THE STRUCTURE AND SPACE OF POSSIBLE MINDS

Aaron Sloman

Cognitive Science Research Paper

Serial no: CSRP 028

The University of Sussex
Cognitive Studies Programme
School of Social Sciences
Falmer
Brighton BN1 9QN

This is a much revised version of a paper to appear in:
Proceedings Anglo-French Philosophy conference
'The Mind and the Machine'
Ed. Steve Torrance,
Middlesex Polytechnic
April 1983

DRAFT -- COMMENTS WELCOME

THE SPACE OF POSSIBLE MINDS

Aaron Sloman
Cognitive Studies Programme
University of Sussex
Brighton BN1 9QN

Introduction

General questions about the nature of mind have long been of central interest to philosophers. On the whole scientists have felt that they can get on with their jobs without engaging in philosophical debates. But it would be desirable for scientific studies of living systems to be based on a general theory concerning the nature of mind, to provide a framework for investigating of the mental processes and behaviour of different sorts of animals. In particular, if we are to understand the evolution of intelligence we must surely have some understanding of the space of possible intelligent systems, to guide hypotheses about how a new system might develop from an old one. This may also help to constrain theories of learning. A good general theory of mind would also aid engineers trying to design intelligent systems.

Clearly there is not just one sort of mind. Besides obvious individual differences between adults there are differences between adults, children of various ages and infants. There are cross cultural differences. There are also differences between humans, chimpanzees, dogs, mice and other animals. And there are differences between all those and machines, which many suppose can have minds. Besides existing animals and artefacts, we can also talk about theoretically possible systems. There is a very large and varied class of possible minds.

My use of the word 'mind' may be thought to beg too many questions. For one thing it is arguable whether the word is really applicable to machines. For another, it is arguable whether animals like dogs and mice, and some would even say chimpanzees, have minds. Further, from a biological standpoint it is not obvious where to draw a line between things with and things without minds. But it is not easy to find a suitable alternative way of characterising the sort of theory we are looking for.

One thing that is common to all the things which we might be tempted to say had minds is that they can produce behaviour. The behaviour may be internal - for instance day-dreaming or planning. So things with minds are behaving systems. But many behaving systems clearly do not have minds in any ordinary sense of the word, for instance carnations, clouds and clocks. It may be useful nevertheless to try to embed a general theory of mind within a still more general theory of behaving systems. This may provide us with a framework in which to characterise the fuzzy boundary between things with and things without minds.

There is still a problem about what should count as a behaving system. Clocks in some sense make things happen, and may control events. Clouds may have far more violent motion within them: yet we are less inclined to regard that as behaviour under the control of the cloud. A chair moves if you push it. Is that an example of behaviour? A bow with a rotten string may suddenly snap straight. Is that behaviour? Is the eruption of a volcano behaviour? The tumult on the surface of the sun?

We seem to need a new subdivision, or perhaps several subdivisions. Should we distinguish behaving things according to whether they use energy stored within themselves or external sources? We might hope thereby to rule out the acceleration of a rain drop - caused by gravity. But this would distinguish clocks with springs or batteries from clocks driven by weights. Moreover much human behaviour, including movement and murder, uses external energy sources: cars and bombs. Can we distinguish things whose behaviour is controlled by stored plans from things whose behaviour isn't? That begs the question whether stored plans underly all the intelligent behaviour we are ultimately interested in? Can we find a less question-begging formula, perhaps by talking about things whose behaviour is controlled by internal structures? But the shape and distribution of mass of a stone controls the way it spins if tossed into the air, and the shape of a kite controls the way it responds to changes in the wind. Do stones and kites control their own behaviour? Can we find a way to exclude them from the systems we wish to study? Why should we try?

The existence of a useful word or phrase in our language can mislead us into thinking that we already understand some clear principle according to which we can distinguish instances from non-instances, even if we find it hard to say what the principle is. But words may serve useful communicative or social purposes even though they are not used according to precise criteria.

For instance, we often find it useful to ask whether someone did something freely. The fact that in some circumstances certain sorts of actions may be described as done freely or not done freely has misled many people into thinking that there is a clear context-free distinction between actions done freely and those which are not. Yet thousands of words have been written by philosophers and theologians arguing about what the distinction is and whether there is a distinction at all.

Similarly, because there are many contexts in which the words 'mind' and 'consciousness' are useful, we are tempted to think that we grasp some clear distinction between things with minds or consciousness and things without, however difficult it may be to explain that distinction. Equally, we may feel there is a distinction between things which control their own behaviour (even if only partly) and things which merely behave, however difficult it may be to explain this distinction.

I suggest that the best way forward is to make no assumptions about whether such words of ordinary language mark clear and general distinctions, and instead to explore the structure of the space within which such distinctions may be made. We may later decide to label boundaries within that space, when we have a full understanding of the nature of the boundaries.

> (Note: in chapter 4 of Sloman [1978] I argued that concepts of ordinary language embody much useful information about the world. This is consistent with my present claim that the distinctions labelled by ordinary words may be of restricted applicability, and may be useful only for certain purposes. Looking at how those distinctions work may unearth much factual (non-linguistic) information of great scientific value even if the words are not sufficiently general or precise to be suitable as scientific technical terms.)

In particular, we should not assume that words like 'mind' and 'consciousness' are useful pointers to major boundaries in the space.

People are often tempted to discuss questions about the evolution of consciousness as if we all understood clearly what is meant by the question 'How did consciousness evolve?' (or the question ('Did it evolve?'). The question is often posed as if consciousness were a thing, like an eye, which an organism either has or doesn't have. Or it may be thought of as being something which can be present to different degrees. The problem is that the concept serves certain limited practical purposes in ordinary conversation, and cannot easily be generalised for scientific purposes. We can often ask and answer questions like: "Has he regained consciousness?' 'When did he lose consciousness?', 'when did your consciousness of your inferiority begin?' 'Were you conscious of how late it was?' But it doesn't follow that every kind of use of the word 'consciousness' or 'conscious' refers to the same state or entity which is definitely present or absent.

Our concept of consciousness is very rich, subtle, and partly incoherent: for it wasn't designed to cope with all sorts of situations. E.g. when a sleep-walker carefully unlocks a door, is he conscious of the door? If he's asleep he can't be conscious. But he can't unlock the door so carefully without seeing the key and the lock and that seems to imply that he is conscious.

Describing the space of behaving systems

If we are to make a sensible selection of a subset of the space of behaving systems, for a general theory of mind, we shall first need a deeper understanding of the structure of that space. Can we find some general characterisation of the larger space? What would such a characterisation be like?

One approach would be to attempt to list possible primitive elements of behaving systems, then formulate rules of composition. In this way we might generate descriptions of all possible behaving sytems. This would provide something like a 'grammar' for the space. But a grammar for a language merely characterises its syntax, i.e. the possible structures of sentences, without saying anything about the meanings, or functions. Similarly a grammar for possible behaving systems would merely characterise the structures of behaving mechanisms, without saying anything about the sorts of behaviour they can produce - what we might call their semantics.

Clearly, any general theory of behaving systems must talk both about the mechanisms and the behaviours, and the relations between the two. This requires notations for describing mechanisms, notations for describing behaviours and notations for representing the relationship. This paper will not present such notations: developing good general notations is a long term research task. Instead the approach will be very informal, attempting only some preliminary groundwork, focusing on issues relevant to understanding the nature of intelligent systems. The ideas are tentative and subject to revision. Criticisms and suggestions are welcome.

Before moving on, it may be useful to look at some old and inadequate theories about the distinction between things with and things without minds.

Some views concerning minds

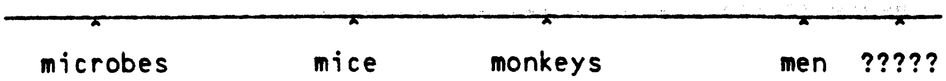
Some philosophers and theologians have argued that there is a simple single division:

Things with minds | Things without minds

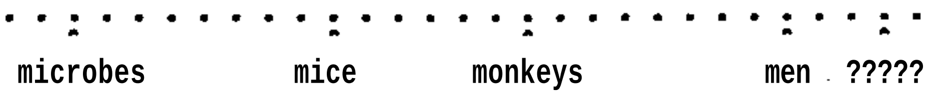
The idea is that having a mind requires, besides physical mechanisms such as brain cells, some additional non-material entity, dubbed 'the ghost in the machine' by Ryle [1949]. The extra entity is either present or absent: hence the simple dichotomy.

There has been much discussion and criticism of this view, which is closely related to the view that consciousness is a kind of stuff that may be present or absent. A major problem is that simply postulating an extra kind of stuff does not explain any of the rich and detailed behaviour that we associate with minds. Nor does it account for individual differences. It explains none of the observed fine-structure of human mental capabilities. We would need the spiritual addition to have a great deal of internal structure, with the capacity to store memories, and many sub-mechanisms able to produce different sorts of thoughts, plans, decisions, desires, etc. In short, we need a machine inside the ghost, even if it is a non-physical machine. Once that possibility is acknowledged, it becomes an open question whether the explanatory mechanism might not after all be a physical mechanism, of a type not considered by those who first postulated the ghostly addition. But for our purposes we shall see that it is the logical structure of the mechanism that is of interest, not whether it is made of physical matter or implemented in some other way -- if there is another way. (It's the structure not the stuff that explains.)

Others, impressed by the huge range of abilities found in living organisms, through microbes, mice, monkeys and men, and unable to find any clear demarcation lines, have argued that there is some sort of <u>continuum</u>, with simple organisms (or possibly machines) at one end, human beings at the other, and other organisms between, something like the colour spectrum.



One problem with this view is that we do not actually find a continuum in the precise mathematical sense. There are gaps and jumps in the abilities of known systems, without any reason to suppose that all intermediate cases are possible. This could lead to the revised hypothesis that the space is not continuous, but discrete, e.g.



But this still assumes a total ordering. There is no basis for such an assumption, any more than for the assumption that physical mechanisms can usefully be thought of as fitting into a linearly ordered space. On the contrary, for any given behaving system we can often envisage a number of different sorts of changes which would produce different systems. We could enlarge the memory, or improve the indexing algorithms, or add new learning mechanisms, or add a new type of perceptual subsystem, etc.

Thus we must think of the space of behaving systems as more like a network, where each node in the net represents a possible behaving system (a clock, a cuckoo, a chess machine....), and links to other nodes in the network represent possible changes which would produce a different system. So the differences between sorts of behaving systems are not merely differences of *degree*. Rather there are many differences of *kind*, corresponding to different transitions. Looking for something measureable which varies continuously or linearly (and assuming that consciousness corresponds to a high value - e.g. degree of complexity of organisation), is therefore misguided, and hinders a deep understanding of the space of possible behaving systems.

Some scientists, influenced by the sorts of mathematics which have proved useful in physics, are prompted by such arguments to suggest that instead of a one-dimensional space we need an N-dimensional 'parametrisable' space, for some yet-to-be-found N. That is, just as we can characterise a point in ordinary space by three numbers (e.g. latitude, longitude and height above sea level), so we may be able to find N measures which can be applied to any behaving system, which will completely characterise that system. N would perhaps have to be a very large number.

This presupposes that the space of behaving systems has a uniform structure, whereas a little reflection suggests that it doesn't. Given any point in an N-dimensional Euclidean space there are N different 'directions' in which the position may be changed, and they are the same for all points. Similarly, given N numerical measures we can consider numerical changes in any of them, no matter what their actual values are. However, if we think of behaving mechanisms, the possible alterations which could be made to one sort of mechanism may be quite unlike the sorts of alterations which could be made to another. For example, a clock can be changed so that it goes faster or slower: this would be a quantitative change. Structural changes are also possible: perhaps with the addition of a few extra components it could be made to indicate more things than it does - e.g. the date and year could be added to hours and minutes. By contrast if we are talking about a computer-based chess machine we can consider quantitative changes such as speeding up the processor, or enlarging the memory, and qualitative changes such as modifying algorithms to enable them to cope with more conditions, or adding a learning component, or changing the interface to the user, etc. Modifications to the chess player are of a type which it just doesn't make sense to try to apply to the clock. We cannot sensibly talk about giving a clock a new strategy for detecting a possible check-mate. In other words, the clock and the chess machine are not in a space with a uniform structure: different directions of change are possible in different locations in the space of behaving systems. In particular, there is no set of N measures which can be used to parametrise the space. This implies, for example, that any evolutionary biologist trying to characterise the behavioural abilities of organisms in terms of an ntuple of parameters is unlikely to produce descriptions of any theoretical value.

Can we say anything more positive about the space of behaving systems? It will probably be useful to try to build on work in theoretical computer science. Such work has, for instance, distinguished finite state machines, machines with N push-down stacks (for various N), machines with bounded randomly addressable memory, machines with infinite randomly addressable memory (Turing machines), deterministic machines, stochastic machines, etc. These categories have been related to various kinds of neural net structures. It will be necesary to extend the theory to include hybrid machines with analog as well as digital components, though analog components can be approximated as closely as required by digital ones. However this approximation may not be of much use for theoretical analysis, just as the possibility of reducing all of number theory to a few logical axioms is not of much use for practising number theorists, although it is of great theoretical interest.

Similarly, it will probably turn out that the very general categories and distinctions so far studied by computer scientists are not very helpful for the practical task of understanding very complex intelligent systems like mice, or monkeys. We may need higher level functional and structural categories. Thus in very general terms we may distinguish machines with a changeable internal store and machines without. But for understanding intelligent systems we may need to talk about different sub-stores used for different purposes, e.g. storing factual information, storing plans, storing current goals, storing goal comparison routines, etc. etc.

Again, there are important differences between single-processor serial machines and systems composed of independent sub-systems which operate in parallel and asynchronously, even though the latter can be approximated as closely as desired by the former, given sufficient processor speed. The serial simulation of asynchronous parallel systems introduces conceptual complications which are best ignored if we are simply trying to understand the parallel systems. So distinctions which may be superfluous in principle may nevertheless be very useful in theory construction.

There are many different sorts of aspects of behaving systems which computing scientists and others have begun to investigate. E.g.

* Different kinds of processing units - e.g. neurones, transistors
* Different kinds of hardware organisation - e.g. serial processors, parallel processors, networks, pipelines, trees
* Different kinds of program architectures.
$ Different resource-allocation strategies.
* Different kinds of symbolic structures - e.g. logical formalisms, procedural languages, 2-D arrays; linear or network like; hierarchical or recursive.

- Different kinds of algorithms for operating on structures - e.g. corresponding to different sorts of programming languages; digital processes or analog processes (e.g. inhibition/excitation)
- Different modes of interpretation to give semantics to symbols - e.g. logical, analogical (iconic), procedural.
- Different kinds of roles which can be given to representations - e.g. storing information, representing goals, representing strategies, etc.
- Different kinds of 'memory' stores - e.g. read only, read and write, write once, deterministic, stochastic.
- Different kinds of learning ability.

Is the space of possible minds something different? That depends on whether we can find some massive boundary within the space of behaving systems which clearly hives off the mental from the non-mental. People often argue as if there were such a boundary even though they cannot agree on where it is. The passion accompanying such debates suggests that more than a search for truth motivates the disputants.

Both sides assume that there is some well defined concept of 'mind', 'consciousness', or whatever, whose boundaries are to be discovered, not created. But these are complex and subtle concepts of ordinary language, not designed for scientific classificatory precision. When using them of our fellow men, or animals, we don't first check that certain defining conditions for having a mind or being conscious are satisfied. Rather we take it for granted that concepts are applicable, and then we make distinctions between quick and slow minds, conscious and unconscious states, feelings of various sorts, etc. Equally we take it for granted (most of the time) that such concepts and distinctions cannot be applied to trees, lakes, stones, clouds. (However, not all cultures agree on this.) But we don't discriminate on the basis of any precise shared definition of the *essence* of mind, consciousness, or whatever. For there is no such precise shared definition.

One traditional way to seek an essence is through introspection. However, nothing learnt in this way about the nature of mind or consciousness could help us distinguish *other* beings with and without consciousness, for we cannot introspect the contents of other minds.

Another approach is to seek behavioural definitions of mental concepts: but these founder on the objection that behaviour merely provides evidence or symptoms and does not *constitute* what are essentially internal states.

The only hard-headed, scientifically acceptable, alternative until recently has appeared to be to locate mind in brain matter. But this ignores important category distinctions: although neuronal states, events or processes may correlate with or underlie my being conscious, they are not themselves consciousness. Mental states and processes are not anything material. This is not because they are something non-material which co-exists, but for the same reason as the horse-power of an engine, or the flexibility of a computer program is not something material: not every category is a 'stuff' category.

We need to explore the similarities and differences between different sorts of systems including those we all agree have minds, those we all agree don't have minds, and those about which there are disputes. Having explored the similarities and differences we should not expect to find some objective essence common to things which really[1] have minds, and absent from the rest.

The theoretical questions can interact strongly with human values. The fervour which which some people argue that machines can be conscious may be bound up with a desire to remove any special status from people. The strength of opposition is bound up with unwillingness to admit that machines may have rights and responsibilities or to enter into truly *personal* relationships with them: e.g. feeling pity when they are harmed.

This interaction with powerful ethical or other values can often lead to discussions which are at cross purposes, in that disagreements which appear to hinge on factual questions (does it have purposes?) really hinge on ethical disagreements (it should/should not be treated as if it had purposes). What is really an ethical disagreement, e.g. an disagreement about how some kind of animal or machine *ought* to be regarded and treated may be disguised as disagreement about *what* sort of thing it *is*.

I suspect that only the confusion of ethical issues and factual issues leads us to think there is some well defined boundary waiting to be discovered. This is because distinctions between different sorts of behaving systems interact strongly with distinctions of an ethical or evaluative type, e.g. distinctions between things whose survival we value highly and those we don't; or between things we believe to have rights and responsibilities, and those we don't. Because we feel a need to make such ethical distinctions, and to treat them as something more than just our personal, or culturally determined, reactions, we assume that there is some absolute, objective, difference between the systems themselves. Yet the fact remains that individuals or cultures differ in their ethical views, without such differences being based on any differences in factual evidence or logic. It follows that if attributing consciousness to certain general classes of things and not to others reflects such ethical reactions, then the word 'consciousness' cannot be taken to refer to something which is objectively present or absent. (This inference is highly tendentious, and detailed elaboration of the argument would have to take account of a great deal of philosophical literature, much of it arguing that ethical theories are as capable of truth or falsity as scientific theories.)

Minimal transitions in the space of behaving systems

One way of exploring the space of possible systems is to look at what sorts of transitions are possible from one system to another. Not all transitions are permissible for all possible systems: we can talk about adding a certain sort of learning algorithm to a chess machine even though it could not be added to a clock or thermostat or even a pocket calculator.

In particular, it may be useful to identify *minimal* meaningful transitions. This concept is difficult to make precise. The key idea is that there are transitions which involve a great many different and independent changes, and these are not 'minimal'. E.g. it may be possible to change system A into system B by changing some of the planning algorithms, and the way certain kinds of information are represented, and the learning strategies. But if each of these changes can occur without the other, then the change from A to B is not a *minimal* change. In the case of quantitative changes, such as adding more memory or increasing the speed of some processing unit, we can treat each change in one dimension as 'minimal', to avoid issues about how small the alteration has to be to be minimal. We are not interested in the differences between small and large changes (except in circumstances where these produce qualitative differences, e.g. if a certain new algorithm cannot run without a minimal addition to the available workspace.)

Obviously the set of non-minimal changes will be very much larger than the set of minimal changes. Only the latter has any hope of being finite: whether it is finite is an interesting research question. (If we include quantitative changes, then the set of possible changes may be infinite (or nearly infinite) for uninteresting reasons.)

There is still a problem in that a change which is minimal relative to one level of description need not be minimal relative to another. For example if we have a list of rules which may or may not be present in a particular sort of expert system, then the addition or removal of a rule would be a minimal change at that level of description. But individual rules may be implemented in terms of complex data-structures, such that the addition of a rule requires a number of alterations to the working memory of the system. Since some of the alterations may be made without others being made, the addition of the rule is not a minimal change.

This problem of levels of description is quite general. Theoretical analysis may proceed at different levels: and issues which are of concern at one level need not be relevant to another. Thus in discussing the relative complexity of two algorithms A and B, one can validly argue that A, which trades space for time, will involve fewer operations, at a certain level of description, than B. This may be true even though certain implementations would make A slower. For instance, if the algorithms are implemented on a machine with a virtual memory system, and a small amount of main memory, then all the structures required for B might fit into the memory, whereas A would need a very much larger 'virtual' memory, and so running A would frequently involve swapping information between the computer's main memory and a disk store. The total number of operations at the level of the physical machine would then be larger for A, whilst for the higher level virtual machine the number of operations is larger for B. Unfortunately, its the physical machine which determines the *actual* speed of computation.

In this sort of case it is perfectly proper to analyse the systems at a high level of abstraction, as long as one does not draw general conclusions about all possible ways of *implementing* the systems. This proviso is important for a scientific study of biological systems, since showing the theoretical advantages of a certain type of mechanism will not tell us anything about evolutionary advantage if such a mechanism cannot itself be efficiently implemented in the classes of organisms under investigation. (Scepticism about the relevance of AI to the study of human psychology may be based on a special case of this general difficulty.)

In spite of all this, it would seem wise not to try to solve all problems at once, and to allow our investigation to proceed in parallel at different levels of abstraction. Thus we can study the pros and cons of transitions between systems at a fairly abstract level of description, while acknowledging that the transitions need not be minimal ones as far as lower level descriptions are concerned, and that the pros and cons may be influenced by how the higher level virtual machine is actually implemented in terms of lower level physical machines.

Subject to these warnings, then, the proposal is that a new sub-task for Cognitive Science is to survey minimal (qualitative) transitions in the space of behaving systems. For any transition, studied it will be necessary to say something about:

(a) the mechanisms presupposed, (since the transition may not be relevant to all possible points in the space),

(b) the changes involved in those mechanisms (at the appropriate level of description), and

(c) the functional consequences of the transition: i.e. what are the advantages and disadvantages of a change from one sort of system to another.

Besides contributing to other disciplines, such as psychology, ethology, philosophy, and AI, the study of the space of behaving systems, and paths through the space, is of interest in its own right as a theoretical study. It could be thought of as either a new branch of philosophy, a new branch of computer science or a new branch of mathematics, depending on the precision and rigour of the analysis!

Virtual machines

Computing Science has provided us with the concept of a virtual machine, within which computational states and processes can occur. This concept is useful for understanding some of the transitions in the space of possible behaving systems. In particular it is possible for a system to change in such a way that it implements a new virtual machine, with new capabilities.

A virtual machine is a structure which can undergo various changes of state, where the state is defined relative to a certain class of descriptions, and certain transitions from one state to another are defined as legal. (E.g. they may be the primitive operations in terms of which the machine can be programmed.)

How the state of a virtual machine changes is determined by rules. Which changes can occur in any given state will depend on certain conditions (e.g. is the word "go" stored at a certain location). If the conditions refer to parts of the system whose states are not fixed then the machine is programmable by specifying the contents of those parts. Programming the machine defines a new virtual machine which is a more specific version of the original machine. Thus, one sort of transition within the space of behaving systems is from a less to a more specific machine, by replacing variable elements with constant (or more constrained) elements. (The change need not be permanent, if the resulting virtual machine is capable of altering its own program later.)

This process of specification may be contrasted with implementation. One virtual machine can be implemented (or embodied) in another virtual machine which satisfies far more detailed state descriptions. The same virtual machine can be embodied in different physical machines. Different virtual machines can be embodied in the same physical machine.

Virtual machine A is implemented in B by specifying a mapping M, where M maps descriptions of the states of B into states of A, and maps 'legal' transitions of B into primitve legal transitions of A. Thus A may only refer to integers as possible values of variables, whereas B may have far more memory locations each capable of having a binary state. A collection of bits of B may be taken to represent a single integer of A. It is not necessary that every state of B correspond to a state of A. For instance, the process of giving a variable of A an integer value may involve changing several locations in B, in sequence, and some of the intermediate states of B need not represent any state defined in A. This also illustrates the fact that a primitive (unanalysable) transition of A may be represented by a succession of state transitions of B.

Totally different languages may be required to describe the two machines. For instance machine A may be described as containing sets of words and performing operations like sorting, forming intersections and unions of sets, etc., whereas the language for describing B makes no mention of sets or set operations but refers only to chained pairs of memory locations and operations such as comparing or altering the contents of individual memory locations.

Machine B implements A relative to a mapping M. Relative to a different mapping, B may implement a different machine. Usually implementation is done by taking a fairly general machine B, then producing a more specific machine B1 by programming it ('specification'). B1 implements A1, relative to mapping M1. A different specification of B may produce a machine B2 which implements another machine A2 via a different mapping M2.

```
B  --specify (by program P1)--> B1
                                 |
                                 --implements(via M1)--> A1

B  --specify (by program P2)--> B2
                                 |
                                 --implements(via M2)--> A2
```

(A Turing machine can be used to implement any other type of discrete machine, by suitable programming. While theoretically interesting this fact is not of much practical significance, for it doesn't tell us how to solve real design problems. Similarly the fact that all of number theory can be reduced to a few axioms doesn't help the number theorist solve most of his problems.)

We have already seen that the mapping of B onto A need not be complete: not all states of B are mapped. The mapping need not be unique either. There may be several different states of B which are all mapped onto the same state of A, for instance if A represents numbers with lower precision than B. Moreover there may be different ways of consistently implementing A in terms of B, via different programs and mappings.

Just as machine A is implemented in B, so can B be implemented in machine C. There may be several levels of implementation. An interesting question is to what extent and in what way either learning processes within an individual or evolutionary processes can bring it about that the number of layers of implementation increases.

The relation of implementation is not an intrinsic relation between two virtual machines, but is relative to a mapping <M1 or H2 above). To this extent it is like the relation of meaning or representation. Two observers of B1, using different mapping rules, may take it to implement different machines A11 and A12, just as two perceivers of a sentence or a picture may arrive at different interpretations. In practice there are not likely to be interestingly different rival interpretations of the same complex system. (Interpretations are rivals if they use the same level of description.) Different rival interpretations are unlikely because the space of legal descriptions is only sparsely inhabited by interesting machines. E.g. in the space of *legal* programs expressible in a given programming language only a small subset will be of any use.

There is, however, an interesting example of such an ambiguity. It is not obvious how one can decide whether a machine with boolean operations uses 1 for true and 0 for false, or the other way round, with •and^1 swapped with ^{f}or^f etc. To get a unique interpretation in terms of such notions we may need to embed the simple logical machine in the context of a larger system which makes *use* of the logical operations.

In the case of systems as complex as a human brain it may well be the case that some parts of the system actually interpret others in accordance with rules. Thus different brains may use different interpretation rules, so that in one brain structures and processes B1 are used to represent or implement a certain higher level machine A, whereas in another brain B2 is used to represent A. Each brain may include an explicit specification of some of the mapping rules it uses. Functionally then the two brains can each have a subsystem implementing a virtual machine A which is not dependent on the subjective interpretations of external observers. But in the two brains (or computers) the physical processes corresponding to operations of A may be utterly different because of the different mapping rules.

Some philosophers like to discuss the implications of hypothetical experiments in which neurosurgeons discover precise correlations between physical brain states and mental states. For instance if brain state Sb correlates with mental state Sm would that mean that Sm really *is*. Sb? What we now see is that if the actual relation between minds and brains is something like the implementation relation, which is relative to a mode of interpretation, and especially if the implementation goes through many layers of virtual machines, then it is extremely unlikely that there will be any discoverable correlations between physical brain states or processes and mental states or processes, just as there will be no empirically discoverable correlations between physical processes in a modern multiprocessing computer and such high level processes as parsing sentences of English. There are far too many different ways in which such high level processes can be mapped onto the physical processes: if we include different states of the virtual memory system in a modern computer then the number of possible mappings will be *astronomical*. In other words, many pages of philosophical writing have been wasted on questions which fail to address the real issues. Questions like whether the abstract machine A implemented on physical machine B, possibly via many intermediate layers, really IS B, or whether the state Sa of A which exists at time t really IS the state Sb of B which exists at time t, are non-questions. The ordinary conception of identity was not designed to cope with such complex relationships as this: the ordinary meaning of 'is' provides no criteria for settling the issue. Moreover, nothing hangs on how it is settled. If we know that the relation is one of implementation what more is there to know? (This is not quite like the question whether the axe I have now really IS the same axe as the one I had a year ago if, in the meantime, I have broken and replaced the handle and then broken and replaced the head. Again there are no criteria built into the meaning of 'is' to answer the question. But something may hang on the question, namely e.g. whether the insurance cover I bought for the original axe still applies to this one. That question can be resolved by taking legal decisions, which is not the same as discovering the 'true' answer.)

Different implementations of the same machine

Work in AI has shown that some virtual machines implemented in terms of stored-program digital computers can produce behaviour which previously had been associated only with minds of living things; for instance such behaviour as producing or understanding language, solving problems, making and executing plans, learning new strategies, playing games. Human and animal brains appear to be very different from such computers. What sorts of virtual machines they implement, and how they implement them, remains an important unsolved problem. It may be that the study of 'connectionist' machines will shed new light on this question. (ref: Ballard, Feldman, Hinton ???)

Different ways of implementing the same virtual machine may have very different implications for learning or evolution. For instance, most AI programs are implemented on computers in such a way that a change to a more powerful system requires considerable alteration, such as adding new complex rules and possibly deleting old ones, or even changing the program architecture. Thus we have a virtual machine A implemented in terms of B in such a way that the transition from A to A' requires very complex changes from B to B', where most of the intermediate states between B and B' would not implement any sensible higher level machine. I.e. some of the minimal meaningful transitions may be very complex. But it is possible that if A were implemented in an entirely different way on machine C then the transition from A to A' could be accomplished by a transition from C to C' through a succession of small steps C, C1, C2, ... Ci ... Cn, C', such that each Ci implements a working system at the level of A and A'. The sequence of high level virtual machines A, A1, ... Ak, A' may be much shorter than the low level sequence (i.e. k << n), if several intermediate machines at the level of C correspond to the same machine at the level of A. In that case a lot of relatively smooth changes at the low level could correspond to a few big jumps at the higher level. If this were so then there would be relatively unstable states of the high level machine without any instability at the low level. A simple but uninteresting example of this is the use of thresholds on quantitative variables.

Combining the best of all philosophies

By using these ideas we can combine the advantages of both behaviourist and mentalist approaches to the study of the mind. The main strength of behaviourism, in all its forms, is that minds are not static things - it's what they do that is so important. But behaviourists tend to be suspicious of talk of internal processes whereas we have learnt that, as mentalists stress, not all important doings are external, some i internal: state changes within a machine.

It is even quite possible for the internal processes to be too ri to be revealed by external behaviour, so that in an important ser external observers cannot know exactly what is going on. This is partial explanation of the old idea that each individual has access tc rich internal world which is not externally observable. There e several different sorts of reasons for this. (Compare my C19783 ch 10,

(1) Tracing everything may lead to infinite regress.
A computer program may be able to print out 'tracing' informati reporting some of its internal states, but the attempt to trace 1 internal processes which produce trace printing can lead to infinite regress.

(2) External behaviour may not be able to keep up with interr behaviour.
Consider a computing system with television camera perform- complex and detailed analyses on large arrays of visual data, I with limited capacity 'output channels' so that any attempt report current visual processing will inevitably get further i further behind. Here perhaps is one root of the sense of a rich I inaccessible inner experience.

(3) Lack of self-monitoring facilities.
One of the important types of transitions in the space of behav systems concerns the difference between systems with and withe various kinds of self-monitoring. The self-monitoring may incomplete. Thus a certain machine may simply lack the ability monitor some of its own internal processes. This is certainly ti of many human mental processes.

(4) Lack of appropriate descriptive language.
More interestingly, the system may not be able to monitor cert processes at a certain level of description. For instance if virti machine A is implemented in terms of B, in such a way that a sin; operation of A corresponds to many operations of B, and a sine object of A corresponds to many objects in B, then there may mechanisms which describe the processes at the level of B, with being able to recognise which events at the level of A occurring. An extreme example would be the neurophysiolog observing the firing of neurones in complete detail without know which higher level virtual machines were using the neuronal mach as an implementation. Alternatively, the system may be able describe some of the processes at a high level of abstraction I not able to represent all the rich detail, e.g. because monitoring processes cannot get at some of the detailed workings the perceptual processes. A simple case of this is where some of i details are 'hard wired'. But there are other cases. Many compute cannot examine their own microcode. Furthermore, although they i be made to monitor certain classes of events, e.g. certain sorts procedure entry or exit, a single processor cannot monitor record every one of its machine instructions, though a sec (possibly much faster) processor could.

This is not a complete list of possibile reasons why self-monitoring may be limited. What it indicates is that there may be several different facts about how a mind is implemented which may separately contribute to the old idea that mental processes are peculiarly indescribable, and externally inaccessible.

Levels of exploration of the space of behaving systems

I stated above that we need to relate the structures of mechanisms (their syntax) to their behaviour or functions (their semantics). This can be related to a distinction between two sorts of tasks for the empirical study of behaving systems: descriptive and explanatory. The _descriptive_ task is empirically to survey and classify the kinds of things different sorts of actual behaving systems can do. This is a classification of different sorts of abilities, capacities or behavioural dispositions - remembering that some of the behaviour may be internal, for instance recognising a face, solving a problem, appreciating a poem. Different sorts of machines, animals, people, can then be described in terms of what they can and can't do.

The _explanatory_ task is to survey different sorts of virtual machines and to show how their properties may explain the abilities and inabilities referred to in the descriptive study. So descriptions give an account of behaviour whereas explanations indicate what the underlying mechanisms are. They may also explain transitions from one sort of system to another, e.g. during individual development.

The explanatory task overlaps with the descriptive task in that explanatory models will have to include descriptions of some of the capabilities of the components. These capabilities will themselves require explanation. Thus we have a recursive process of explanation and description corresponding to the idea of layers of implementation mentioned above.

Examples of divisions in the space

I have suggested that instead of one major boundary between things with and things without minds we can expect to find many different transitions, at different levels of abstraction, in the space of behaving systems. I would like to be able to provide an elegant theoretical overview of the set of such transitions, or the structure of the space, but cannot (yet). Nevertheless we can list some distinctions whose importance has emerged in recent years.

• Some systems, like a thermostat, have only quantitative representations of states, processes etc. For instance, a very simple organism may be able to measure temperature, or the density of useful chemicals in the surrounding medium. Others, like some computer programs and perhaps people, can build structural descriptions, like parse-tree representations of sentences or chemical formulae.

* A closely related distinction can be made between systems whose internal processing consists only of continuous variation of quantitative measures and systems which in addition can perform a variety of discrete operations on discrete structures, e.g. matching them, rearranging them, storing them in a memory etc. (This should not be confused with discontinuous jumps in values of scalar variables, as in catastrophe theory.)

* Some computational systems can construct formulae of predicate calculus and perform logical inferences. Other systems lack this ability.

* More generally, we can distinguish systems according to the types of symbolisms they are able to use for a variety of internal or external purposes, such as storing information, formulating goals, making inferences, storing plans or strategies, communicating information or questions, etc. to others, and so on. There may be some limitations which are inherent in the physical architecture of the systems (e.g. they do not have sufficient internal variability to support certain sorts of languages) whereas other limitations may be due simply to how the system has been programmed.

* Some systems (unlike a thermostat, for instance) have the ability to store complex sequences of symbolic instructions. Different sorts of instructions provide different sorts of behavioural capacities. For instance conditional instructions are crucial for flexible, context sensitive performance. Instructions for modifying stored instructions may play an important role in learning processes.

* Some systems, like conventional digital computers, can essentially do only one thing at a time, albeit very quickly in some cases. Others are parallel machines. The study of different sorts of parallelism and their properties is now in its infancy. One consequence of certain sorts of parallel architecture is the ability to monitor (internal or external) behaviour while it is being produced. It also permits •postponed[1] conditional instructions of the form 'If ever X occurs do Y[1]. This seems to be crucial to many features of human and animal intelligence.

* Some parallel systems are composed of a network of serial machines whereas others are massively and fundamentally parallel in that they consist of very large collections of processing units, no one of which performs any essential computing function. What would normally be thought of as a computational state is distributed over large portions of the network. The implications of this sort of distinction are at present hardly understood, though it seems clear that at least the more complex animal brains are of the massively parallel type. The gain seems to be that for certain sorts of task, including pattern recognition, very great speed can be achieved, along with the ability to generalise from old to new cases and to degrade gracefully as input information degrades. Other sorts of task, for instance long chains of deductions, may only be achieveable on this sort of machine by indirect, clumsy and unreliable strategies. We see here an echo of the current fashion for distinguishing left and right brain activities: except that both halves of the human brain seem to be massively parallel systems.

* Some systems merely perform internal manipulations, except possibly for receiving some input to start things off and producing some output at the end. Others are linked to sensors which continuously receive information from the environment, which affects the pattern of internal processing. The •environment[1] may include the physical body in which the virtual machine is instantiated.

* Some systems are embodied in a complex physical machine with many sensors and motors which are controlled to perform complex actions in the environment. Others must merely react internally to what the environment offers, like a paralysed person.

* Some perceptual mechanisms essentially only recognise patterns in the sensory input. Others interpret the input by building descriptions of other things which may have produced the input. Thus two-dimensional images may be interpreted as produced by three-dimensional structures, and various forms of observable behaviour may be interpreted as produced by unobservable mental states in other agents. So some systems can represent only observable or measurable properties and relations between things, whereas others can construct hypotheses which go beyond the given, including hypotheses about causal connections between observable events. In particular, some systems can postulate that other objects may themselves be agents with internal programs, motives, beliefs, etc., and take these internal states into account in their own planning, perception, etc.

* Some systems have a fixed collection of programs, whilst others have the ability to reprogram themselves so as radically to alter their own abilities - possibly under the influence of the environment.

* Some systems, especially applications of AI, are essentially presented with a single goal at a time, from outside, and all they can do is pursue that goal and sub-goals generated by it. Other systems, notably living organisms, have a motley of 'motive-generating' mechanisms so that current motives, preferences, principles, constantly need to be re-assessed in the light of new ones which may have nothing to do with previous ones. This seems to be another of the computational properties underlying the ability to have emotions.

* Some systems have a fixed set of motive generators, whereas others may have motive-generator-generators. Can this hierarchy be extended indefinitely?

* Some systems can select goals for action, yet postpone action because there will be better opportunities later. Others can only act immediately on selected goals. The former need databases in which postponed goals and plans are stored, and monitors which can react to new opportunities. This ability to postpone intended actions would seem to be one of the differences between more and less sophisticated animals, and perhaps between human infants and adults.

* Some systems, once they have begun to execute a plan or program cannot do anything else, whereas others can, where appropriate, interrupt execution, and switch to another plan if necessary, and then continue execution of the original later, if appropriate. This requires mechanisms for storing what has been done so far and some indication of where to continue an interrupted plan.

* Some systems can only be interrupted at pre-determined stages of processing, e.g. when they explicitly pause to examine new input. Others allow asynchronous interrupts: i.e. monitoring of interrupt conditions is done in parallel with the main activity, possibly by a second processor which can stop and re-direct the main one. The use of asynchronous interrupts allows programs to be simpler and more general, but also makes it harder to predict exactly what will happen when a program runs.

* Properly assessing new information may cause dangerous disruption of current activities. So an intelligent system may have a way of using a relatively low level system do some preliminary filtering of interrupts,

in parallel with and independently of the main activities, so that only relatively important interrupts disturb higher levels. In Sloman and Croucher [1981] it is suggested that this sort of mechanism, essential for intelligent systems with multiple motive-generators in a complex and fast moving environment, may account for many emotional states.

* Some systems can monitor only the subsequent effects of their actions, e.g. a thermostat. Some can monitor the behaviour itself, e.g. placing a paw carefully on a potentially dangerous object. Some can monitor internal as well as external processes, for instance a computer checking which of its routines are used most frequently, or a person detecting and classifying some emotional state. Different kinds of monitoring provide different opportunities for self-assessment, self-modification, self-understanding.

These are merely examples of some of the more obvious discontinuities in the space of possible explanatory mechanisms - virtual machines. Although the descriptions are general and vague, it is already clear how we can design machines which illustrate most of these distinctions, at least in relatively simple tasks. We don't yet have a full understanding of all the different ways of doing this, nor what their implications are. Moreover, many more detailed distinctions need to be explored -- distinctions between sorts of languages, sorts of operating systems, sorts of algorithms, sorts of data-structures, sorts of computer architecture.

In terms of such mechanisms, we can begin to account for different abilities found in human beings and other animals, as well as constructing machines which display such abilities. In terms of the sorts of differences alluded to here we can find objective reasons for describing some systems but not others in intentional terms, i.e. saying that they have believes, goals, knowledge, skills etc. This is in contrast to the view of Dennett (1978) that taking up the 'intentional stance' in describing a system is a matter of convenience. Further analysis would also show, I believe, that intentional descriptions can be justified by the type of computational architecture in a system, i.e. the sorts of functional divisions of different data-bases and mechanisms which operate on them, independently of how *rational* the system is in its actual use of such mechanisms. In short, in contrast with Dennett I claim that we have to adopt what he calls the 'design stance' as a basis for adopting the 'intentional stance' in any systematic study of behaving systems. (This point needs further elaboration.)

What we still need to do is explore which combinations of mechanisms are required to account for the characteristically human abilities which have puzzled philosophers and psychologists and provide much of the motivation for research in AI. A tentative list of such characteristics in need of explanation follows:

Salient features of intelligent systems.

What follows is an attempt to describe, at a very general and abstract level, the union of the kinds of abilities which people in the field of AI have begun to try to understand and replicate. This gives a very rough and provisional characterisation of an intelligent system as one which has some combination of the features listed below. It is perhaps worth stressing that the list reflects the spread of research which is already in progress, though not all aspects have been pursued to the same depth. Thus, the list represents, from an AI viewpoint, an answer to the question: what are the features of human beings (and some other animals) which make them different from inanimate mechanisms and unintelligent plants and animals?

Characteristics of intelligent systems: a tentative overview

(The order is not significant.)

- Having a *general* range of abilities, including
  (a) the ability to cope with *varied* objects in a domain
  (b) the ability to cope with *a variety* of domains of objects
  (c) the ability to perform a *variety* of tasks in relation to any object,
  (d) the ability to recognise which sub-ability to use.

  NOTE: 'object' here is a neutral term, covering such diverse things as physical objects, spoken or written sentences, stories, images, scenes, mathematical problems, social situations, programs, etc. 'Coping' includes such diverse things as perceiving, interpreting, producing, using, acting in relation to, predicting, etc.

- Various forms of discovery, learning, or self-improvement, including: qualitative extensions to new domains, new kinds of abilities, and quantitative improvements in speed of performance, complexity of tasks managed, etc. Important special cases include the discovery of new concepts, heuristics or generalisations within a domain, the creation of new domains, and the novel combination of information about several different domains to solve a new class of problems. The more complex examples overlap with what we ordinarily refer to as 'creativity'.

- Performing inferences, including not only logical deductions but also reasoning under conditions of uncertainty, non-monotonic reasoning (e.g. making use of implicit assumptions which may be cancelled by new information), reasoning with non-logical representations e.g. maps, diagrams, networks.

- Being able to communicate and co-operate with other intelligent systems, especially human beings.

- Being able to co-ordinate and control a variety of sensors and manipulators in achieving a task involving physical movement or manipulation.

- Coping flexibly with an environment which is not only complex and messy, but also partly unpredictable, partly friendly, partly unfriendly and often fast moving. This includes the ability to interrupt actions and abandon or modify plans when necessary, e.g. to grasp new opportunities or avoid new dangers.

- Self-awareness, including the ability to reflect on and communicate about at least some of one's own internal processes. This includes the ability to explain one's actions.

- Coping with a multiplicity of "motivators", i.e. goals, general principles, preferences, constraints, etc. which may not all be totally consistent in all possible circumstances. This need can arise either because a single high-level goal can generate a multiplicity of inter-related sub-goals, or because a system has a collection of independent sources of goals, requirements, etc.

- Having motivator-generators and motivator-modifiers. I.e. being able to change the collection of goals, preferences, principles, etc. which guide decision making.
- The ability to generate, or appreciate, aesthetic objects. This is often thought of as distinct from cognitive abilities, but there are reasons for thinking that aesthetic processes are involved in many cognitive processes, and vice-versa. E.g. elegant proofs not only give pleasure: they generally provide more insight than messy ones.

The notion of intelligence is bound up not only with what can be done, but also with how it is done (i.e. the style, or manner). For example:

- When confronted with messy, ill-defined problems and situations, and incomplete or uncertain information; an intelligent system should degrade gracefully as the degree of difficulty/complexity/noise/incompleteness etc. increases, rather than merely 'crashing', or rejecting the problem. Degrading gracefully may involve being slower, less reliable, less general, less accurate, or producing less precise or complete descriptions etc.
- Using insight and understanding rather than brute force or blind and mechanical execution of rules, to solve problems achieve goals, etc. E.g. instead of exhaustive trial and error searching there should be selection of alternatives based on some analysis and description of the current state of a problem-solving process. This is closely connected with a requirement for speed and generality.
- Plans should not be created simply by applying pre-defined rules for combining primitive actions to achieve some goal, but should rely on the ability to use inference to answer hypothetical questions about 'What would happen if..'. This should also play a role in the ability to make predictions, or test generalisations.
- Conflicting goals should not be dealt with simply by means of a pre-assigned set of priority measures, but for example by analysing the reasons for the conflict and making inferences about the consequences of alternative choices or compromises.
- Unexpectedly good or bad performance should feed back into a learning process.

These lists are not proposed as a definition of 'intelligence'. The list merely summarises salient aspects of the most intelligent systems we already know, namely (adult?) human beings. Having compiled a list of features of intelligent systems, we can then move on to ask what underlying mechanisms or capabilities may be required for the production of these features.

No existing AI system fulfils even a subset of these criteria, except in very restricted domains, with rather generous interpretations of concepts like 'generality', 'graceful degradation', 'flexibly', etc. Nevertheless there are many examples of fragmentary progress.

There is no sharp boundary between such work and other fields of computer-science and engineering, and perceived boundaries change as our understanding deepens. For instance compilers capable of accepting algebraic expressions were once thought of as intelligent because previously only human beings had been able to do such things. It is to be expected that as our understanding and technical achievements progress so will the boundary between what we do and don't regard as intelligent change.

There is still a lot more to be done to discover precisely what sorts of computational and represetational mechanisms are capable of accounting for what sorts of abilities.

Conclusion

Instead of arguing fruitlessly about where to draw major boundaries to correspond to concepts of ordinary language like 'mind' and 'conscious', we should analyse the detailed implications of the many intricate similarities and differences between different systems. To adapt an example of Wittgenstein's: there are many ways in which the rules of a game like chess might be modified, some major some minor. However to argue about which modifications would cause the essence of chess to be lost would be a waste of time, for there is no such thing as the essence. What is more interesting is what the detailed effects of different modifications would be on possible board states, possible strategies, the difficulty of the game etc. Similarly, instead of fruitless attempts to divide the world into things with and things without the essence of mind, or consciousness, we should examine the many detailed similarities and differences between behaving systems.

This is a multi-disciplinary exercise. Psychologists and ethologists can help by documenting the characteristics of different types of systems to be found in nature (e.g. Lorentz 1977), including the many detailed differences between humans of different ages, and the results of various types of brain damage, which produce systems not normally found in nature. Anthropologists can help by drawing attention to different sorts of minds produced by different cultural contexts. Linguists and other students of the structures perceived and produced by human minds can help to pin down more precisely what needs to be explained. Computer scientists can help by proposing and investigating detailed mechanisms capable of accounting for the many kinds of features of human minds, animal minds, robot minds. Philosophers can help in a number of ways. They can analyse the many complex implicit assumptions underlying ordinary concepts and thereby help to indicate what exactly it is that we need to explain: for instance those who start from an over-simplified analysis of consciousness or emotion concepts will over-simplify the explanatory task. More generally, a philosophical stance is needed to criticise conceptual confusions and invalid arguments, and to assess the significance of all the other work. E.g. does a computational model of mind really degrade us, as some suggest, or does it reveal unsuspected richness and diversity?

By exploring the structure of the space of possible mental mechanisms we may achieve a deeper understanding of the nature of our own minds, by seeing how they fit into a larger realm of possibilities. We may also hope to get a better understanding of the evolutionary processes which could have produced such minds. We will learn that there is neither a continuum of cases between ourselves and a thermostat or amoeba, nor an impassable gulf either.

Such a study should be of interest to engineers trying to design behaving systems, as it will help to improve their understanding of available options. It is of interest to psychologists and ethologists, as providing a conceptual framework for describing differences and similarities between different organisms, but also for describing behavioural development within individual organisms. And it is crucial for evolutionary biology that we develop a theory concerning which sorts of transitions between behaving systems are possible, for fossils can tell us little, if anything, about the behaviour of organisms, especially their internal behaviour, or the details of brain structures. Therefore, without a rich theory constraining hypotheses, speculation about the evolution of behaviour and mind is likely to be totally undisciplined.

So much for methodology. The really hard and interesting work remains to be done.

References

TO BE EXTENDED

Boden, M.A. 'Artificial Intelligence and Animal Psychology', in *New Ideas in Psychology*, Vol 1. No. 1, 1983 (Pergamon Press).
Dennett, D.C. *Brainstorms*, Harvester Press 1978.
Feldman, J. and Ballard D. ???? in *Cognitive Science* ????
Hinton, G. and J. Anderson *Parallel Models of Associative Memory*, Earlbaum 198??
Lorenz, K. *Behind the mirror*, Methuen, 1977.
Ryle, G. *The Concept of Mind*, Hutchinson, 1949. (Also Penguin Books).
Sloman, A. *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind*, Harvester Press and Humanities Press, 1978.
Sloman, A. and M. Croucher, 'Why robots will have emotions', in *Procedings 7th IJCAI*, Vancouver, 1981.