* a mind, or part of it, is an “automated formal system” (I will explain what is meant by these terms below) *then* it can in principle be simulated by a computer. Most of Haugeland’s effort in his essay goes into explicating the exact meaning of the premise and various reasons why cognitive scientists can assume it to be true and try to implement the conclusion. He also points out reasons to doubt its truth. In the following I will try to summarize the remarks that seem most significant to me.

First of all I will define, in a very abstract manner, what is meant by an “automated formal system” (AFS hereafter). A formal system is defined by an initial state, and a set of rules according to which states may be manipulated to yield new states. If the system applies the rules to the symbols by itself it is “automated”.

What is so special about AFSs? It can be proven, using the notion of the Turing machine, that any such system can be implemented on a computer<sup>1</sup>, which can then be said to “formally imitate” it. The obvious consequence is that if minds were AFSs, they could be formally imitated by computers. However, it is less obvious why the latter assumption should be true. I will now consider three objections to the thesis that a mind could be an AFS and how Haugeland relates to them.

First, Human minds are not necessarily “digital” as AFSs are. “Being digital” is defined by Haugeland as being self-contained (relating only to its own states), perfectly definite (without any uncertainties) and finitely checkable (only a limited number of actions can be done in a certain state). On the contrary, human minds might very probably be “analog” systems, which implies the denial of at least one part of the definition. Haugeland shows that for a digital system it is not even possible to approximate an analog system if the latter is sufficiently complex. Thus it is inherently impossible to do a molecular simulation of the human mind in the digital computer, but whether this prevents us from finding abstractions that still work as desired does not follow from the criticism. Haugeland thinks that we should stick to digital computers instead of switching to analog ones since the former scale better: Small errors that occur in real analog systems variables are prone to increase when these values are further processed. Since this kind of error cannot occur in digital systems, they are preferable to analog ones.

Second, Human beings know the semantics (say true or false relative to the real world according to some interpretation function) of their “states”, while AFSs do not, since they purely work on a syntactic level. To meet this objection Haugeland first explains the logical notion of soundness of the rule system: If symbol manipulation rules only modify a given state that is interpreted as “true” in a way that produces another state that can also be interpreted as “true”, and we start the system with such a state, then we can assign the value “true” to any state that can be produced by the rules. Obviously, the AFS still does not “know” the interpretation of the states, but we get the desired semantics anyway, for “free” as Haugeland puts it.

---

<sup>1</sup>for Turing machines we have to assume infinite memory to work, which we cannot provide in reality, but today's computers can have almost arbitrarily large memories, probably more than a mind has at its disposal, so this does not seem to be a real restriction

---

Third, Human beings do more than express truth. In a conversation, for example, it is not sufficient to state true statements, it is rather required to contribute to the topic while obeying certain social rules. For this to work in an AFS, according to Haugeland, it needs four ingredients: A common sense reasoning framework (a set of rules that implements “rationality”), a way to translate incoming information into adequate symbols and to translate states into adequate actions (input and output transducers, respectively) and the social rules (“conversational cooperativeness”). It is not obvious what should prevent the cognitive scientist from adding these ingredients to his AFS, although it is probably far from easy to do so.

If, regardless of these objections, we assume the mind to be an AFS, what do we gain? Haugeland euphorically claims that the most widely discussed dilemmas in philosophy of mind would be circumvented: (i) the mind-body problem; (ii) the problem of explaining the relevance of meanings; and (iii) the problem of objectively verifying mental explanations. I will shortly expand on how the computational idea deals with these issues.

The mind-body problem is “solved” by claiming that a mind is, speaking bluntly, basically a biological computer. This is, as Haugeland points out, not as big a restriction as it might seem: Formal systems can reach any level of complexity by operating on states of lower level formal systems as their symbols. Their performance in a computer can only be seen as “dealing with ‘ones’ and ‘zeros’” at the lowest level, the upper levels can have much more advanced tasks.

Meaning, that is, semantics could be introduced when the system is created and preserved by its rules, thus in operation there would be no need for interpretation of the states and rules. Our impression that we actually do more than stating true statements could be reflected in an AFS by implementing rationality, input and output transducers and social rules as central parts.

Mental explanations would be expressed in terms of symbol manipulations of the AFS, regardless of whether it is implemented on a computer or a human mind. Again, Haugeland points to a common misconception: Just because a computer follows its rules infallibly, it does not follow that it will not do any mistakes. It can rather “infallibly follow quite fallible rules”. This allows it to err just as human minds do.

To summarize Haugeland’s essay, I would say as follows: There are reasons to believe that a mind cannot be simulated on a digital computer, but none of them can be used in a proof. However, in my opinion these reasons are only part of the story, and important aspects are left out.

First, I believe the artificial intelligence approach could be much more directed to real human beings than Haugeland suggests. The brain, for example, seems to have no problems in using analog computation and thresholds and controlling errors by introducing redundancies. I doubt that there exist digital systems with the same interpretation that are easier to compute than a complete (molecular) simulation, which Haugeland ruled out.

Second, it seems to be a distinguishing mark of intelligence to be self-reflexive, to be able to look at oneself from “above”. This is not possible in the stacked formal system view Haugeland proposes: There must be a “highest” system which cannot be looked at from “above”.

Thus, although the advantages of the mind being a formal system are very appealing, I do not believe that the formal system approach has a chance to succeed.

---