IJCAI SYMPOSIUM
UNDER WHAT CONDITIONS CAN A MACHINE USE SYMBOLS WITH MEANING?

W. A. Woods
Bolt Beranek and Newman Inc.
10 Moulton Street
Cambridge, MA 02238

Most existing computer systems Use representations that have meaning to their designers, but do not stand in any relationship to their intended meaning whereby one could say that they had that meaning to the machine itself. I will be concerned here with the question of when symbols*actually have meaning to the machine. This question presupposes the answer to a prior question: why should one maintain a distinction between the meaning of a symbol and rules that simply imply its applicability or inapplicability? I will attempt first to answer this question, from which an answer to the former will follow. Briefly, I will argue that the distinction is necessary in order for a system to make proper adjustments to its knowledge base as a result of experience.

In a previous work [1], I have argued that the meaning of a symbol resides in an abstract procedure, not necessarily executable, linking the symbolic expression to the physical world through the (computational / inferential) operations of a physical interpreter operating on a combination of internal representations and sensory/motor connections to the world. Moreover, I argued that this meaning specification (the meaning function) is distinct from the procedures that one ordinarily uses to conclude the truth of a description in everyday situations (the recognition functions).

I will assume here that both meaning and recognition functions can be represented by collections of rules of the form "p implies q'', provided that a distinction is maintained by the system between those rules that it applies as part of the meaning (i.e., "meaning postulates") and other implications between symbols (which I will refer to as hypotheses).

If a system is to try to develop a model of the world in which it operates, then it needs to record rules of the form "p implies q" where, in the simplest case, p denotes some observable attribute of the world and q denotes some other observable attribute. The utility of such a rule lies in the ability to conclude q without actually perceiving it. Let us call p a sensory symbol (for a system) if there are processes in the world that (physically) cause the system to conclude p or to conclude not p. I will take the physical processes which operate in this way as determining the meanings of such sensory symbols.

A system with no sensory symbols could not be expected to acquire a model of the world except through the intervention of an external agent (i.e., a programmer). In this case the meanings of its symbols would reside solely in the head of the programmer, and the system would be open to different interpretations by different external observers, however consistent its internal operations were with some particular interpretation.

Consider a system that has both sensory and nonsensory symbols and that hypothesizes rules of the form "p implies q" under some circumstances. A validated hypothesis "p Implies q" is useful to the extent that one knows under what conditions to invoke the hypothesis (i.e., one knows the meaning of p) and one knows what claims are being made thereby (i.e., one knows the meaning of q). Moreover, in order to determine the validity of the hypothesis, one needs to be able to determine the truth and falsity of p and q in most circumstances.

Suppose that this system encounters a violation to a pair of rules "p implies q" and "q implies r" by perceiving that p is true and r is not. If q is a sensory symbol and its truth can be determined by perception in the situation in question, then the appropriate response is simple — either the "p implies q" rule or the "q implies r" rule should be modified, depending on whether q is false or true. If q Is not a sensory symbol, the system needs some criterion to decide which of the two rules to modify. Since one rule implies q is true, while the other implies it is false, what the symbol q means (either to the system or to an external agent) becomes important at this point in deciding what to give up.

To claim that the symbols in a machine have meaning is to claim that above and beyond the rules and procedures that lead them to be deduced by the system, there is some further claim about the state of the world that is being made by virtue of such a deduction. That is, regardless of one's reasons for believing p, there is some claim about the world that p makes, which may be either true or

I will use the word "symbol" here as a generic for the entire class of symbolic expressions used by the system, including designating expressions, predicates, propositions, etc.

false. Without such a distinction, a system would be unable to formulate a predictive model of its world, and could at best keep a record of its experiences; no deduction that it made could ever claim anything more than that it had made the deduction.

One of the main benefits from having a notion of the claim made about the world by a symbol (i.e., its meaning) is that if one encounters a sequence of deductions from some sensory symbol p to some sensory symbol q, while perceiving that p is true and q is false, then one can attempt to localize the erroneous rule by examining the truth of the intermediate symbols involved in the deduction. In particular if some intermediate symbol s is determined to be true (by virtue of its meaning) then the deduction leading from p to s may be left intact and attention directed to finding the error in the deduction from s to q (and conversely if s is determined to be false).

Without some such criterion for localizing the error, any one of the intermediate rules could be changed to eliminate the inconsistency in the momentary situation, but with little confidence of the effect on other hypotheses (perhaps already strongly verified). That is, it would be inappropriate to change the meaning rules (and thereby the meaning) of any of the symbols already involved in partially verified hypotheses, thereby changing the claim made by those hypotheses and either invalidating them or weakening their previous validation.

One can imagine situations in which the minimal change to an entire belief system in order to bring it into line with some body of observations is to make a change in the meaning of some symbol, thereby eliminating inaccuracies in, or increasing the coverage of, existing hypotheses. However, such a change would have to come from a complex analysis of a large portion of the system's knowledge base. This would be a revolutionary rather than evolutionary modification and would not be an appropriate mechanism for performing routine, incremental modifications to knowledge.

If one considers such modifications in an internal language of thought, coupled to symbols in the external vocabulary by connections that can themselves be learned and modified, then any such revolutionary change could also be effected with no changes in the meanings of internal symbols by merely moving the affected hypotheses (and appropriate external symbols) to new (internal) symbols with the desired meanings rather than changing the meanings of the old ones. Thus, in the internal language there appears to be no need to ever change the meaning of a symbol.

An obvious question, at this point, is whether an analysis along the above lines can apply to natural kind terms as well as to terms with obvious necessary and sufficient conditions. I believe that this is the case, although the story now becomes a little more complex. In this case, the question of interest about the meaning of a symbol has to do with what natural kind the system has got hold of when it forms a hypothesis. (In general a satisfactory account of natural kinds must allow the meaning rules to leave "gaps" in which the system has no intentions as to whether some hypothetical situation embodies an instance of a symbol or not, and must permit meanings to be extended in certain circumstances by narrowing the gap — see [1] for further discussion.)

Consider the hypothesis "snakes are harmless" expressed in New England, where there are virtually no poisonous snakes. One may ask whether the meaning of the internal symbol for snakes in this case refers to only (nonpoisonous) snakes of the kind that live in New England, or to all kinds of snakes. (I assume that both these classes are natural kinds). The answer, I would argue, depends on the intentions of the person when the concept snake was formulated and the preservation of those intentions in the meaning rules for the symbol (i.e., whether the scope of the meaning rules was sufficiently broad to cover all snakes or included enough specific characterization of the range of snakes experienced to rule out the poisonous kinds that exist elsewhere).

Once again, the meaning intended and encoded in the meaning rules will determine the reaction to encountering an anomalous situation — in this case, whether to attach the external word "snake" to a new internal concept which includes poisonous snakes, or to subclassify the the internal concept and move the "snakes are harmless" hypothesis to a more specific subclass.

If we take the above simplified story as the reason for maintaining a distinction between the meaning of a symbol and a mere collection of conditions that imply it or that it implies, then it follows that for symbols to have meaning to a system, there must be sensory symbols to which the nonsensory symbols are related by a distinguished set of meaning rules or some equivalent mechanism. These rules together with their interpreter and the physical processes that govern the senses determine the meanings of the system's symbols.

By this account, a system that read stories and answered questions about them, based solely on the manipulation of internal representations with no experiential base, would, from its own perspective, be manipulating meaningless symbols. Similarly a pocket calculator would be a manipulator of symbols whose meanings were externally attributed (even though its internal structure does faithfully model the intended interpretation). On the other hand, a simple robot creature with primitive perceptions can be said to have meanings for its symbols that are not externally imposed if its symbols are tied to those perceptions by some distinguished mechanism equivalent to meaning rules.

Reference

[1] Woods, W.A., "Procedural Semantics as a Theory of Meaning," in A. Joshl, I. Sag, and B. Webber (eds.), Elements of Discourse Understanding. Cambridge University Press, 1981, pp 300-334.