Contents

1 PHILOSOPHICAL AND SCIENTIFIC PRESUPPOSITIONS OF LOGICAL AI 3

1.1 Philosophical Presuppositions 3

1.2 Scientific Presuppositions 10

Chapter 1

PHILOSOPHICAL AND SCIENTIFIC PRESUPPOSITIONS OF LOGICAL AI

> Extinguished theologians lie about the cradle of every science as the strangled snakes beside that of Hercules.
> —T. H. Huxley (Darwin's bulldog)[1]

Abstract: Many ideas from philosophy, especially from recent analytic philosophy, are usable for AI. However, some philosophical points of view make assumptions that have the effect of excluding the possibility of AI. Likewise work on AI is not neutral with regard to philosophical issues. This chapter presents what we consider the presuppositions of logical AI and also some scientific presuppositions, i.e. some results of science that are relevant. We emphasize the relation to AI rather than philosophy itself.

1.1 Philosophical Presuppositions

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

[1] Progress in AI may extinguish some philosophies, but don't stand on one foot.

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 than philosophers customarily consider relevant.[2]

In principle, an evolutionary approach need not involve philosophical presuppositions. However, many putative evolutionary approaches are crippled by impoverished philosophical assumptions. For example, the systems often only admit patterns in appearance and can't even represent reality behind appearance. (McCarthy) presents a challenge to learning systems to learn reality behind appearance.

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 we propose to build intelligent machines, i.e. via logical AI, makes more presuppositions, some of which may be new.

This chapter concentrates on stating the presuppositions and their relations to AI without much philosophical argument. A later chapter presents arguments and discusses other opinions.

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 is even more odd to program a robot to regard the world as its own construct. What the robot believes about the world in general 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—and hence what we can get them to do for us.

[2] Compare the treatment of counterfactual conditional sentences in CostelloMcC99 with that in (Lewis 1973).

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, though not usually as detailed 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 entities need to correspond to the entities so that the sentences will express facts about these entities. We have in mind both material objects and other entities, e.g. plans.

Already this involves a philosophical presupposition—that which is called the *correspondence theory of truth*. AI also needs a *correspondence theory of reference* , i.e. that a mental structure can refer to an external object and can be judged by the accuracy of the reference.

As with science, a robot's theories are tested experimentally, but the concepts robots use are often not defined in terms of experiments. Their properties are partially axiomatized, and some axioms relate 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. In order that a robot should not always believe what it sees with its own eyes, it must distinguish between appearance and reality. (McCarthy) presents a challenge problem requiring the discovery of reality behind appearance.

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 scientific theories as good approximations to reality. Much "postmodern philosophy" is in opposition to this.

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. However, the dualist intuition of separation between mind and body is related to the sometimes weak connections between thought and action. Dualism has some use as a psychological abstraction.

common sense Common sense ways of perceiving the world and common opinion are also substantially correct. When general common sense errs, it can often be corrected by science, and the results of the correction may 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.

From Socrates on philosophers have found many inadequacies in common sense usage, e.g. common sense notions of the meanings of words. The corrections are often elaborations, making distinctions blurred in common sense usage. Unfortunately, there is no end to philosophical elaboration, and the theories become very complex. However, some of the elaborations seem essential to avoid confusion in some circumstances. Here's a candidate for the way out of the maze.

Robots will need both the simplest common sense usages and to be able to tolerate elaborations when required. For this we have proposed two notions—contexts as formal objects (McCarthy 1993) and (McCarthy and Buvač 1997) and *elaboration tolerance* (McCarthy 1999)[3]

[3] Hilary Putnam (Putnam 1975) discusses two notions concerning meaning proposed by previous philosophers which he finds inadequate. These are

> (I) That knowing the meaning of a term is just a matter of being in a certain "psychological state" (in the sense of "psychological state" in which states of memory and psychological dispositions are "psychological states"; no

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 (Searle 1984) or intrinsically impossible (Dreyfus 1992) or (Penrose 1994). The methodological basis of these arguments has to be wrong and not just the arguments themselves. We hope to deal with this elsewhere.

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. (McCarthy 1979) treats this. Ascribing a few beliefs to thermostats has led to controversy.

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.

> one thought that knowing the meaning of a word was a continuous state of consciousness, of course.)
>
> (II) That the meaning of a term (in the sense of "intension") determines its extension (in the sense that sameness of intension entails sameness of extension).

Suppose Putnam is right in his criticism of the general correctness of (I) and (II). His own ideas are more elaborate.

It may be convenient for a robot to work mostly in contexts within a larger context C_{phil1} in which (I) and (II) (or something even simpler) hold. However, the same robot, if it is to have human level intelligence, must be able to *transcend* C_{phil1} when it has to work in contexts to which Putnam's criticisms of the assumptions of C_{phil1} apply.

It is interesting, but perhaps not necessary for AI at first, to characterize those contexts in which (I) and (II) are correct.

This presupposes the correspondence theory of truth. It applies to how we look at robots, but also to how we want robots to reason about the knowledge of people and other robots. Some philosophers, e.g. John Searle, insist with Descartes on a first person point of view.

rich ontology Our theories involve many kinds of entity—material objects, situations, properties as objects, contexts, propositions, indivdual concepts, wishes, intentions. 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. AI needs the relation between the man and the appearance of a dog, the relation between the man and the dog and also the relation between dogs and appearances of them. None is most fundamental.

natural kinds 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. AI systems also have to distinguish between sets of properties that suffice to recognize an object in particular situations and the natural kinds of some objects.

To a child, all kinds are natural kinds, i.e. kinds about which the child is ready to learn more. The idea of a concept having an if-and-only-if definition comes later—perhaps at ages 10–13. Taking that further, *natural kind* seems to be a context relative notion. Thus some part of income tax law is a natural kind to me, whereas it might have an if-and-only-if definition to an expert.

Curiously enough, many of the notions studied in philosophy are not natural kinds, e.g. proposition, meaning, necessity. When they are regarded as natural kinds, then fruitless arguments about what they really are take place. AI needs these concepts but must be able to work with limited notions 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.

Our 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. An article on approximate theories and approximate entities is forthcoming.

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 non-determinist approximations to a determinist world—or at least a world more determinist than is needed in the approximation. The philosophical name for this view is *compatibilism*. I think compatibilism is a requisite for AI research reaching human-level intelligence.

In practice, regarding an observed system as having choices is necessary when ever a human or robot knows more about the relation of the system to the environment than about what goes on within the system. This is discussed in (McCarthy 1996).

mind-brain 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 (Searle 1984). 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.

If one settles for a Chinese conversation on the level of Eliza (Weizenbaum 1965), then, according to Weizenbaum (1999 personal communication), the program can be hand simulated with reasonable performance.

1.2 Scientific Presuppositions

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.

innate knowledge The human brain has important innate knowledge, e.g. that the world includes three dimensional objects that usually persist even when not observed. This was learned by evolution. Acquiring such knowledge by learning from sense data will be quite hard. It is better to build it into AI systems.

Different animals have different innate knowledge. Dogs know about permanent objects and will look for them when they are hidden. Very likely, cockroaches don't know about objects.

Identifying human innate knowledge has been the subject of recent psychological research. See (Spelke 1994) and the discussion in (Pinker 1997) and the references Pinker gives. In particular, babies and dogs know innately that there are permanent objects and look for them when they go out of sight. We'd better build that in.

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 requires something like alpha-beta pruning. Perhaps this should be regarded as a scientific opinion (or bet) rather than as philosophical.

universal expressiveness of logic This is a proposition analogous to the Turing thesis that Turing machines are computationally universal—anything that can be computed by any machine can be computed by a Turing machine. The *expressiveness thesis* is that anything that can be expressed, can be expressed in first order logic. Some elaboration of the idea is required before it will be as clear as the Turing thesis.[4]

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 (Quine 1969), (Putnam 1975), (Dennett 1971), (Dennett 1998)), 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.

We expect these philosophical and scientific presuppositions to become more important as AI begins to tackle human level intelligence.

[4] First order logic isn't the best way of expressing all that can be expressed any more than Turing machines are the best way of expressing computations. However, with set theory, what can be expressed in stronger systems can apparently also be expressed in first order logic.

Bibliography

Dennett, D. 1998. *Brainchildren: Essays on Designing Minds.* MIT Press.

Dennett, D. C. 1971. Intentional systems. *The Journal of Philosophy* 68(4):87–106.

Dreyfus, H. 1992. *What Computers still can't Do.* M.I.T. Press.

Lewis, D. 1973. *Counterfactuals.* Harvard University Press.

McCarthy, J. n.d. **appearance and reality**[5]. *web only for now, and perhaps for the future.* not publishable on paper, because it contains an essential imbedded applet.

McCarthy, J. 1979. Ascribing mental qualities to machines[6]. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence.* Harvester Press. Reprinted in (McCarthy 1990).

McCarthy, J. 1990. *Formalizing Common Sense: Papers by John McCarthy.* 355 Chestnut Street, Norwood, NJ 07648: Ablex Publishing Corporation.

McCarthy, J. 1993. Notes on Formalizing Context[7]. In *IJCAI-93.*

McCarthy, J. 1996. Making Robots Conscious of their Mental States[8]. In S. Muggleton (Ed.), *Machine Intelligence 15.* Oxford University Press. to appear in 1999.

McCarthy, J. 1999. Elaboration tolerance[9]. *to appear.*

[5] http://www-formal.stanford.edu/jmc/appearance.html

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

[7] http://www-formal.stanford.edu/jmc/context.html

[8] http://www-formal.stanford.edu/jmc/consciousness.html

[9] http://www-formal.stanford.edu/jmc/elaboration.html

McCarthy, J., and S. Buvač. 1997. Formalizing context (expanded notes). In A. Aliseda, R. v. Glabbeek, and D. Westerståhl (Eds.), *Computing Natural Language*. Center for the Study of Language and Information, Stanford University.

Penrose, R. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Pinker, S. 1997. *How the Mind Works*. Norton.

Putnam, H. 1975. The meaning of "meaning". In K. Gunderson (Ed.), *Language, Mind and Knowledge*, Vol. VII of *Minnesota Studies in the Philosophy of Science*, 131–193. University of Minnesota Press.

Quine, W. V. O. 1969. Propositional objects. In *Ontological Relativity and other Essays*. Columbia University Press, New York.

Searle, J. R. 1984. *Minds, Brains, and Science*. Cambridge, Mass.: Harvard University Press.

Spelke, E. 1994. Initial knowlege: six suggestions. *Cognition* 50:431–445.

Weizenbaum, J. 1965. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery* 9(1):36–45.