What has AI in Common with Philosophy?

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

August 28, 1995

Abstract

AI needs many ideas that have hitherto been studied only by philosophers. This is because a robot, if it is to have human level intelligence and ability to learn from its experience, needs a general world view in which to organize facts. It turns out that many philosophical problems take new forms when thought about in terms of how to design a robot. Some approaches to philosophy are helpful and others are not.

1 Introduction

Artificial intelligence and philosophy have more in common than a science usually has with the philosophy of that science. This is because human level artificial intelligence requires equipping a computer program with some philosophical attitudes, especially epistemological.

The program must have built into it a concept of what knowledge is and how it is obtained.

If the program is to reason about what it can and cannot do, its designers will need an attitude to free will. If it is to do meta-level reasoning about what it can do, it needs an attitude *of its own* to free will.

If the program is to be protected from performing unethical actions, its designers will have to build in an attitude about that.

Unfortunately, in none of these areas is there any philosophical attitude or system sufficiently well defined to provide the basis of a usable computer program.

Most AI work today does not require any philosophy, because the system being developed doesn't have to operate independently in the world and have a view of the world. The designer of the program does the philosophy in advance and builds a restricted representation into the program.

Building a chess program requires no philosophy, and Mycin recommended treatments for bacterial infections without even having a notion of processes taking place in time. However, the performance of Mycin-like programs and chess programs is limited by their lack of common sense and philosophy, and many applications will require a lot. For example, robots that do what they think their owners want will have to reason about wants.

Not all philosophical positions are compatible with what has to be built into intelligent programs. Here are some of the philosophical attitudes that seem to me to be required.

1. Science and common sense knowledge of the world must both be accepted. There are atoms, and there are chairs. We can learn features of the world at the intermediate size level on which humans operate without having to understand fundamental physics. Causal relations must also be used for a robot to reason about the consequences of its possible actions.

2. Mind has to be understood a feature at a time. There are systems with only a few beliefs and no belief that they have beliefs. Other systems will do extensive introspection. Contrast this with the attitude that unless a system has a whole raft of features it isn't a mind and therefore it can't have beliefs.

3. Beliefs and intentions are objects that can be formally described.

4. A sufficient reason to ascribe a mental quality is that it accounts for behavior to a sufficient degree.

5. It is legitimate to use approximate concepts not capable of **iff** definition. For this it is necessary to relax some of the criteria for a concept to be meaningful. It is still possible to use mathematical logic to express approximate concepts.

6. Because a theory of approximate concepts and approximate theories is not available, philosophical attempts to be precise have often led to useless hair splitting.

7. Free will and determinism are compatible. The deterministic process that determines what an agent will do involves its evaluation of the consequences of the available choices. These

choices are present in its consciousness and can give rise to sentences about them as they are observed.

8. Self-consciousness consists in putting sentences about consciousness in memory.

9. Twentieth century philosophers became to critical of reification. Many of the criticism don't apply when the entities reified are treated as approximate concepts.

2 The Philosophy of Artificial Intelligence

One can expect there to be an academic subject called the philosophy of artificial intelligence analogous to the existing fields of philosophy of physics and philosophy of biology. By analogy it will be a philosophical study of the research methods of AI and will propose to clarify philosophical problems raised. I suppose it will take up the methodological issues raised by Hubert Dreyfus and John Searle, even the idea that intelligence requires that the system be made of meat.

Presumably some philosophers of AI will do battle with the idea that AI is impossible (Dreyfus), that it is immoral (Weizenbaum) and that the very concept is incoherent (Searle).

It is unlikely to have any more effect on the practice of AI research than philosophy of science generally has on the practice of science.

3 Epistemological Adequacy

Formalisms for representing facts about the world have to be adequate for representing the information actually available. A formalism that represented the state of the world by the positions and velocities of molecules is inadequate if the system can't observe positions and velocities, although such a formalism may be the best for deriving thermodynamic laws.

The common sense world needs a language to describe objects, their relations and their changes quite different from that used in physics and engineering. The key difference is that the information is less complete. It needs to express what is actually known that can permit a robot to determine the expected consequences of the actions it contemplates.

4 Free Will

An attitude toward the free will problem needs to be built into robots in which the robot can regard itself as having choices to make, i.e. as having free will.

5 Natural Kinds

Natural kinds are described rather than defined. We have learned about lemons and experienced them as small, yellow fruit. However, this knowledge does not permit an **iff** definition. Lemons differ from other fruit in ways we don't yet know about. There is no continuous gradation from lemons to oranges. On the other hand, geneticists could manage to breed large blue lemons by tinkering with the genes, and there might be good reasons to call the resulting fruit lemons.

6 Four Stances

Daniel Dennett named three *stances* one can take towards an object or system. The first is the *physical stance* in which the physical structure of the system is treated. The second is the *intentional stance* in which the system is understood in terms of its beliefs, goals and intentions. The third is the *design stance* in which the system is understood in terms of its composition out of parts. One more stance we'll call the *functional stance*. We take the functional stance toward an object when we ask what it does without regard to its physics or composition. The example I like to give is a motel alarm clock. The user may not notice whether it is mechanical, an electric motor timed by the power line or electronic timed by a quartz crystal.[1] Each stance is appropriate in certain conditions.

7 Ontology and Reification

Quine wrote that one's ontology coincides with the ranges of the variables in one's formalism. This usage is entirely appropriate for AI. Present philosophers, Quine perhaps included, are often too stingy in the reifications they permit. It is sometimes necessary to quantify over beliefs, hopes and goals.

When programs interact with people or other programs they often perform *speech acts* in the sense studied by Austin and Searle.

[1] I had called this the design stance, and I thank Aaron Sloman for pointing out my mistake and suggesting *functional stance*.

Quantification over promises, obligations, questions, answers to questions, offers, acceptances and declinations are required.

8 Counterfactuals

An intelligent program will have to use counterfactual conditional sentences, but AI needs to concentrate on useful counterfactuals. An example is *"If another car had come over the hill when you passed just now, there would have been a head-on collision."* Believing this counterfactual might change one's driving habits, whereas the corresponding material conditional, obviously true in view of the false antecedent, could have no such effect. Counterfactuals permit systems to learn from experiences they don't actually have.

Unfortunately, the Stalnaker-Lewis closest possible world model of counterfactuals doesn't seem helpful in building programs that can formulate and use them.

9 Philosophical Pitfalls

There is one philosophical view that is attractive to people doing AI but which limits what can be accomplished. This is logical positivism which tempts AI people to make systems that describe the world in terms of relations between the program's motor actions and its subsequent observations. Particular situations are sometimes simple enough to admit such relations, but a system that only uses them will not even be able to represent facts about simple physical objects. It cannot have the capability of a two week old baby.

10 Philosophers! Help!

Previous philosophical discussion of certain conecpts has been helpful to AI. In this I include the Austin-Searle discussion of speech acts, Grice's discussion of conversational implicatures, various discussions of natural kinds, modal logic and the notion of philosophy as a science. Maybe some of the philosophical discussions of causality and counterfactuals will be useful for AI. In this paragraph I have chosen to be stingy with credit.

Philosophers could help artificial intelligence more than they have done if they would put some attention to some more detailed conceptual problems such as the following:

belief What belief statements are useful?

how What is the relation between naming an occurrence and its suboccurrences? *He went to Boston. How? He drove to the airport, parked and took UA 34.*

responsiveness When is the answer to a question responsive? Thus *"Vladimir's wife's husband's telephone number"* is a true but not responsive answer to a request for Vladimir's telephone number.

useful causality What causal statements are useful?

useful counterfactuals What counterfactuals are useful and why? *"If another car had come over the hill when you passed, there would have been a head-on collision."*

References

There is not space in this article nor have I had the time to prepare a proper bibliography. Such a bibliography would refer to a number of papers, some of mine being reprinted in my *Formalizing Common Sense* Many are available via my Web page http://www-formal.stanford.edu/jmc/. I would also refer to work by the following philosophers: Rudolf Carnap, Daniel Dennett, W. V. O. Quine, Hilary Putnam, Paul Grice, John Searle, Robert Stalnaker, David Lewis, Aaron Sloman, Richard von Mises. Much of the bibliography in Aaron Sloman's previous article is applicable to this one.

Acknowledgement: Work partly supported by ARPA (ONR) grant N00014-94-1-0775.