The Little Thoughts of Thinking Machines

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1983

When we interact with computers and other machines, we often use language ordinarily used for talking about people. We may say of a vending machine, 'It wants another ten cents for a lousy candy bar.' We may say of an automatic teller machine, 'It thinks I don't have enough money in my account because it doesn't yet know about the deposit I made this morning.' This article is about when we're right or almost right in saying these things, and when it's a good idea to think of machines that way.

For more than a century we have used machines in our daily lives whose detailed functioning most of us don't understand. Few of us know much about how the electric light system or the telephone system work internally. We do know their external behavior; we know that lights are turned on and off by switches and how to dial telephone numbers. We may not know much about the internal combustion engine, but we know that a car needs more gas when the gauge reads near EMPTY.

In the next century we'll be increasingly faced with much more complex computer based systems. It won't be necessary for most people to know very much about how they work internally, but what we will have to know about them in order to use them is more complex than what we need to know about electric lights and telephones. As our daily lives involve ever more sophisticated computers, we will find that ascribing little thoughts to machines will be increasingly useful in understanding how to get the most good out of them.

Much that we'll need to know concerns the information stored in computers, which is why we find ourselves using psychological words like 'knows', 'thinks', and 'wants' in referring to machines, even though machines are very different from humans and these words arose from the human need to talk about other humans.

According to some authorities, to use these words, the language of the mind, to talk about machines is to commit the error of anthropomorphism. Anthropomorphism is often an error, all right, but it is going to be increasingly difficult to understand machines without using mental terms.

Ever since Descartes, philosophically minded people have wrestled with the question of whether it is possible for machines to think. As we interact more and more with computers — both personal computers and others — the questions of whether machines can think and what kind of thoughts they can have become ever more pertinent. We can ask whether machines remember, believe, know, intend, like or dislike, want, understand, promise, owe, have rights or duties, or deserve rewards or punishment. Is this an all-or-nothing question, or can we say that some machines do some of these things and not others, or that they do them to some extent?

My answer is based on work in the field of artificial intelligence (usually abbreviated AI) which is the science and engineering of making computers solve problems and behave in ways generally considered to be intelligent.

AI research usually involves programming a computer to use specific concepts and to have specific mental qualities. Each step is difficult, and different programs have different mental qualities. Some programs acquire information from people or other programs and plan actions for people that involve what other people do. Such programs must ascribe beliefs, knowledge and goals to other programs and people. Thinking about when they should do so led to the considerations of this article.

AI researchers now believe that much behavior can be understood using the principle of rationality:

It will do what it thinks will achieve its goals.

What behavior is predicted then depends on what goals and beliefs are ascribed. The goals themselves need not be justified as rational.

Adopting this principle of rationality, we see that different machines have intellectual qualities to differing extents. Even some very simple machines can be usefully regarded as having some intellectual qualities. Machines have and will have varied little minds. Long before we can make machines with human capability, we will have many machines that cannot be understood except in mental terms. Machines can and will be given more and more intellectual qualities; not even human intelligence is a limit. However, artificial intelligence is a difficult branch of science and engineering, and, judging by present slow progress, it might take a long time. From the time of Mendel's experiments with peas to the cracking of the DNA code for proteins, a hundred years elapsed, and genetics isn't done yet.

Present machines have almost no emotional qualities, and, in my opinion, it would be a bad idea to give them any. We have enough trouble figuring out our duties to our fellow humans and to animals without creating a bunch of robots with qualities that would allow anyone to feel sorry for them or would allow them to feel sorry for themselves.

Since I advocate some anthropomorphism, I'd better explain what I consider good and bad anthropomorphism. Anthropomorphism is the ascription of human characteristics to things not human. When is it a good idea to do this? When it says something that cannot as conveniently be said some other way.

Don't get me wrong. The kind of anthropomorphism where someone says, 'This terminal hates me!' and bashes it, is just as silly as ever. It is also common to ascribe personalities to cars, boats, and other machinery. It is hard to say whether anyone takes this seriously. Anyway, I'm not supporting any of these things.

The reason for ascribing mental qualities and mental processes to machines is the same as for ascribing them to other people. It helps understand what they will do, how our actions will affect them, how to compare them with ourselves and how to design them.

Researchers in artificial intelligence (AI) are interested in the use of mental terms to describe machines for two reasons. First we want to provide machines with theories of knowledge and belief so they can reason about what their users know, don't know, and want. Second what the user knows about the machine can often best be expressed using mental terms.

Suppose I'm using an automatic teller machine at my bank. I may make statements about it like, 'It won't give me any cash because it knows there's no money in my account,' or, 'It knows who I am because I gave it my secret number'. We need not ascribe to the teller machine the thought, 'There's no money in his account,' as its reason to refuse to give me cash. But it was designed to act as if it has that belief, and if I want to figure out how to make it give me cash in the future, I should treat it as if it knows that sort of thing.

It's difficult to be rigorous about whether a machine really 'knows', 'thinks', etc., because we're hard put to define these things. We understand human mental processes only slightly better than a fish understands swimming.

Current AI approaches to ascribing specific mental qualities use the symbolism of mathematical logic. In that symbolism, speaking technically, a suitable collection of functions and predicates must be given. Certain formulas of this logic are then axioms giving relations between the concepts and conditions for ascribing them. These axioms are used by reasoning programs as part of the process whereby the program decides what to do. The formalisms require too much explanation to be included in this article, but some of the criteria are easily given in English.

Beliefs and goals are ascribed in accordance with the the principle of rationality. Our object is to account for as much behavior as possible by saying the machine or person or animal does what it thinks will achieve its goals. It is especially important to have what is called in AI an *epistemologically adequate* system. Namely, the language must be able to express the information our program can actually get about a person's or machine's 'state of mind' — not just what might be obtainable if the neurophysiology of the human or the design of the machine were more accessible.

In general we cannot give definitions, because the concepts form a system that we fit as a whole to phenomena. Similarly the physicist doesn't give a definition of electron or quark. Electron and quark are terms in a complicated theory that predicts the results of experiments.

Indeed common sense psychology works in the same way. A child learns to ascribe wants and beliefs to others in a complex way that he never learns to encapsulate in definitions.

Nevertheless we can give approximate criteria for some specific properties relating them to the more implicit properties of believing and wanting.

Intends — We say that a machine intends to do something if we can regard it as believing that it will attempt to do it. We may know something that will deter it from making the attempt. Like most mental concepts, intention is an intermediate in the causal chain; an intention may be caused by a variety of stimuli and predispositions and may result in action or be frustrated in a variety of ways.

Tries — This is important in understanding machines that have a variety of ways of achieving a goal including possibly ways that we don't know about. If the machine may do something we don't know about but that can later be explained in relation to a goal, we have no choice but to use ‘is trying’ or some synonym to explain the behavior.

Likes — As in ‘A likes B’. This involves A wanting B’s welfare. It requires that A be sophisticated enough to have a concept of B’s welfare.

Self-consciousness — Self-consciousness is perhaps the most interesting mental quality to humans. Human self-consciousness involves at least the following:

1. Facts about the person’s body as a physical object. This permits reasoning from facts about bodies in general to one’s own. It also permits reasoning from facts about one’s own body, e.g. its momentum, to corresponding facts about other physical objects.

2. The ability to observe one’s own mental processes and to form beliefs and wants about them. A person can wish he were smarter or didn’t want a cigarette.

3. Facts about oneself as a having beliefs, wants, etc. among other similar beings.

Some of the above attributes of human self-consciousness are easy to program. For example, it is not hard to make a program look at itself, and many AI programs do look at parts of themselves. Others are more difficult. Also animals cannot be shown to have more than a few. Therefore, many present and future programs can best be described as partially self-conscious.

Suppose someone says, ‘The dog wants to go out’. He has ascribed the mental quality of wanting to the dog without claiming that the dog thinks like a human and can form out of its parts the thought, ‘I want to go out’.

The statement isn’t shorthand for something the dog did, because there are many ways of knowing that a dog wants to go out. It also isn’t shorthand for a specific prediction of what the dog is likely to do next. Nor do we know enough about the physiology of dogs for it to be an abbreviation for some statement about the dog’s nervous system. It is useful because of its connection with all of these things and because what it says about the dog corresponds in an informative way with similar statements about people. It doesn’t commit the person who said it to an elaborate view of the mind of a dog. For example, it doesn’t commit a person to any position about whether the dog has the mental machinery to know that it is a dog or even to know that it wants to go out. We can make similar statements about machines.

Here is an extract from the instructions that came with an electric blanket. *“Place the control near the bed in a place that is neither hotter nor colder than the room itself. If the control is placed on a radiator or radiant* *heated floors, it will 'think' the entire room is hot and will lower your blanket temperature, making your bed too cold. If the control is placed on the window sill in a cold draft, it will 'think' the entire room is cold and will heat up your bed so it will be too hot."*

I suppose some philosophers, psychologists, and English teachers would maintain that the blanket manufacturer is guilty of anthropomorphism and some will claim that great harm can come from thus ascribing to machines qualities which only humans can have. I argue that saying that the blanket thermostat 'thinks' is ok; they could even have left off the quotes. Moreover, this helps us understand how the thermostat works. The example is extreme, because most people don't need the word 'think' to understand how a thermostatic control works. Nevertheless, the blanket manufacturer was probably right in thinking that it would help some users.

Keep in mind that the thermostat can only be properly considered to have just three possible thoughts or beliefs. It may believe that the room is too hot, or that it is too cold, or that it is ok. It has no other beliefs; for example, it does not believe that it is a thermostat.

The example of the thermostat is a very simple one. If we had only thermostats to think about, we wouldn't bother with the concept of belief at all. And if all we wanted to think about were zero and one, we wouldn't bother with the concept of number.

Here's a somewhat fanciful example of a machine that might someday be encountered in daily life with more substantial mental qualities.

In ten or twenty years Minneapolis-Honeywell, which makes many thermostats today, may try to sell you a really fancy home temperature control system. It will know the preferences of temperature and humidity of each member of the family and can detect who is in the room. When several are in the room it makes what it considers a compromise adjustment taking account who has most recently had to suffer having the room climate different from what he prefers. Perhaps Honeywell discovers that these compromises should be modified according to a social rank formula devised by its psychologists and determined by patterns of speech loudness. The brochure describing how the thing works is rather lengthy and the real dope is in a rather technical appendix in small print.

Now imagine that I went on about this thermostat until you were bored and you skipped the rest of the paragraph. Confronted with an uncomfortable room you form any of the following hypotheses depending on what other information you had.

1. It's trying to do the right thing, but it can't because the valve is stuck. But then it should complain.

2. It regards Grandpa as more important than me, and it is keeping the room hot in case he comes in.

3. It confuses me with Grandpa.

4. It has forgotten what climate I like.

A child unable to read the appendix to the user's manual will be able to understand a description of the 'climate controller' in mental terms. The child will be able to request changes like *'Tell it I like it hotter'* or *'Tell it Grandpa's not here now'*. Indeed the designer of the system will have used the mental terms in formulating the design specifications.

The automatic teller is another example. It has beliefs like, 'There's enough money in the account,' and 'I don't give out that much money'. A more elaborate automatic teller that handles loans, loan payments, traveler's checks, and so forth, may have beliefs like, 'The payment wasn't made on time,' or, 'This person is a good credit risk.'

The next example is adapted from the University of California philosopher John Searle. A person who doesn't know Chinese memorizes a book of rules for manipulating Chinese characters. The rules tell him how to extract certain parts of a sequence of characters, how to re-arrange them, and how finally to send back another sequence of characters. These rules say nothing about the meaning of the characters, just how to compute with them. He is repeatedly given Chinese sentences, to which he applies the rules, and gives back what turn out, because of the clever rules, to be Chinese sentences that are appropriate replies. We suppose that the rules result in a Chinese conversation so intelligent that the person giving and receiving the sentences can't tell him from an intelligent Chinese. This is analogous to a computer, which only obeys its programming language, but can be programmed such that one can communicate with it in a different programming language, or in English. Searle says that since the person in the example doesn't understand Chinese — even though he can produce intelligent Chinese conversation by following rules — a computer cannot be said to 'understand' things. He makes no distinction, however, between the hardware (the person) and the process (the set of rules). I would argue that the set of rules understands Chinese, and, analogously, a computer program may be said to understand things, even if the computer does not. Both Searle and I are ignoring practical difficulties like how long it would take a person with a rule book to come up with a reply.

Daniel Dennett, Tufts University philosopher, has proposed three attitudes aimed at understanding a system with which one interacts.

The first he calls the physical stance. In this we look at the system in terms of its physical structure at various levels of organization. Its parts have their properties and they interact in ways that we know about. In principle the analysis can go down to the atoms and their parts. Looking at a thermostat from this point of view, we'd want to understand the working of the bimetal strip that most thermostats use. For the automatic teller, we'd want to know about integrated circuitry, for one thing. (Let's hope no one's in line behind us while we do this).

The second is called the design stance. In this we analyze something in terms of the purpose for which it is designed. Dennett's example of this is the alarm clock. We can usually figure out what an alarm clock will do, e.g. when it will go off, without knowing whether it is made of springs and gears or of inegrated circuits. The user of alarm clock typically doesn't know or care much about its internal structure, and this information wouldn't be of much use. Notice that when an alarm clock breaks, its repair requires taking the physical stance. The design stance can usefully be applied to a thermostat — it shouldn't be too hard to figure out how to set it, no matter how it works. With the automatic teller, things are a little less clear.

The design stance is appropriate not only for machinery but also for the parts of an organism. It is amusing that we can't attribute a purpose for the existence of ants, but we can find a purpose for the glands in an ant that emit a chemical substance for other ants to follow.

The third is called the intentional stance, and this is what we'll often need for understanding computer programs. In this we try to understand the behavior of a system by ascribing to it beliefs, goals, intentions, likes and dislikes, and other mental qualities. In this stance we ask ourselves what the thermostat thinks is going on, what the automatic teller wants from us before it'll give us cash. We say things like, *'The store's billing computer wants me to pay up, so it intends to frighten me by sending me threatening letters'*. The intentional stance is most useful when it is the only way of expressing what we know about a system.

(For variety Dennett mentions the astrological stance. In this the way to think about the future of a human is to pay attention to the configuration of the stars when he was born. To determine whether an enterprise will succeed we determine whether the signs are favorable. This stance is clearly distinct from the others — and worthless.)

It is easiest to understand the ascription of thoughts to machines in circumstances when we also understand the machine in physical terms. However, the payoff comes when either no-one or only an expert understands the machine physically.

However, we must be careful not to ascribe properties to a machine that the particular machine doesn't have. We humans can easily fool ourselves when there is something we want to believe.

The mental qualities of present machines are not the same as ours. While we will probably be able, in the future, to make machines with mental qualities more like our own, we'll probably never want to deal with machines that are too much like us. Who wants to deal with a computer that loses its temper, or an automatic teller that falls in love? Computers will end up with the psychology that is convenient to their designers — (and they'll be fascist bastards if those designers don't think twice). Program designers have a tendency to think of the users as idiots who need to be controlled. They should rather think of their program as a servant, whose master, the user, should be able to control it. If designers and programmers think about the apparent mental qualities that their programs will have, they'll create programs that are easier and pleasanter — more humane — to deal with.

References

Dennett, Daniel (1981).True Believers: the Intentional Strategy and Why it Works, *The Herbert Spencer Lectures*, A. Heath (ed.), Oxford University Press. This non-technical article describes the physical, design and intentional stances in philosophical language.

Kowalski, Robert (1979). *Logic for Problem Solving*, New York: North Holland. This book describes the use of logical formalism in artificial intelligence.

McCarthy, John (1979). Ascribing Mental Qualities to Machines[1] in *Philosophical Perspectives in Artificial Intelligence*, Ringle, Martin (ed.), Humanities Press. This is the technical paper on which this article is based.

McCarthy, John (1979). First Order Theories of Individual Concepts and Propositions,[2] in Michie, Donald (ed.) *Machine Intelligence 9*, Ellis Horwood. (Reprinted in this volume, pp. 000–000.) This paper uses the mathe-

[1]http://www-formal.stanford.edu/jmc/ascribing.html

[2]http://www-formal.stanford.edu/jmc/concepts.html

matical formalism of first order logic to express facts about knowledge.

Newell, Allen (1982). The Knowledge Level, *Artificial Intelligence*, Vol. 18 No.1, pp. 87–127. This article clearly expounds a different approach to ascribing mental qualities.

Searle, John (1980). Minds, Brains and Programs, *Behavioral and Brain Sciences*, Vol.3 No. 3, pp. 417–424. This article takes the point of view that mental qualities should not be ascribed to machines.

/@steam.stanford.edu:/u/ftp/jmc/little.tex: begun 1996 May 14, latexed 1996 May 14 at 1:40 a.m.