In D. Michie (ed)
Expert Systems in the Microelectronic Age
(Edinburgh University Press, 1979)

Epistemology and Artificial Intelligence

Aaron Sloman
Cognitive Studies Programme
School of Social Sciences
University of Sussex, Brighton

(Now School of Computer Science, University of Birmingham)
http://www.cs.bham.ac.uk/~axs/

Introduction

For readers unfamiliar with epistemology a very brief overview of the subject may be useful prior to reading the papers by Hayes and McCarthy, linking epistemology and Artificial Intelligence.

Epistemology is the study of knowledge. There are many different ways the subject can be approached, including the following:

1. A study of the knowledge, beliefs and concepts of different societies, or different epochs. (Sociology, Anthropology, History of Ideas.)

2. A study of the cognitive development of human individuals. Piaget calls his work in this field 'Genetic Epistemology'.

3. A study of the evolution of different kinds of knowledge and the mechanisms for using them. I call this bio-epistemology. The subject is in its infancy and will remain so until biologists accept that not only animal behaviour but also the underlying mental structures and mechanisms are worthy of scientific study.

4. Philosophical studies of knowledge. These are often not so much concerned with finding out *facts* as with establishing *norms* or *standards.* So philosophers try to find criteria for deciding whether certain beliefs are *rational*, or whether inferences from evidence to conclusions are *valid.* A closely related topic is the study of the systems of concepts required for formulating beliefs of different kinds, for instance concepts of space, time, matter, cause, and mind. As we shall see, philosophical studies of concepts and knowledge are closely related to work in Artificial Intelligence.

5. Work in Artificial Intelligence, whether aimed at modelling human minds or designing smart machines, necessarily includes a study of knowledge. Knowledge about particular domains is the basis of the expertise of all the expert systems described in this volume.

General knowledge about how knowledge is acquired represented and used, has to be embodied in flexible systems which can be extended, or which can explain their actions. A machine which communicates effectively with a variety of humans will have to use information about what people can be expected to know in various circumstances. This is especially true of a teaching machine.

Problems about Knowledge

The study of knowledge, in all the aforementioned fields can be shallow or deep. A shallow study is concerned simply with what is known by some system - a person, community, or machine capable of performing certain tasks. Deeper investigations are concerned with the *basis* of such knowledge. We can identify a series of questions of increasing depth and difficulty, about some knowledgeable system X, actual or hypothesised.

1. What are the things which X knows, or believes? What does X need to know in order to perform some task?

2. What system of *concepts* is used in such knowledge? (The system of concepts will usually define a more general class of possible facts, rules, etc, than that actually grasped by X. See my 1978-chapter 2.)

3. What formalism, or set of formalisms is used for expressing such concepts and combining them into beliefs, rules, intentions, etc.? What inference rules does the formalism use? (In general a formalism e.g. Predicate calculus, or the notation of maps, or a programming language, will support a more general class of concepts than those actually used by X).

4. What are the mechanisms by which the formalism is implemented and used? (The underlying mechanism may be capable of supporting a wider range of formalisms - symbolisms, notations - than those actually used by X.)

For each of these questions there is an additional dimension in which new questions of increased depth and difficulty can be asked. We can ask about a class of beliefs, or concepts, or notations, or mechanisms, what the alternatives are, and how one can choose rationally between alternatives. E.g. why is it better to use this information than that? Why are these concepts more useful than some others? Why is this a good formalism to use? This subsumes much philosophical discussion of questions like "Is it rational to believe there is an external world?" "Is it rational to accept predictions based on extrapolation from past experience?" "Is it rational to use concepts not definable in terms of the contents of experiences?"

These questions are all clearly relevant to the design of working expert systems, but the depth of study is related to the ambitions of such a design. For instance if a system is merely intended to be a data-base of information on some topic, able to answer questions about what has been stored explicitly, then question (1) is all that need be considered, and the question ‘Why store these facts rather than some others?’ need not arise. The set of items to be stored defines the task, unlike cases where the task is defined by some other goal, such as accurate diagnosis of a range of diseases.

If the system is to be able to check for consistency, or to make inferences to conclusions not explicitly stored, then the designers will have to go into questions about how the concepts used are related, what general patterns of inference are valid, methods for weighting uncertain evidence or merely probable conclusions, etc.

If the system is to be able to store many different kinds of knowledge, and to exercise some independence in choosing which subset is relevant to a given problem, then issues about generality of the formalisms, trade-offs between different sorts of formalisms and underlying mechanisms, etc become paramount.

Finally, if the system is to be able to explain its decisions to people, then the form in which facts and inference rules are stored must relate closely to what people can understand. This may not always be what is computationally most economical.

We can illustrate the importance of question 3 above, not only for philosophical and AI studies of knowledge, but for all science, by noting that until recently there were few good ways of expressing theories about complex processes. The formalisms available were mainly derived from the concepts of physics, and were therefore mostly concerned with changes in scalar (discrete or continuous) variables. But although many social scientists and psychologists continue to use such descriptive methods, it is now clear that they are grossly inadequate for the description and analysis of processes involving the manipulation of complex symbolic structures. No computer scientist would dream of trying to describe the behaviour of a compiler or operating system in terms of a few equations relating a handful of measurable quantities. More complex ideas, such as are embodied in programming languages (concepts like conditional instructions, recursion, data-structures, networks, etc.) are required for adequate descriptions. And yet many scientists studying human beings and social systems continue to try to force their theories into the quantitative framework derived from physics.

A favourite area of disagreement among AI researchers concerns the selection of a formalism for specifying computational mechanisms. Many of the papers in this book report systems based on the idea of a set of production-rules, namely rules with a condition and an action. The underlying mechanism is, in the simplest cases, a machine which repeatedly looks for a rule whose *condition* matches some data-base, and then obeys the *action.* By contrast the papers by McCarthy and Hayes tend to favour explicit use of predicate calculus as a formalism for expressing knowledge, where the contents of rules will normally be expressed in some universally quantified assertion. Detailed comparative studies of such alternative formalisms and mechanisms are still awaited, though the papers by Warren and Young begin to compare predicate calculus and production systems (respectively) with conventional programming languages. The criteria to be used in such comparisons may well be radically revised as we acquire a deeper understanding of distributed computing systems, in which many processors collaborate, working in parallel. (For more on this question of alternative representations, see Hayes 1974, Bobrow 1975, Sloman 1978, chapter 7.)

Different kinds of experts

Most of the expert systems described in this volume embody highly specialised concepts and knowledge, of medicine, chemistry, geology or some other topic which is not widely understood. By contrast, the papers by McCarthy and Hayes are concerned with concepts and knowledge possessed by all ordinary people. We all know that people can know things, can believe things, can think about things. We all know a great deal about physical objects events

and processes that can occur in our environment.

Machines which are able to think about people and take decisions about them (or give advice about them) will need to use mentalistic concepts of the sorts discussed by McCarthy, especially if they are to be able to explain their reasoning. In theory, they might function, up to a point, by simulating physical and chemical processes in human brains - but such a system would be as hard to understand as a human brain. When people think about, and reason about, other people (which we all do in our every day lives) it seems unlikely that we unconsciously simulate the brain processes of other people. Rather, we function with abstract descriptions of their mental states.

Similarly, in thinking and taking decisions about objects in our environment, for instance in deciding how to hold a teapot whilst pouring tea into a cup, it seems certain that we do not simulate the underlying atomic or sub-atomic events. Rather we perceive the world, think about the world, and act on the world using macroscopic concepts like *spout, rim, falling, tilting, spill, splash, nearly full,* and so on. Intelligent machines capable of explaining their actions to people, and capable of receiving instructions from people, will have to use similar concepts for representing physical states and processes. Such concepts are not required for machines which plan and monitor trajectories of space probes and moon shots, since the normal idealisations of physics suffice for that sort of job. Paradoxically, designing a domestic robot, able to clear up the dishes after dinner and wash them in the sink, is much harder.

The analysis of such everyday concepts, described by Hayes, like McCarthy's analysis of mental concepts, is very close to the kind of "conceptual analysis" practised by philosophers. (See Sloman 78, chapter 4). However, the motivation is different, since philosophers are often concerned merely to refute scepticism about the physical world, or simply to understand the conceptual frameworks used by people. It is at first sight curious that the philosophical activity of trying to understand aspects of human thought and language should have so much in common with the activity of designing intelligent machines. But there may be unique solutions to some complex design problems.

One issue that arises in the design of intelligent machines is how knowledge should be represented so as to be efficiently and effectively useable. (McCarthy and Hayes 1969 called this the problem of heuristic adequacy.) This issue does not arise when philosophers analyse concepts. They merely wish to record results in a form useful for communication with other philosophers. In addition they may wish to prove theorems *about* certain bodies of knowledge, or systems of concepts. A mode of representing knowledge that is useful for such philosophical and mathematical enquiries may be very different from what is practical for a machine which has to *use* the knowledge. The requirements of such a machine may also be closer to the requirements of ordinary people in their ordinary thinking and acting -- as opposed to the explicit theorising of philosophers and AI researchers.

Such practical uses may require elaborate indexing, explicit storage of much redundant information, "procedural embedding" of knowledge in strategies or programs for performing actions, and explicit storage of cues and procedures for controlling the selection of knowledge relevant to current problems. When Hayes and McCarthy recommend the use of predicate calculus as a formalism it is not clear whether they are talking about the needs of human theorists, or the needs of a working system. It is hard to believe that ordinary people and other intelligent animals use predicate calculus. We currently know very little about what sorts of formalisms and mechanisms are possible.

It is perhaps worth noting that not all forms of human expertise are readily articulated and formalised. Facts and theories explicitly studied by experts can often be expressed in English after a certain amount of reflection. Not so our knowledge of how to understand English, our knowledge of how to see, our knowledge of how to acquire new skills and concepts. This knowledge, if it can be called such, is very widely shared, and frequently used in everyday life, yet it is extremely difficult to explain how we understand speech, see objects or learn concepts. Paradoxically the types of expertise which are hardest for ordinary people to *acquire* (e.g. knowledge of science, mathematics, chess, and medicine) are easiest to *articulate* for the benefit of designers of expert systems. Visual expertise, shared between humans and many other animals, is particularly difficult to account for.

Tools vs theories

Richard Young's paper stresses the contrast between the goal of producing a useful machine and the goal of modelling and explaining human abilities. There is an enormous amount of overlap between the two sorts of research. In part this is because both require detailed analysis of *task* requirements.

For instance designing a speech-understanding machine and designing a model of human speech processing both require a study of the significant structures and relations in speech sounds, and an analysis of the syntax and semantics of natural languages. Just as ordinary people don't need to simulate details of each others brain processes, so also can scientists studying some aspect of the human mind represent their theories at a level which ignores details of the working of the brain. For this reason, computational theories cannot be criticised simply because computers are very different from brains, any more than a meteorologist can be criticised for not trying to represent the movements of individual sub-atomic particles in a weather-map.

Where there are alternative means of achieving the same task, the question arises: which of them best explains human abilities? Experimental studies of the externally observable behaviour of the system may be very indecisive. For instance very many different algorithms can all give the same input-output relations. The methodological issues are discussed at length by Pylyshyn (1978) and the commentators in the same journal. My view is that at the present time existing theories are much too simple to be taken seriously as explanations of human abilities. We don't yet know how to build sufficiently complex theories.

In particular human abilities involve a degree of generality, flexibility and extendability not matched by any present-day expert system. The ability to "degrade gracefully" in the face of increasing difficulties, which humans often display, will be especially important when expert systems are put to real use in the field. Sudden and inexplicable failures of complex systems arising out of a combination of circumstances not anticipated by the designers could be very annoying, if not disastrous. Similarly a consultant program unable to answer a question should be able to notice partial analogies with problems it can solve, and give hints or suggest relevant lines of exploration, instead of simply saying "I don't know". More generally, programs designed with a natural-language interface should be able to cope with mis-spellings, metaphors, bad grammar, false starts, elliptical sentences and other vagaries of human linguistic production. Notice how the task of explaining human abilities merges with the task of designing cleverer machines than we already have.

These requirements impose constraints on the way knowledge is represented and organised, and the amount of meta-knowledge (knowledge a system has about its own knowledge) needed by an expert system. This aspect of heuristic adequacy is a new field of epistemology whose study is still in its infancy. Examples of the issues involved include: the need for such systems to include mechanisms by which "demons" or "monitors" can be set up to watch out for special circumstances and take appropriate action, the need for methods of describing and drawing conclusions from similarities and differences between complex descriptions which are partly but not completely alike, the need for ways of extracting useful information from unsuccessful attempts at solving a problem, the need to be able to describe classes of problems or situations at a suitable level of generality, and so on.

My own impression is that production systems, predicate calculus and other existing formalisms are at best only relatively low-level "machine-languages" out of which far more powerful descriptive formalisms and inference systems will need to be constructed. Alternatively they may turn out to be limiting special cases. For instance both use matchers - but not network matchers, only tree-matchers.

Conclusion

This paper has attempted to supplement the detailed studies of expertise in the other conference papers with some general remarks about epistemological issues arising out of Artificial Intelligence. It should be clear that these issues are not unique to the design of clever artefacts. They are central to the study of the human mind. For this reason AI can have a powerful impact on philosophy, psychology, biology and perhaps even social studies of knowledge, combatting their theoretical poverty, arising out of a lack of good concepts and formalisms for representing complex processes.

But as the ideas diffuse through our culture they may also begin to revolutionise much less specialised modes of thought. Weizenbaum (1976) has argued that this could be very harmful. Boden (1977) and Sloman (1978) argue that on the contrary computationally informed thinking will enrich our understanding of ourselves and other people. Perhaps only time will tell who is right, since the science of social forecasting is still waiting to be born.

Bibliography

Bobrow, D.G. ‘Dimensions of Representations’ in D.G. Bobrow and A. Collins (Eds.) *Representation and Understanding* Academic Press 1975.

Boden, M.A. *Artificial Intelligence and Natural Man* Harvester Press and Basic Books 1977.

Hayes, P.J. ‘Some Problems and non-problems in representation theory’ in proceedings AISB Summer Conference 1974.

McCarthy, J. & Hayes, P.J. ‘Some Philosophical Problems from the standpoint of Artificial Intelligence’, in B Meltzer and D. Michie (Eds.) Machine Intelligence 4, Edinburgh University Press, 1969.

Pylyshyn, Zenon ‘Computational models and empirical constraints’, in *The Behavioural and Brain Sciences,* Vol 1, No 1, 1978.

Sloman, Aaron *The Computer Revolution in Philosophy* Harvester Press and Humanities Press, 1975.

Weizenbaum, Joseph *Computer Power and Human Reason* W.H. Freeman & Co. 1976.