Originally entitled: ‘‘Artificial Intelligence and Popper’s Three Worlds’’.
Retitled by editors, for inclusion in:
Problems, Conjectures, and Criticisms: New Essays in Popperian Philosophy,
Eds. Paul Levinson and Fred Eidlin,
Special issue of *ETC:A Review of General Semantics*, (42:3) Fall 1985.

A Suggestion About Popper’s Three Worlds In the Light of Artificial Intelligence

(Previously: Artificial Intelligence and Popper’s Three Worlds)

Aaron Sloman
Cognitive Studies Programme
University of Sussex
(Now at University of Birmingham. http://www.cs.bham.ac.uk/~axs)

Abstract

Materialists claim that world2 is reducible to world1. Work in Artificial Intelligence suggests that world2 is reducible to world3, and that one of the main explanatory roles Popper attributes to world2, namely causal mediation between worlds 1 and 3, is a redundant role. The central claim can be summed up as: ‘‘Any intelligent ghost must contain a computational machine.’’ Computation is a world3 process. Moreover, much of AI (like linguistics) is clearly both science and not empirically refutable, so Popper’s demarcation criterion needs to be replaced by a criterion which requires scientific theories to have clear and definite consequences concerning what is possible, rather than about what will happen.

Introduction

Having always admired Popper and been deeply influenced by some of his ideas (even though I do not agree with all of them) I feel privileged at being invited to contribute to a volume of commentaries on his work. My brief is to indicate the relevance of work in Artificial Intelligence (henceforth AI) to Popper’s philosophy of mind. Materialist philosophers of mind tend to claim that world2 is reducible to world1. I shall try to show how AI suggests that world2 is reducible to world3, and that one of the main explanatory roles Popper attributes to world2, namely causal mediation between worlds 1 and 3, is a redundant role. The central claim of this paper can be summed up by the slogan: ‘‘Any intelligent ghost must contain a computational machine.’’

In passing, I shall comment on the relevance of AI to Popper’s demarcation criterion, suggesting that Popper’s views on the nature of scientific theories need to be modified. This essay does not attempt detailed justification of the theses presented: that will require an extended research programme.

Popper's Ontology

For the sake of argument I shall provisionally accept Popper's metaphysical theory that there are three worlds. World1 contains physical things, like atoms, lightning flashes, skyscrapers, eyelashes and planets. World2 contains ''subjective'' mental events, processes, and entities, like pains, acts of deciding, processes of imagining, and mental images. World3 contains ''objective'' propositions, theories, problems, proofs, numbers and the *contents* which may be common to the thoughts of two or more people. World3 objects are generally formulated or expressed in some concrete physical object, like a book, uttered sound, pattern of light and shade on a television screen. However, it is not *necessary* for their existence that they be so formulated, though Popper does say that they are brought into existence by the human mind (1976 - page 186). There are details of the theory that are obscure or controversial, but for my purposes such difficulties can be ignored. A fairly up to date summary of the main ideas can be found by following up the index entries under ''world'' in Popper(1976).

What is Artificial Intelligence?

For detailed accounts of work in AI readers are referred to Boden (1977) and Winston (1977). The former is philosophically more sophisticated, whereas the latter gives more technical details. I have discussed some of the philosophical implications in my (1978). For the present, a very brief summary of AI will have to suffice. Here are the main points.

1. AI is the study of actual and possible mechanisms underlying mental processes of various kinds - perceiving, learning (including concept-formation), inferring, solving problems, deciding, planning, interpreting music or pictures, understanding language, having emotions, etc. The basic assumption is that all such processes are essentially computational - that is they involve symbol-manipulations of various kinds, such as building, copying, comparing, describing, interpreting, storing, sorting, and searching.

2. Although theories about such mechanisms are usually embedded in computer programs, since they are too complex for their consequences, strengths, and weaknesses, to be explored ''by hand'', there is no commitment to any *specific* kind of underlying computer. In particular if a structure in Popper's world2 admitted a rich enough set of internal states, and laws of transition between states, it would suffice as a ''computer'' to run typical AI programs.

3. AI programs usually bear little resemblance to either the ''number-crunching'' programs familiar to most scientists and engineers, or to the Turing machines and ''effective'' procedures familiar to logicians and mathematicians.

For instance, many AI programs blur the distinction between program and data, since programs may be data for themselves, especially self-modifying programs. Further, the larger programs are often complex, messy and not necessarily fully intelligible to their designers, so that test runs rather than proofs or theoretical analysis are required for establishing their properties, especially after self-modification in a complex environment. Finally, they normally use several layers of ''virtual machine'' between the physical computer and the highest level program.

4. It is usual to interpret AI programs as constructing and manipulating complex

symbols to represent hypotheses, plans, possible states of affairs, goals, problems, criteria of adequacy etc. The programs go through symbol-manipulations which are best described as inferences, consistency checks, searches, interpretations etc. In other words, the processes are essentially concerned with world3 objects, properties and relations. Because of the complexity already alluded to, the programmer is not usually aware of all the hypotheses, inferences, decisions, etc. involved in a run of the program. Their existence therefore does not depend on their actually being the content of the programmer's thoughts.

5. Existing programs, though they constitute major advances in our understanding of the problems of explaining human abilities, are nevertheless pathetically limited by comparison with people and many other animals. In part this is due to limitations of computers available for such research: their memories are far too small and their computational power inadequate. Recent developments in micro-electronics and distributed processing may alter this. Another source of inadequacies is the piece-meal amateurish development characteristic of a young subject. Significant progress will require much more integration of theories and methods from psychology, linguistics, philosophy, anthropology and computer science.

Is AI a science?

It is worth noting that AI theories, although rich in content, since they are capable of explaining widely varying and intricately specified possibilities, often do not meet Popper's criterion for scientific status, since they are mostly at a stage of development where empirical falsification is not possible. As far as I am concerned this merely helps to show the inadequacy of Popper's criterion. The important thing for science is not that theories should have empirically refutable consequences, but that there should be varied and detailed consequences, and that *whether* something is a consequence of the theory should be objectively decidable. The consequences will not be empirically refutable, for instance, if they are of the form: X *can* occur or exist. I have argued elsewhere (Ch.2 1978) that explaining *possibilities* is a major function of scientific theories, and not just a hangover from their metaphysical pre-history.

Notice that I am not totally rejecting Popper's criterion. Like him, I claim that scientific significance of a theory is to be assessed in terms partly of number, variety, and types of consequences. Popper's mistake was simply to limit the range of admissible types to empirically refutable consequences. He could instead have criticised his ''metaphysical'' opponents simply by showing that their theories generated too few consequences or that which alleged consequences did and which did not follow from the theories was not usually objectively decidable because of the inherent vagueness or openness of the theories. I see no point in striving for a *monolithic* demarcation between science and non-science. It is not usually fruitful to divide the world into ''goodies'' and ''baddies'' when in fact there are many overlapping spectra of merit, with the same individuals often occupying different locations in different spectra. (Although Popper himself does not link his demarcation criterion with an *evaluative* distinction, it has been so used by many of his followers.)

Thus, the important question is *not* ''Are AI theories scientific or metaphysical, or whatever?'' - but ''what are their specific merits and faults and how can they be improved?'' Their study will prove a rewarding field for philosophers of science, as they are so different from theories in more familiar branches of science.

Reducing World2 to World3

Nobody could sensibly claim that any existing machine has experiences closely analogous to those of people. Certainly there are at present no pains, tingles, thrills of delight, anxieties, loves, hates despair etc. But there are the beginnings of visual experiences - quite richly structured internal states in which images are interpreted as first this then that, with attention shifting between different parts and aspects of the scene. There are also processes in which references are assigned to English phrases, processes involving exploration of alternative actions, and choices between such alternatives (for examples, see Boden and Winston). Some hypothesis are rejected and others accepted as true. Programs which represent goals and use them to generate new subgoals indicate how systems which have their own motives might be built. (See also Boden 1972.) Above all, there are the beginnings of self-awareness, in programs which can monitor their own performance and modify themselves accordingly. (See Boden, 1977, Part V; and Sussman 1975).

It is already possible to see in very rough outline how the phenomenology of physical pain could be replicated in a robot by subsystems which monitor parts of the ‘‘body’’ and on detecting disturbances and malfunctions generate new symbols within a special store of motives which control the direction of attention and influence priorities in decision-making. The ‘‘warnings’’ produced by such monitors would need to have fairly rich descriptive content - concerning location, spread, urgency, and nature of the malfunction or injury. This would account for some of the qualitative differences between different sorts of pains. (Dennett 1978, Chapter 11, provides an illuminating discussion of the problems.)

It is possible to see, again in rough outline, that much of the phenomenology of emotions could be replicated in machines which are able to detect the presence or absence of factors which help or hinder the attainment, preservation, or prevention of events or states which fulfil or conflict with the machines motives. Whether there is an emotion will depend on the extent and nature of the disturbance, within the machine’s internal processing, produced by its discovery.

Notice that it is not mere replication of the *external* behaviour of some human being that we need to aim for. A huge condition-action table could do that. The structure of the internal processing is what is important, for instance in determining the potential for alternative lines of development in an indefinitely large and varied set of circumstances.

These sketches could already be amplified in some detail. Nevertheless there are still many unsolved problems, so I do not claim that it is *established* that much of the structure of our internal or ‘‘subjective’’ experience, could be mirrored in processes inside a symbol-manipulating system of the same general character as existing AI programs, though with far greater complexity. (In collaboration with a research student at Sussex University, Monica Croucher, I am exploring some of these ideas in more detail, and hope to publish reports later on.)

Would world2 events, objects, processes, etc. exist in such a symbol manipulating system? I do not believe that any rational *argument* can answer this question decisively. Ultimately, it is a question for moral decision not factual discovery. If a robot were to be made whose internal design and verbal and non-verbal behaviour indicated decisively that computational processes structurally similar to typical human mental processes occurred within it, this would still leave some people saying: but it is only a *machine* and so, by definition ordinary mentalistic language is inapplicable to it. (E.g. Boden 1977, pp 418-426).

At that stage, with the machine, pleading or ‘‘pleading’’ for friendship, for civil rights, for a good education, for a less uncomfortable elbow-joint, or whatever, we’d be faced with what I can only describe as a *moral* or *political* disagreement between those who asserted or denied, that it was conscious and suffered. (Similar problems arise with animals, brain-damaged people, etc.)

Popper’s position on this issue is unclear to me. I suspect he would not join the society for prevention of cruelty to robots, not because he does not oppose cruelty, but because he appears to be convinced, in (Popper & Eccles 1978) that *only* an animal brain can provide a basis for world2.

However, I see no reason to share this conviction, and neither would many people who had grown up with such robots as playmates, nannies, house-servants, etc. Thus for me, and for such people, it would seem morally right, to attribute subjective mental states, processes and events, to an individual with sufficiently rich and human-like internal computations. Apart from racial, or species, prejudice, we’d have as much reason for doing this as for treating people, cats, monkeys, etc. as conscious.

But as indicated above, the computations in such a machine essentially involve states, processes, events and objects in world3. That is, they involve such things as symbols, theories, plans, decisions, refutations, inferences, and such world3 relations as implication, inconsistency, representation, denotation, validity and the like.

If, from a certain moral standpoint, the existence of certain sorts of such world3 phenomena is a sufficient condition for the existence of much of world2, then we have a kind of reduction of world2 to world3. More generally much work in AI can be interpreted as an attempt to show that world2 processes in people are really world3 processes. This may seem paradoxical to philosophers who normally think of computers and AI in the context of attempting to reduce world2 to world1, the physical world of brains and atoms. But such a reduction is of little interest in the light of AI, since, as stated previously, it is relatively unimportant whether AI programs run on a physical computer, a spiritual or world2 computer, or a ‘‘virtual machine’’ constituted by software (i.e. more programs) in a physical computer.

Levinson has suggested, in correspondence, that a moral distinction can be made between apparently intelligent artefacts and humans or other animals, since the former are deliberately created and fully understood whereas the emergence of life and intelligence in the animal world remains an unexplained mystery, and therefore a suitable basis for awe and reverence. (I have paraphrased his argument.) My answer is two-fold. First the ‘‘therefore’’ expresses a debatable moral position. Secondly, and more importantly, successful design of a human-like robot would remove much of the mystery, or at least help to remove it. The remaining task would be to extend existing biological studies to account for the evolution of *programs* in living things.

The causal dispensability of world2

Finally, whatever the nature of the human mind, it is indubitable that AI programs do run in physical computers at present, and that physical events (e.g. light entering TV cameras) can cause computational events to occur, and that computational events can cause physical events to occur, such as switching motors on and off in mechanical arms. Hence, within already *existing* systems we have causal interactions between world1 and world3, whereas according to Popper (e.g. 1976 page 185) this should require the mediation of world2. Popper could rescue his claim by granting my main thesis concerning the reduction of world2 to world3, and then stating that

what is required for causal mediation between world1 and certain sorts of world3 objects, such as theories, proofs, problems, etc. is the presence of *other* kinds of world3 entities, namely the sorts of symbol manipulations which constitute mental processes!

Levinson, in correspondence, has suggested that Popper would argue that since AI programs (and any robots that may one day descend from them) are products of human minds, they depend on world2 for their existence. World2 is therefore not causally dispensable. This, however, presupposes that human mental processes are somehow essentially different from the world3 processes in A.I. systems. As I pointed out above, this is essentially a *moral* view. The existence of computational systems with human-like abilities, strongly suggests that instead of being mysterious other-worldly entities, human mental states, processes, events, etc. are computational entities on a biological computer. I.e. they are World3 entities too. However, I do not claim that this has been established. It is a conjecture linked to a flourishing, exciting and revolutionary research programme.

Concluding remarks

I have tried to live up to Popper's recommendation that conjectures should be bold and rich in content. The theses sketched above have plenty of consequences for traditional problems about the relation of mind and body. For instance they imply that certain sorts of physical conditions may be (morally) *sufficient* for the existence of mental processes, but that the existence of a physical world anything like this one is not *necessary* for the existence of mind. There are many more implications of AI research, concerning knowledge, reasoning, and the nature of free decisions, some of which I have begun to explore elsewhere. (See also Boden 1977, ch. 14.)

Another of Popper's admirable recommendations is that one should expose potential weaknesses of one's theories, instead of merely displaying their strong points. I must therefore end by acknowledging that there are two aspects of the phenomenology of conscious experience which I do not yet see how to account for in terms of internal symbol-manipulations, namely pleasant physical sensations, and finding something funny. Perhaps pleasures in general and physical pleasures in particular can, like pains, be accounted for in terms of conscious and unconscious perceptual processes which interact in a suitably rich way with motivations, decision making, priorities and the control of attention. (Analysis in terms of excitation of a physical subsystem labelled a ''pleasure centre'' of course explains nothing). As for finding something funny - I suspect that is essentially connected with being a *social* animal. If so, a robot with a sense of humour will have to have an awareness of, and a concern with, other sentient beings built deep into its system of beliefs and motivations. But this idea remains to be clarified and tested by detailed exploration.

Bibliography

Boden, M. *Purposive Explanation in Psychology* Harvard University Press 1972, Harvester Press 1978.

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

Dennett, D.C. *Brainstorms: Philosophical Essays on Mind and Psychology,* Bradford Books and Harvester Press.

Popper, K.R. *Unended Quest,* Fontana/Collins, 1976.

Popper, K.R. & J. Eccles *The Self and Its Brain* Springer Verlach, 1978.

Sloman, A. *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind* Harvester Press and Humanities Press 1978.

Sussman, G.J. *A Computational Model of Skill Acquisition* American Elsevier 1975

Winston, P. *Artificial Intelligence* Addison Wesley, 1977.