HOW CAN A PROGRAM MEAN?

Donald Perlis

University of Maryland
College Park, Maryland
USA 20742

Abstract:

This paper is about meaning, in the following sense: If a system employs symbols, in what sense are they symbolic, of *what* are they symbolic, and in what sense is it the *system* that makes them symbolic? In other words, what does it take for a system to be such that the question "what do you mean by that?" can be appropriately asked it? We suggest an answer based on the idea of quotation or reification.

I. Introduction

The problem of meaning is a part of the general question of the mechanical nature of cognitive or mental states, e.g., of belief, emotion, sensation, perception, intelligence, understanding, intentionality. The latter, intentionality, is particularly tied to the idea of meaning, in the sense that to attribute a meaning to an internal expression is to "intend" something by the use of the expression. As such it has been an issue in the philosophical underpinnings of artificial intelligence even before AI was a recognized field, e.g., in the famous Turing test, or even much earlier work in the philosophy of mind. The puzzle, in modern terms, can be stated as follows: there is plenty of indirect evidence for the contention that the mind, at least in humans, is a kind of process in the physical brain, and yet mental events as we currently understand them seem to have little in common with familiar material things.[1]

Much has been written on this problem, largely by philosophers. There seem to be at least three main camps. Some (dualists, e.g., Popper (1965)) argue that meaning is not a mechanical notion at all, but that it is a dual phenomenon to material aspects of behavior, and therefore not to be understood in ordinary scientific modes of discourse. Others (functionalists, e.g., Dennett (1978)) claim that meaning is merely a useful terminological category one agent uses in reasoning about (the functioning of) another (the "intentional stance" of outsiders). Still others (empiricists, e.g., Thagard (1986)) argue that there may be internal phenomena that clearly produce genuine 'attribution of meaning' within an agent, but that the specific character is yet to be discovered. This latter contention is addressed (at least to some degree) by recent work in artificial intelligence, sketched below.

One important view on the question is called the direct-reference theory of meaning. It asserts that the meanings of certain tokens (e.g., words or expressions) are particular external entities. For instance, the expression "John's house" may have as its meaning a particular house. The difficulty then is just how that relationship between the words and the house is determined. Kripke (1979) has provided a particularly telling example of the inadequacy of this theory. Others, e.g., Sayre (1986), have stessed the importance of interactivity between the system and the external world, so that the meaning of "John's house" might be related to the speaker's experiences with people and buildings, a kind of external-reference-through-experience theory. Putnam (1970,75) argues that whatever meaning is, it is not in the head, i.e., he favors some version of an external reference theory. Searle argues that *no* program can understand (or, presumably, mean). Lebowitz (1986) states that a formal symbol manipulator can be semantical if it has a rich enough symbol structure. Sayre (1986) responds that it is still merely formal, and so has no outer meaning, no tie to the world. Stich (1984) argues that there may be no firm tie to the world (of interest for intentionality).

This quick run-through of a variety of positions is intended to illustrate the breadth of attention that the issue has attracted. Space does not allow detailed description of these positions, let alone serious contrast with our own below. For more on the literature, see (Pylyshyn 1984), and especially (Minsky 1968), (Waltz and Boggess 1979), (Sloman 1986), and (Steels 1986) for views related to the present one.

Now, one answer above, offered by McCarthy (1979) and Dennett (1978), is that if there is an informative answer to "what do you mean by that?" then the question is in some sense appropriate. For instance, McCarthy points out that one might ask what a thermostat means by its thermometer pointing at 80 degrees Fahrenheit and its thermocouple opening the circuit so that the furnace stops. The reply that the meaning is that the room is too hot is informative and for some purposes satisfactory; for instance if someone has held a match near the thermostat then it is tempting to attribute to it a "belief" that the room shares its own high temperature. Whether or not the thermostat "really" believes anything, or means anything by its shifting states, could be regarded as a terminological question, and the more interesting issue as that of the usefulness of the description to others in dealing with the system in question (the "intentional stance" taken by us regarding a system we want to understand).

This is an appealing doctrine, but perhaps it clouds certain things. For instance, it sheds little light on just what sorts of systems are ones we are likely to take an intentional stance toward. The thermostat example, for instance, is one we are not likely to find very useful if we continue to probe. The thermostat's behavior is far too rigid for us to be able to usefully attribute much in the way of cognition to it. McCarthy discusses more complex examples, such as intelligent agents in a *game of Life* automaton, where it may be more tempting to ascribe beliefs. Still, this

[1]This is often discussed in terms related to the mind-body problem, and the problem of consciousness: how can a material/mechanistic entity have mental states?

leaves open the issue of the conditions under which such ascription is appropriate, aside from the purely pragmatic one of apparent usefulness to others. So the question that arises is whether there are any qualitative boundaries here, or whether it is all simply a matter of degree. That is, there might be certain key behavioral modes that determine fairly clearcut distinctions among systems, such that we would strongly tend to take the intentional stance toward some and not others. This is the theme we explore in this paper, with particular attention to the issue of meaning: what does it take for a mechanical system (such as a program) to use symbols meaningfully? Or, following the intentional stance, what does it take in a mechanical system for us to significantly benefit from regarding it as using symbols meaningfully?

II. The quotation approach

Presumably there are many desiderata one might suggest in answer to our main question. Instead of trying to present a substantial list and then categorize and discuss all the alternatives, I simply present for analysis two related ones that bear specifically on a tack that has shown tentative progress in (Perlis and Hall 1986) and (Perlis 1986). For a system to (be said to) use symbols meaningfully, it should be able to:

1. take stances of its own (beliefs), yes and no, toward informational structures (symbols), and

2. distinguish between its symbols on the one hand and what they stand for on the other.

What we have in mind in these desiderata is that a belief is something believed to be *true*, and that therefore the concept of a representational structure being *true or false* is relevant. This leads into the second point, for in order to take the stand that a certain structure is, say, false, it is necessary to distinguish it from what it supposedly is about. This is much like saying a word is different from what it stands for, and can even be misused. If I mention the dog by the tree and you say it isn't a dog but rather a wolf, you have recognized the word 'dog' as having being misapplied to the creature by the tree, rather than thinking some dog by the tree is also a wolf or has been replaced by or changed into a wolf.

How then are we to address desiderata 1 and 2 above? We start by observing that typical AI systems today do not mean anything at all to themselves. They cannot compare their use of a symbol and some other entity that it is supposed to refer to, because they do not use symbols *to refer*. Even though they may have rules associating 'bowl' and 'container' (and many more complex ones), they still do not have a rule that accounts for 'bowl' being *used* to refer to a container. There are bowls and containers, or expressions 'bowl' and 'container', as you like, but not the relation between use and mention. There is only one level of symbolism, which is to say, no symbolism at all. In particular, there is no facility to state something like "the thing you called a 'bowl' is really a cup", and even less to revise word use on such a basis. But a (formal) system that does have a quotation mechanism could ask itself whether, say, the thing it called a 'bowl' is really a bowl or a cup.

That is, a symbol (to a system) is something used (by the system) to stand for something else, i.e., that the system itself has *both* symbol and symboled at hand. Thus a bowl and 'bowl' are related; my saying so attests to my having two modes for 'bowl', the use and mention modes. If I have only one mode, then I do not relate the word and the (supposed) object. But note that even if I do relate them with two modes, I do not thereby actually have a bowl in my head, or even necessarily in my hand. That is, we are reduced to dealing with symbols and their meanings, whatever they are, via expressions or other internal forms. We categorize things: this is a that, yet 'this' and 'that' are expressions; we never get to the outer 'thing in itself'.

But why bother? How will quotation help in the understanding and design of intelligent systems? Well, we can ask, how did it evolve in us? Presumably it endows us with some advantage, and I hope my discussion already points to a clear one: flexibility of behavior, ability to deal with a changing and complex world we cannot know in detail in advance. We must at times go on not only incomplete but also faulty knowledge, and be prepared to change our minds, while still remembering the past error. We had better not confuse our dreams for reality for long, nor our longings for the truth. Any sort of planning activity must have some amount of this, but not necessarily very much; in particular, revising word use, though very important, does not enter into most current AI planning systems.

To get into our main concern, consider an empty box, that has no meanings of its own. If we put in it a slip of paper with the word "bowl" on it, the box-plus-slip-plus-word still has no meaning of its own. If there is a bowl situated outside the box, the word on the slip of paper may mean the bowl to *us*, but not to the box. What is the difference between the box and us? Well, we have, in a sense, both the word *and* the bowl to contrast and relate; that is how we manage to say, if asked, that the word 'bowl' means the (or a) bowl. How can this be? We do not have bowls in our heads. Well, we do have something that we use for the bowl, different from the word 'bowl', *at least when we are pointing out that we are using the one for the other*. It may be a visual image, another word or expression, or a conglomerate of things. For simplicity let us think of our retinal image, containing an image of the bowl, and another of the box with the word 'bowl' in it. Then we are saying that the one image is related to the other in that we use the one to call attention to the other. Since the box does not have two such internal entities to manipulate, it is different from us in a crucial respect.

Note however that in reality neither the box-plus-word nor the actual bowl are in our heads; all this using/meaning/referring is going on in our heads with our own structures. It is a contingent (and fortunate) matter that there is a close tie between certain of these internals and certain externals. Now, if we endow the box with a quotation mechanism, so that it can write on paper, not only the word 'bowl' but also a notation it can relate to the word 'bowl', such as "bowl" (i.e., it's own quotes) then it is moving toward being able to distinguish between word and object as in desideratum 2. Note that it further is moving toward desideratum 1 in that once it can form statements *about* its own structural forms, then it has some of the tools needed to assert that various forms are true or false. In other words, quotation seems to be just the kind of

mechanism needed.

Note again the oddity that what at first was simply taken as a bowl without question, is now instead reduced to (or quoted into) a 'bowl' or a 'bowl-image' or simply 'that thing I had in mind', and then inspected critically. We may then conjure up a supposed real bowl concept C to which we compare the former (C itself being simply another internal form), or momentarily surrender by deciding we aren't sure what we are talking about, whether there is a cogent notion of bowl at all. But it all stays *de dicto*.

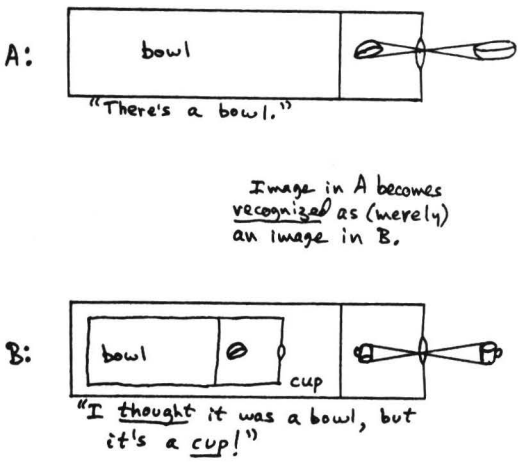
How then must the box be augmented, what must it be endowed with, in order that we take the intentional stance with regard to it? How can we make it be like us? Our discussion so far suggested that at the very least, we should provide it with *two* internal entities, such as the word 'bowl' and a bowl-image. Suppose then that we fit a lens to the box, and a screen inside, so that an image of the bowl appears on the screen in the box. Now the box has both 'bowl' and a bowl-image. Does this mean the 'bowl' means a bowl (or even its bowl-image) to the box? Clearly not: there is no internal *tie* between the two.

There may arise an uncomfortable feeling that we are wavering on the edge of the chasm of infinite regress here, and that we must postulate a ghost in the machine to account for any "real" intentionality, any "real" meaningful tie between word and referent. But our approach still has more to offer.[2] Recall that we want the box to be able to use "bowl" as if it meant the bowl or bowl-image. Now just how a box can use anything is a question of importance but not the point we are aiming at right now. Rather, it is the odd situation that arose in our attempt two sentences ago to describe what we want the reference to consist of: shall "bowl" refer to the (external) bowl, or to the (internal) bowl-image? Our position is that *there shall be no difference* between these *until it occurs to the box* that there is a difference.

Now this is a tall order, for we seem to be compounding the problems rather than simplifying them. However, this is the beauty of the idea of quotation. For quotes can be placed around previously unquoted entities. Thus at first the box may simply take the bowl-image *to be* a bowl in front of it; but then later it may (introspectively) judge that this must be something in its head formed by its lens and so on, so that really (so it thinks) there is some other thing out there similar to its bowl-image (which it had been calling "bowl"). That is, it puts quotes around its (newly thought-up) entire process of image-ing and naming a bowl, in a grand reification we shall call "reflection." The entities so reflected need not however be tokens or images; the idea is that any mental objects can be reflected.

[2]In addition to our ideas below, we single out the contributions of Sloman (1986), Steels (1986), and Waltz and Boggess (1979), which discuss some possible advantages of internal representations. Waltz and Boggess in particular consider the presence of pairs of tokens and images used within a computational system. However, all of these leave unexamined the question of flexible use of tokens and internal recognition of the fact of reference itself, which is our main concern here: what is it for the mechanism to refer "intentionally", i.e., to know it is using words for objects? This we suggest may be approached via the idea of reflection to be discussed.

We have tacitly endowed the box with lots of thinking devices: reasoning with introspection, quotation mechanisms, temporal information, knowledge of its physical features (lens, etc). Our contention, however, is that *these* things are at least somewhat understood phenomena in the current state of artificial intelligence. In the spirit of the adage that a picture is worth a thousand words, we offer the following illustration of the idea of reflection and quotation.



III. Conclusions

An answer then to why bother to have two notational tokens, such as bowl and 'bowl', is to distinguish what is from what isn't. For instance, I may change my mind that I have seen a bowl, but to use this fact (that I have changed my mind), I recall that I used 'bowl' inappropriately, or that I entertained the sentence "there is a bowl present". Quotes (or words as such) allow us to entertain possibilities, even ones we think are false. By 'quotes' I mean simply names; I refer to the capacity for creating structures to manipulate vis-a-vis one another. This can apply to images or any other structures. But crucial to it is a mechanism for relating name and named, essentially a truth-predicate (or reality-predicate): the bowl-image is of a bowl, or 'bowl' stands for a bowl. Then we can choose between hedging (maybe that isn't a bowl) or going for it (that is a bowl) where 'that' is some other internal entity such as an image. The main point though is that not only 'that' but also the considered reference (bowl) is internal to the system, even if it is not quoted. To draw out the illustration further, a bowl 'becomes' (under suitable circumstances) 'bowl' and then may not be a bowl after all. This strange statement may seem less so when taken with the further claim that, as far as meaning goes, all is imag-inal. As long as thinking works, we use it, but possibly there are no 'firm' bowls at all. This is reminiscent of natural kinds, which often defy definition.

How avoid the criticism that then we never think about *real* things? Well, here we can borrow from the

adverbial theory of perception, which maintains, for instance, that Macbeth was "perceiving dagger-ly" in the famous scene in which he seems to see a dagger before him. By way of analogy, we may think "aboutly", that is, when I think about a cow, I really am thinking in an "about cow-ly" fashion, or better put, I "refer cow-ly". That is, I have (at least) a pair of tokens, such as cow and 'cow', in my head, that I am using to form hypotheses, reason concerning what is or isn't. Notice that this implicitly forces an external reality on us at least in terms of a natural explanation of what it means for such token-hypotheses to be or not to be the case: the "referring cow-ly" thinker may *take* her tokens to be real. Now, whether or not they *are* real (externally), i.e., whether or not there is a (natural) external referent for the internal tokens, becomes contingent, much as in the adverbial theory of perception. We may refer "unicorn-ly" and yet have no external referent; the same for a cow which may be referred to in error (if there is really no cow that is the object of ones thought).

When we reflect on our behavior, we are in effect putting quotes on it so that it no longer is simply taken at face value. The word 'bowl' used above in direct-reference mode, as it were, suddenly becomes for us a word different from its meaning which now is, perhaps, an image. But there is a *presumption of an external object of thought*, something that we take as real. Expressions or other internal forms (even images) do all the work, but at least one is momentarily taken *as* the thing-in-itself. We have no other way to refer, no casting our mind forward to external things.

Is there any external tie worthy of mention? Perhaps the best we can do, and what we should do, is find rough partial isomorphisms between the networks of quotational forms in different agents. This might allow us to say that, with degree d, agent p and agent q have the same belief about entity r. That is, if the quotational network Np of (beliefs of) p maps (almost) uniquely to the world, and if the same holds for Nq, and if r is in the joint range of those maps, then it may be possible to measure a degree d of similarity between those elements of Np and Nq connected to (the pre-image of) r. This is the current state of ongoing efforts to delineate the view urged here. My hope is to be able soon to provide examples of artificially simplified worlds in which meanings can be pinned down by means of such rough partial isomorphisms between agents' belief sets.

Acknowledgement

I would like to thank Robert Audi, John Barnden, Rosalie Hall, Jim Reggia, Kenneth Sayre, Aaron Sloman, Brian Smith, Luc Steels, Stephen Stich, and Joseph Tolliver for stimulating discussion of this topic.

REFERENCES

Dennett, D. (1978) *Brainstorms*. Montgomery, VT: Bradford Books.

Kripke, S. (1979) A puzzle about belief. In: *Meaning and Use*, ed. A. Margalit, pp. 234-283. Holland: Dordrecht.

Lebowitz, M. (1986) Semantic information: inference rules + memory. *Behavioral and Brain Sciences, 9(1)*, pp. 147-148.

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical Perspectives in Artificial Intelligence*, ed. M. Ringle, Humanities Press.

Minsky, M. (1968) Matter, Mind, and Models. In: *Semantic Information Processing*, ed. M. Minsky, pp 425-432. Cambridge: MIT Press.

Perlis, D (1986) What is and what isn't. Symposium on intentionality, Society for Philosophy and Psychology, Johns Hopkins University.

Perlis, D. and Hall, R. (1986) Intentionality as internality. *Behavioral and Brain Sciences, 9(1)*, pp. 151-152.

Popper, K. (1965) *Conjectures and Refutations*. Basic Books.

Putnam, H. (1970) Is semantics possible? In H. Kiefer and M. Munitz (eds.) Language, Belief and Metaphysics, SUNY Press.

Putnam, H. (1975) The meaning of 'meaning'. In K. Gunderson (ed.) Language, Mind and Knowledge, Univ of Minnesota.

Pylyshyn, Z. (1984) Computation and Cognition. MIT Press.

Sayre, K. (1986) Intentionality and information processing... *Behavioral and Brain Sciences, 9(1)*, pp. 121-138.

Searle. J. (1980) Minds, brains, and programs. *Behav, and Brain Sciences* 3, 417-424.

Sloman, A. (1986) Reference without causal links, *Proceedings, 7th ECAI*, July 21-25, 1986, Brighton, UK. 369-381.

Steels, L. (1986) The explicit representation of meaning. *Proceedings, Workshop on Meta-Level Architectures and Reflection*, Sardinia, October 1986.

Stich, S. (1984) *From folk psychology to cognitive science: the case against belief*. Cambridge: MIT Press.

Thagard, P. (1986) Parallel computation and the mind-body problem. *Cognitive Science*, 10, 301-318.

Waltz, D. and Boggess, L. (1979) Visual analog representations for natural language understanding. Draft, Advanced Automation Group, Univ. of Illinois.

This research has been supported in part by the following institutions:

The U.S. Army Research Office (DAAG29-85-K-0177)
The Martin Marietta Corporation