PANEL DISCUSSION
UNDER WHAT CONDITIUNS CAN A MACHINE ATTRIBUTE MEANINGS TO SYMBOLS?

Drew McDermott
Yale University

1. In what sense do humans assign meanings to symbols?

2. In what sense do humans think they assign meanings to symbois?

3. Could machines assign meanings the way humans do?

4. Could machines think *of* their meaning-assignments the way humans do?

Many humans think that the first step in perception is sensation, and that larger percepts are made of atomic sensations, we now know that it is not necessary to appeal to sensation to explain perceptions. (You could arbitrarily label the first events in perception "sensations", but there is no compelling reason to.) It looks now as if we could design a complete vision system that had no intuitions whatever about "atomic feelings" occurring inside it. tony then do we think sensations play a role? How could we augment the vision system so it could think this about itself?

Similarly, we should expect in studying meaning (and other mental phenomena) to find discrepancies between what actually happens and the way we perceive it. One such discrepancy accompanies our overwhelming intuition that when we know that a symbol refers to something, we are "connecting" an abstract symbol to a concrete reality. I don't just have a formal theory about an individual named Reagan; I know who this name refers to.

But when we build robots, there is nothing inside the robot to actually "connect" a symbol to. In fact, the major intellectual achievement of computer science in western culture is to demonstrate that a device can manipulate symbols incorrectly" without knowing what they mean. Previlusly, mind-theorists who tried to build theories based on symbol manipulation kept stumbling over the homunculus required to understand the symbols being manipulated.

Does this mean that a robot does not succeed in referring to the world at all? No, as the following sketch should demonstrate.

A robot can be modelled as one or more formal systems, connected to the world by way of sensors and effectors. They will be formal, systems in that they operate by applying elementary operations to uninterpreted symbol structures, deriving new uninterpreted symbol structures. The sensors create the first symbol structures; the effectors receive some of the inferred symbol structures, and execute them as instructions.

This works because the formal system has an interpretation, which matches the real world closely. I.e. there is an assignment of real-world entities, properties, and propositions to the symbol structures of the formal system, with the following properties:

* Input soundness: the sensor apparatus is constructed so that the symbolic structures it constructs usually correspond (under the interpretation) to actual states of affairs; that is, they are usually true.

* Inferential soundness: the formal system is constructed so that it tends to infer true symbolic structures from other true symbolic structures.

* Output soundness: the effector apparatus is constructed so that it tends to perform the actions that the interpretation assigns to the symbol structures given to it.

For example, when a car is coming, the robot's sensors put an expression (CAR-COMING) into the formal system, which infers (SHOULD-DO (JUMP)), and sends (JUMP) to the effectors, which actually causes a jump.

having an interpretation, does not mean that the formal system has access to something, but that there exists an interpretation in the mathematical sense. It is futile to ask for more. If the robot had access to a thing purporting to "contain the meanings" of its symbols, then either this thing would be just another formal system, and wouldn't contain the meanings after all; or the robot would not be a Turing machine, but some more mystical entity. If we want to stick with the formal robots that have revolutionized our thinking about the mind, we must place the interpretations of their symbols outside their heads.

So the interpretation can play no functional role in how the robot works. It is simply an analytical tool, to explain how it works.

what if there is more than one interpretation of the formal system? After all, any interesting system will have an infinite number of different models. But most of them will fail to satisfy the three soundness conditions. There is an interpretation in which (CAR-COMING) means "Rice pudding present", and (SHOULD-DO (JUMP)) means "My mother is standing on her head", but the sensors, effectors, and inference machinery do not work correctly under this interpretation. 1 expect that for any robot, there is a "standard sound interpretation" that does satisfy the soundness conditions in our world. Since these conditions are stated as engineering precepts ("component X usually operates according to the interpretation"), there are probably lots of unimportant variants of the sound interpretation, but 1 will ignore this issue.

1 have now answered my original questions 1 and 3, how meanings actually work. We must now ask how humans and robots might think about how they work, hirst, let me acknowledge that there is a mystery here about why humans have any opinion at all about whether they assign meanings to symbols. The answer might be that humans are just naturally inquisitive, and have opinions about everything, but perhaps there is some special reason to have opinions of certain kinds about oneself.

The symbols people think they manipulate are not those in the robot's formal system. When people think of symbols, they think of words, names, mathematics, and road signs. Suppose a human, to be specific, Edwin Meese, is mediating on the meaning of "Reagan", he knows that this symbol refers to his boss, person he sees often. He believes he can think about Reagan any time he wants, and that this name refers to that person.

The truth is that for Meese to think about Reagan is for him to manipulate symbol structures that refer to Reagan in his standard sound interpretation. Ironically, for him to think about the name "Reagan" is to manipulate symbol structures in much the same way. (Since the name is a social object about which various things are known, e.g., it's pronounced differently from the name of the Treasury Secretary.) The idea that the one object refers to the other is a third symbol structure, used mainly by natural-language modules.

The mind conceals such facts about itself. (It is usually pretty easy to see why, but that isn't my main topic.) It tells itself that to think about Reagan is to "directly apprehend" him. You can think about an object if you are "acquainted" with it, if you know "which object it is".

This is contrasted with another situation we have all been in, where we know someone (or something) by name only. Suppose someone has been marooned on a desert island for twenty years, and, having returned, hears people blame someone named "Reagan" for all our troubles. Obviously, ,this is someone he ought to know about; he ought to know who the name refers to. Somehow, just knowing that Reagan is "the person everyone is blaming for our troubles" is inadequate. After finding several more facts, and seeing Reagan on television, he begins to feel tht he is "knows who Reagan is", that he can think about him any time he wants.

People feel a sharp difference between only knowing something's name and knowing the object directly. 1 think this is an illusion; in reality, one accumulates information about an object gradually. There is no well defined point at which one is "really" acquainted with it. The sharp feeling is akin to a sharp feeling of hunger; there is no qualitative difference between an empty stomach and a full one, but it feels like there is; if it didn't, you wouldn't work so hard at finding food. In the case of acquaintance, you need a reason to work hard to gather information.

whatever the source of this feeling, it leads to disbelief that all knowing is mediated by a formal system. If all you have is symbols, then you aren't "really" acquainted with anything, and you don't "really" understand anything. In fact, the sketch 1 started with explains quite satisfactorily how a purely formal system can nonetheless deal with the real world, and, in a certain sense, have its symbols mean things in that world. Being "directly connected", *or* "knowing what the symbols mean", plays no role.

This answers my second question, How do people think they assign meanings to symbols? The fourth question, how might we get machines to think about themselves this way?, 1 will leave unanswered.