Philosophical Premises of Logical AI

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

1998 Nov 20, 9:55 a.m.

Abstract

1 Philosophical presuppositions of logical AI

> Extinguished theologians lie about the cradle of every science as the strangled snakes beside that of Hercules.
> —T. H. Huxley

Q. Why bother stating philosophical presuppositions? Why not just get on with the AI?

A. AI shares many concerns with philosophy—with metaphysics, epistemology, philosophy of mind and other branches of philosophy. This is because AI concerns the creation of an artificial mind. However, AI has to treat these questions in more detail.

AI research not based on stated philosophical presuppositions usually turns out to be based on unstated philosophical presuppositions. These are often so wrong as to interfere with developing intelligent systems.

That it should be possible to make machines as intelligent as humans involves some philosophical premises, although the possibility is probably

accepted by a majority of philosophers. The way this book proposes to build intelligent machines makes more presumptions, some of which may be new.

This section concentrates on stating the premises without much argument. Chapter **??** presents arguments and discusses other opinions. We give some references to sections in the book formalizing some of the philosophical notions where this is appropriate.

objective world The world exists independently of humans. The facts of mathematics and physical science are independent of there being people to know them. Intelligent Martians and robots will need to know the same facts. A robot also needs to believe that the world exists independently of itself. Science tells us that humans evolved in a world which formerly did not contain humans. Given this, it is odd to regard the world as a human construct. It would be even more odd to program a robot to regard the world as its own construct. The problem doesn't arise for the limited robots of today, because the languages they are programmed to use can't express assertions about the world in general. This limits what they can learn or can be told.

correspondence theory of truth and reference A logical robot represents what it *believes* about the world by logical sentences. Some of these beliefs we build in; others come from its observations and still others by induction from its experience. Within the sentences it uses *terms* to refer to objects in the world.

In every case, we try to design it so that what it will believe about the world is as accurate as possible. Debugging and improving the robot includes detecting false beliefs about the world and changing the way it acquires information to maximize the correspondence between what it believes and the facts of world. The terms the robot uses to refer to objects need to correspond to the objects so that the sentences will express facts about the objects.

Already this involves a philosophical presupposition—that which is called the *correspondence theory of truth*. AI also needs a *correspondence theory of reference*.

As with science, a robot's theories are tested are experimentally, but the concepts robots use are not defined in terms of experiments. They are axiomatized, and some axioms relate the terms to observations.

The important consequence of the correspondence theory is that when we design robots, we need to keep in mind the relation between *appearance*, the information coming through the robot's sensors, and *reality*. Only in certain simple cases, e.g. the position in a chess game, does the robot have sufficient access to reality for this distinction to be ignored.

Some robots react directly to their inputs without memory or inferences. It is our scientific (i.e. not philosophical) contention that these are inadequate for human-level intelligence, because the world contains too many important entities that cannot be observed directly.

A robot that reasons about the acquisition of information must itself be aware of these relations. Some formalizations of these relations are given in Section **??**.

The correspondence theory of truth may be contrasted with *pragmatic theories of truth* in which beliefs are regarded as true if they result in success in achieving goals. Each kind of theory has adherents among philosophers. Roughly speaking, pragmatic theories of truth correspond to making *reactive robots* that respond directly to inputs. Some behaviors can be programmed this way, but logical robots are appropriately designed to *do what they think will advance their goals.*

science Science is substantially correct in what it tells us about the world, and scientific activity is the best way to obtain more knowledge. 20th century corrections to scientific knowledge mostly left the old theories as good approximations to reality.

mind and brain The human mind is an activity of the human brain. This is a scientific proposition, supported by all the evidence science has discovered so far.

common sense Common sense ways of perceiving the world and common opinion are also substantially correct. When common sense errs, it can be corrected by science, and the results of the correction often become part of common sense if they are not too mathematical. Thus common sense has absorbed the notion of inertia. However, its mathematical generalization, the law of conservation of momentum has made its way into the common sense of only a small fraction of people—even among the people who have taken courses in physics.

science embedded in common sense Science is embedded in common sense. Galileo taught us that the distance s that a dropped body falls in time t is given by the formula

$$s = \frac{1}{2}gt^2.$$

To use this information, the English (or its logical equivalent) is just as essential as the formula, and common sense knowledge of the world is required to make the measurements required to use or verify the formula.

possibility of AI According to some philosophers' views, artificial intelligence is either a contradiction in terms [Sea84] or intrinsically impossible [Dre92] or [Pen94]. See Chapter **??** for some polemics.

mental qualities treated individually AI has to treat mind in terms of components rather than regarding mind as a unit that necessarily has all the mental features that occur in humans. Thus we design some very simple systems in terms of the beliefs we want them to have and debug them by identifying erroneous beliefs. Some philosophers reject this.

third person point of view We ask "How does it (or he) know?", "What does it perceive?" rather than how do I know and what do I perceive.

rich ontology Our theories involve many kinds of entity—material objects, situations, properties as objects, contexts, concepts. When one kind A of entity might be defined in terms of others, we will often prefer to treat A separately, because we may later want to change our ideas of its relation to other entities.

We often consider several related concepts, where others have tried to get by with one. Suppose a man sees a dog. Is seeing a relation between the man and the dog or a relation between the man and an appearance of a dog? Some purport to refute calling seeing a relation between the man and the dog by pointing out that the man may actually see a hologram or picture of the dog.

rich entities The entities the robot must refer to often are *rich* with properties the robot cannot know all about. The best example is a *natural*

kind like a lemon. A child buying a lemon at a store knows enough properties of the lemons that occur in the stores he frequents to distinguish lemons from other fruits in the store. Experts know more properties of lemons, but no-one knows all of them.

approximate entities Many of the philosophical arguments purporting to show that naive common sense is hopelessly mistaken are wrong. These arguments often stem from trying to force intrinsically approximate concepts into the form of if-and-only-if definitions. This point will be discussed more fully in Section **??** about *approximate objects*.

The emphasis on the first class character of approximate entities may be new. It means that we can quantify over approximate entities and also express how an entity is approximate.

compatibility of determinism and free will A logical robot needs to consider its choices and the consequences of them. Therefore, it must regard itself as having *free will* even though it is a deterministic device. We discuss our choices and those of robots by considering nondeterminist approximations to a determinist world—or at least a world more determinist than is needed in the approximation. The formalism is discussed in section **??**. The philosophical name for this view is *compatibilism*. I think compatibilism is a requisite for AI research reaching human-level intelligence.

mind-body distinctions I'm not sure whether this point is philosophical or scientific. The mind corresponds to software, perhaps with an internal distinction between program and knowledge. Software won't do anything without hardware, but the hardware can be quite simple. Some hardware configurations can run many different programs concurrently, i.e. there can be many minds in the same computer body. Software can also interpret other software.

Confusion about this is the basis of the Searle Chinese room fallacy. The man in the hypothetical Chinese room is interpreting the software of a Chinese personality. Interpreting a program does not require having the knowledge possessed by that program. This would be obvious if people could interpret other personalities at a practical speed, but Chinese room software interpreted by an unaided human might run at 10^{-9} the speed of an actual Chinese.

2 Scientific premises of logical AI

Some of the premises of logical AI are scientific in the sense that they are subject to scientific verification. This may also be true of some of the premises listed above as philosophical.

middle out Humans deal with middle-sized objects and develop our knowledge up and down from the middle. Formal theories of the world must also start from the middle where our experience informs us. Efforts to start from the most basic concepts, e.g. to make a basic ontology are unlikely to succeed as well as starting in the middle. The ontology must be compatible with the idea that the basic entities in the ontology are not the basic entities in the world. More basic entities are known less well than the middle entities.

universality of intelligence Achieving goals in the world requires that an agent with limited knowledge, computational ability and ability to observe use certain methods. This is independent of whether the agent is human, Martian or machine. For example, playing chess-like games effectively efficiently requires something like alpha-beta pruning. Perhaps this should be regarded as a scientific opinion (or bet) rather than as philosophical.

sufficient complexity yields essentially unique interpretations A robot that interacts with the world in a sufficiently complex way gives rise to an essentially unique interpretation of the part of the world with which it interacts. This is an empirical, scientific proposition, but many people, especially philosophers (see Quinexx, Putnamxx, [**?**], [Den98]), take its negation for granted. There are often many interpretations in the world of short descriptions, but long descriptions almost always admit at most one.

The most straightforward example is that a simple substitution cipher cryptogram of an English sentence usually has multiple interpretations if the text is less than 21 letters and usually has a unique interpretation if the text is longer than 21 letters. Why 21? It's a measure of the redundancy of English. The redundancy of a person's or a robot's interaction with the world is just as real—though clearly much harder to quantify.

References

[Den98] Daniel Dennett. *Brainchildren: Essays on Designing Minds.* MIT Press, 1998.

[Dre92] Hubert Dreyfus. *What Computers still can't Do.* M.I.T. Press, 1992.

[Pen94] Roger Penrose. *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford University Press, Oxford, 1994.

[Sea84] John R. Searle. *Minds, Brains, and Science.* Harvard University Press, Cambridge, Mass., 1984.

/@steam.stanford.edu:/u/ftp/jmc/philosophical.tex: begun Wed Sep 16 12:51:22 1998, latexed November 20, 1998 at 9:55 a.m.