In Proceedings 14th International Joint Conference on AI
Montreal, August 1995

A Philosophical Encounter

Aaron Sloman
School of Computer Science & Cognitive Science Research Centre
The University of Birmingham, B15 2TT, England
A.Sloman@cs.bham.ac.uk, http://www.cs.bham.ac.uk/~axs

(Reformatted July 28, 2014)

Abstract

This paper, along with the following paper by John McCarthy, introduces some of the topics to be discussed at the IJCAI95 event 'A philosophical encounter: An interactive presentation of some of the key philosophical problems in AI and AI problems in philosophy.' Philosophy needs AI in order to make progress with many difficult questions about the nature of mind, and AI needs philosophy in order to help clarify goals, methods, and concepts and to help with several specific technical problems. Whilst philosophical attacks on AI continue to be welcomed by a significant subset of the general public, AI defenders need to learn how to avoid philosophically naive rebuttals.

1 AI as philosophy

Most AI researchers regard philosophy as irrelevant to their work, though some textbooks (e.g. (Boden, 1978; Russell & Norvig, 1995)) treat the two as strongly related, as does McCarthy, one of the founders of AI. If we ignore explicit statements of objectives, and survey the variety of research actually to be found in AI conferences, AI journals, AI books and AI departments, we find that AI includes: The general study of self modifying information-driven control systems,

- both natural (biological) and artificial,
- both actual (evolved or manufactured) and possible (including what might have evolved but did not, or might be made at some future date).

This is extraordinarily close to a major concern of philosophers, namely asking what sort of minds are possible, and what makes them possible in a physical world. Some (like Kant) make the mistake of assuming that there is a *unique* set of necessary conditions for a mind, whereas AI research suggests that human-like mentality is not a simple all-or-nothing feature, but amounts to possession of a very large number of distinct capabilities, such as: many kinds of learning, seeing occluded surfaces as continuing behind obstructions, using quantifiers, making conditional plans, using nested sentences, and deferring goals. Different subsets can occur in different organisms or machines. Even humans have different subsets, according to age, culture, inherited dispositions, and whether they have suffered brain damage or disease. Thus 'mind' is a *cluster* concept referring to an ill defined collection of features, rather than a single property that is either present or absent.

Since different collections of capabilities define different kinds of minds, the old philosophical task of explaining *what a mind is*, is replaced by exploration of *what minds are*, through a study of their mechanisms, their capabilities, how they develop, and how some of them might evolve. I have described this ((Sloman, 1994a, 1995)) as exploring mappings between 'design space' and 'niche space', where niche space is the space of sets of requirements and constraints which may be satisfied, in varying ways and to varying degrees, by diverse designs.

This undermines two opposing philosophical views: (a) that there is a single major division between things with and things without minds and (b) that there is a continuum of cases with only arbitrary divisions. Both are wrong because there are many *dis*continuities in design space, corresponding to the presence or absence of particular capabilities (e.g. those listed above) that do not admit of degrees.

Another topic on which AI can advance philosophy concerns 'qualia', sometimes also referred to as 'raw feels'. These are defined variously as the contents of our experience, the answer to what it is like to feel, see or want something, and so on ((Dennett, 1991)). Some philosophers require that qualia have no physical effects and claim that different people may have different qualia without any objectively detectable evidence existing for the difference.

One reaction is to argue against their existence, as Dennett does. A deeper response will emerge from detailed work on the design of human-like agents. From an AI viewpoint it is obvious that a complete autonomous agent, unlike simple expert systems, must have myriad distinct, coexisting, interacting, information stores, including both long term collections of general information, personal history, procedural information, and short term stores corresponding to current goals and plans, suppositions, imaginings, thoughts, different levels in perceptual processing ((Marr, 1982; Minsky, 1987; Sloman, 1989)), and motor control. What is not so obvious is that an agent needs to be able to attend to and control some of its internal databases ((Minsky, 1987; Sloman, 1990; McCarthy, 1995)) and may need to be able to inform others about them, which we can do with varying degrees of accuracy (e.g. describing how we feel or how things look to us, or painting pictures, or setting up a situation that recreates the experience for others). By describing one's discomfort one can sometimes enable an expert (e.g. parent, or doctor) to prescribe a remedy. Attention to internal states may also play an important role in learning.

Whatever *they* may think, I claim that philosophers who talk about qualia are actually referring to internally detected states that are essential to the high level functional architecture of a sophisticated agent. Fleas may not need them. Of course, internal perception, like external perception, is liable to error, omission or oversimplification. In both cases, we can distinguish how things appear to the perceiver and how they actually are (e.g. from the standpoint of a scientist). Similarly a software system may misreport the contents of its data-structures. Of course, the agent or the system, cannot be wrong about how things appear to it, not because of privileged access but because that's what 'how they appear to it' means. Our ability sometimes to switch attention from the environment to these *internal* information states will not be explained until we have a detailed account of an information processing architecture that replicates and explains typical human capabilities, including introspection. On that basis we shall (in principle) be able to build a robot that has qualia and may wish to talk about them and may even propose the philosophical thesis that qualia exist in a non-physical realm.

But the robot's qualia, like ours, will be complex information processing states, whose identity depends on an intricate web of causal and functional relationships to other states and processes, just as the identity of a spatial location depends on a complex web of spatial relationships with other things. In both cases, if we change the relationships the question whether we still have the same thing becomes undetermined.

There is a powerful illusion that, by focusing attention on the thing itself, we can uniquely identify what we are talking about and ask whether some other thing (another's experiences, a location seen later) is the same as the original. Arguments showing the absurdity of this tendency are powerfully articulated in (Dennett, 1991). In some philosophers, the tendency is incurable. Perhaps teaching them how to design robots with qualia will finally cure some who resist all other treatments. But some incurables will always remain. One day, their ranks will include robot philosophers who claim to have qualia. Only when we understand why this is inevitable, will we have a complete theory of qualia.

There are many other ways in which AI can (and will) contribute to philosophy. There are unanswered questions about the nature of mathematical concepts and knowledge, discussed for centuries by philosophers in their armchairs. We shall gain a deeper understanding by doing experimental epistemology and studying designs for human-like information processing architectures that can learn about numbers in the ways that children do, including learning to distinguish between (a) empirical discoveries (e.g. adding two drops of water to three drops can sometimes produce one large patch of water, and counting the same set twice sometimes gives different answers) and (b) non-empirical discoveries (e.g. counting elements of a set in two different orders should give the same result, two plus three equals five, there is no largest prime number). Such mechanisms will require forms of learning and discovery not yet addressed in AI, including the ability to reflect on the nature of their own discovery processes, e.g. distinguishing results where the environment's input is essential from those determined entirely by the structure of the mechanisms and processes (as Kant argued).

Designing testable working systems will teach us new, detailed, precise, answers to questions in other areas of philosophy. A good specification of a mind-like architecture can be used systematically to generate a family of concepts of mental states, processes and capabilities, just as our theory of the architecture of matter enabled us to create new concepts of kinds of stuff, and the architecture of an operating system allows us to define states it can get into, e.g. deadlock and thrashing. Such a taxonomy of mental states will be far more complex and open-ended than the periodic table: for there is but one physical reality while there are many kinds of minds supporting different families of concepts.

A new potentially important area of influence of AI on both philosophy and psychology concerns the study of motivation and emotions. As designs for complete or 'broad' ((Bates, Loyall, & Reilly, 1991)) agent architectures develop, we can expect to obtain a much deeper grasp of how motivational and emotional states arise, along with moods, attitudes, personality, and the like. These are all important aspects of the mind as a control system, a point made in Simon's seminal paper (Simon, 1967) and developed in various ways since then e.g. (Sloman & Croucher, 1981; Minsky, 1987; Beaudoin & Sloman, 1993).

Philosophy benefits also from computer science and software engineering, which provide concepts such as 'virtual' or 'abstract' machine, 'implementation' and 'implementation hierarchy', and show how causal relations can hold between information states. I've argued in (Sloman, 1994b) that this answers philosophical questions about 'supervenience' (the converse of implementation) and shows how supervenient states can have causal powers, contrary to the view that only physical events have causal relations.

This undermines a common interpretation of Newell's and Simon's 'physical symbol system hypothesis' (e.g. (Newell, 1982)), for most of the symbols AI is concerned about are not physical, but structures in virtual machines. In fact, data-structures like sparse arrays show that there can be symbols that exist in a virtual machine without having any separable physical implementation: a large sparse array may contain far more items than the computer has memory locations. Only in the context of the whole implementation do all the array locations exist. Similar but more subtle global implementation relations probably hold between mental states and brain states, making the search for physical correlates of individual mental phenomena, including the detailed contents of qualia, futile. And yet these indirectly implemented structures can exist, and have causal powers.

2 Philosophy as AI

Not only does philosophy need AI to help with age-old problems, AI needs philosophy. To mis-quote Santayana: those who are ignorant of philosophy are doomed to reinvent it, often badly.

In fact, much AI already builds on work by philosophers. An obvious example is the use of speech act theory, developed originally by philosophers such as John Austin, John Searle and Paul Grice. There are also various uses of specialised logics, e.g. deontic logic, epistemic logic, and modal logics, originally developed by philosophers in an attempt to clarify concepts like 'permission' and 'obligation' (deontic logic), 'knows' and 'believes' (epistemic logic), and 'necessarily' and 'possibly' (modal logic). These contributions from philosophy are not passively accepted in AI: putting them to use in designing working systems often reveals shortcomings and suggests further development.

There are much older contributions from philosophy. One was Kant's proof in *Critique of Pure Reason* that learning from experience was impossible without some sort of prior (innate) conceptual apparatus. Another was Frege's heroic (but unsuccessful) attempt a century ago to show that all arithmetical concepts could be reduced to logical concepts and all arithmetical knowledge could be derived from logical axioms and rules. This led him to a number of extremely important results, including the first ever accurate analysis of the role of variables in mathematical expressions, discovery of the notion of higher order functions and invention of predicate calculus (accomplished independently by C.S.Peirce). This led (via work by Russell, Church and others) to lambda calculus, type theory, and other important notions in computer science and formalisms for AI. More recently the old philosophical controversy about varieties of forms of representations (e.g. logical and pictorial), which I discussed in (Sloman, 1971), has become a topic of active AI research ((Narayanan, 1993)).

Another recent development is recognition of deep connections between the AI task of understanding what sort of knowledge an intelligent system requires and the older philosophical activities of metaphysics, especially what Strawson (Strawson, 1959) described as 'descriptive metaphysics', including ontology, the attempt to characterise in a systematic way what exists. The word 'ontology' is now commonplace in the DARPA knowledge sharing effort ((kqml, 1994)). This is required both as part of the methodology of knowledge elicitation for expert systems, and also for design of robots intended to communicate with humans, act on human goals, use human criteria for resolving conflicts and deal with the unexpected in ways that are acceptable to humans ((McCarthy, 1990)). This extends the process outlined in chapter 4 of (Sloman, 1978), linking conceptual analysis in philosophy with articulation of knowledge for intelligent artefacts. McCarthy's paper gives more examples of connections between AI and philosophy. See also (McCarthy & Hayes, 1969; Hayes, 1985).

3 Two way influences, and more

I have listed some topics on which AI informs philosophy and others on which philosophy informs AI. In fact this is a spurious separation, for in all these areas the two activities inform each other, and as the depth of analysis increases, the amount of feedback increases, the work becomes more technical and specialised and the boundary between AI and philosophy will disappear.

Philosophers and AI theorists have worked independently on the role of rationality in intelligence. Much work by philosophers has been directed at clarifying conditions for rationality. Dennett's 'intentional stance' (Dennett, 1978) chapter 1, attributes beliefs and desires to agents on the assumption that they are rational. Newell's knowledge level ((Newell, 1982, 1990)) is also defined in terms of a presupposition of rationality. However deeper analysis shows ((Sloman, 1994b)) that mechanisms of intelligence can be understood at the information processing level without assuming rationality. Something closer to the design stance than to the intentional stance underpins ordinary concepts like 'belief', 'desire', 'intention'. The designs implicitly presupposed by folk psychology will, of course, need to be superseded.

A design for an intelligent agent may be constrained by resource limits and inevitable gaps in knowledge, requiring mechanisms and strategies that mostly work but cannot be justified as 'rational'. Sometimes the designer of a system can be regarded as rational even when the system isn't. More generally, though biological evolution (in effect) uses a fitness function to select the mechanisms on which our mental states and processes depend, the function need not be one that serves *our* goals. Evolution's goals are not our goals, except when the mechanisms it implants in us serve its wider (implicit) purposes. An example is the drive to produce, feed and shelter young, often at great cost to parents.

Human information processing mechanisms are extremely complex and unstable and easily diverted into states that serve neither the individual nor anything else. Only from the design stance can we understand the resulting pathological behaviour, where the assumption of rationality is clearly invalid, despite efforts of some therapists to portray mental illness as rationally based. (Insights from AI will eventually make a deep impact on psychotherapy.)

The disappearing boundary between AI and philosophy is nothing new. It is often said that as philosophers discover how to make progress in some area, that area ceases to be philosophy and becomes a new technical discipline: e.g. physics, biology, psychology, logic, linguistics, or political science. Compare the absorption of AI concepts and techniques by computer science.

This illustrates the artificiality of academic boundaries: often they exist only because of academic politics, or the organisation of research funding agencies, rather than because the problems and techniques have clear boundaries. In fact, the topics discussed here in the overlap between AI and philosophy will increasingly have to merge with studies in other disciplines, not least neuroscience, psychology, social science, and the empirical and theoretical analysis of how complex information processing systems like ourselves and other animals could have evolved in a world that originally contained only physical processes.

This short paper barely begins to list the myriad links between AI and philosophy. There are many topics I have not had room to address, including: consciousness and free will (both of them 'cluster' concepts rather than names for something that is either present or absent); issues raised by Searle and Penrose in their attacks on AI; how machines can understand the symbols they use ((Sloman, 1985)); the relevance of metamathematical incompleteness theorems; confusions surrounding the Turing test; the role of states like pain and pleasure in intelligent agents; ethical issues about the rights and responsibilities of intelligent artefacts; debates about the philosophical significance of the choice between connectionist implementations and symbolic implementations (I have argued elsewhere ((Sloman, 1994b)) that *architecture dominates mechanism*); whether mentality requires causal embedding in an external physical environment (as argued in the 'systems' reply to Searle); whether AI needs non-computational as well as computational mechanisms; analysis of the concept of 'computation'; and prospects for future forms of intelligence, including distributed minds. Some of these issues may turn up during discussions at IJCAI95. Many will recur at future AI conferences.

References

Bates, J., Loyall, A. B., & Reilly, W. S. (1991). Broad agents. In *Paper presented at aaai spring symposium on integrated intelligent architectures.* ((Available in SIGART BULLETIN, 2(4), Aug. 1991, pp. 38–40))

Beaudoin, L., & Sloman, A. (1993). A study of motive processing and attention. In A.Sloman, D.Hogg, G.Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for artificial intelligence* (pp. 229–238). Amsterdam: IOS Press.

Boden, M. A. (1978). *Artificial intelligence and natural man*. Hassocks, Sussex: Harvester Press. (Second edition 1986. MIT Press)

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Allen Lane: Penguin Press.

Hayes, P. (1985). The second naive physics manifesto. In (pp. 1–36). Norwood, NJ: Ablex.

(1994). (The KQML project and related activities are described in Web documents accessible via http://www.cs.umbc.edu/kqml)

Marr, D. (1982). *Vision*. Freeman.

McCarthy, J. (1990). *Formalising common sense*. Norwood, New Jersey: Ablex.

McCarthy, J. (1995). Making robots conscious of their mental states. In *Aaai spring symposium on representing mental states and mechanisms.* (Accessible via http://www-formal.stanford.edu/jmc/)

McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of ai. Edinburgh: Edin. Univ. Press.

Minsky, M. L. (1987). *The society of mind*. London: William Heinemann Ltd.

Narayanan, E. N. (1993). The imagery debate revisited. *Special issue of* Computational Intelligence, *9*(4), 303–435. ((Paper by J.Glasgow, and commentaries))

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*(1), 87–127.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Russell, S., & Norvig, P. (1995). *Artificial intelligence, a modern approach*. Prentice Hall.

Simon, H. A. (1967). *Motivational and emotional controls of cognition.* (Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979)

Sloman, A. (1971). Interactions between philosophy and ai: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai.* London. (Repr in *Artificial Intelligence*, 1971)

Sloman, A. (1978). *The computer revolution in philosophy: Philosophy, science and models of mind*. Hassocks, Sussex: Harvester Press (and Humanities Press).

Sloman, A. (1985). What enables a machine to understand? In *Proc 9th ijai* (pp. 995–1001). Los Angeles.

Sloman, A. (1989). On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, *1*(4), 289–337.

Sloman, A. (1990). Notes on consciousness. *AISB Quarterly*(72), 8–14. (Also presented at Rockefeller foundation workshop on consciousness, Villa Serbelloni, Bellagio March 1990, organiser D.C.Dennett.)

Sloman, A. (1994a). Explorations in design space. In *Proceedings 11th european conference on ai.* Amsterdam.

Sloman, A. (1994b). Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering*, *349*(1689), 43–58.

Sloman, A. (1995). Exploring design space & niche space. In *Proc. 5th scandinavian conf. on ai, trondheim.* Amsterdam: IOS Press.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proc 7th int. joint conf. on ai.* Vancouver.

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.