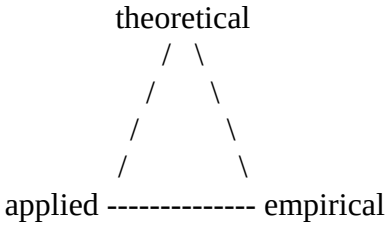
In *Artificial Intelligence - Human Effects,* eds. M. Yazdani and A. Narayanan, Ellis Horwood, 1984, pp 173-182.

Towards a Computational Theory of Mind

Aaron Sloman
Cognitive Studies Programme,
School of Social Sciences,
University of Sussex,
Brighton, BN1 9QN
At University of Birmingham since 1991
http://www.cs.bham.ac.uk/~axs

1. Introduction

Cognitive Science has three interrelated aspects



Work in all three areas depends on and feeds back into the other two. Theoretical work explores possible computational systems, possible mental processes and structures, attempting to understand what sorts of mechanisms and representational systems are possible, how they differ, what their strengths and weaknesses are, etc. Empirical work studies existing intelligent systems, e.g. humans and other animals. Applied work is both concerned with problems relating to existing minds (e.g. learning difficulties, psychopathology) and also the design of new useful computational systems. This paper sketches some of the assumptions underlying much of the theoretical work, and hints at some of the practical applications. In particular, education and psychotherapy are both activities in which the computational processes in the mind of the pupil or patient are altered. In order to understand what they are doing, educationalists and psychotherapists require a computational theory of mind. This is not the dehumanising notion it may at first appear to be.

2. A theory of what?

Concepts of mind and consciousness are complex, partly ill-defined, and their use very much context dependent. For instance, it is quite correct, for normal purposes to say of someone who is asleep

‘‘He is now unconscious’’

whereas the very same person may later say, quite correctly,

‘‘In my dream I was conscious of a pain in my leg’’

Neither speaker is in error. They are providing information of different sorts, about different aspects of the sleepers’ mental state. Similarly there is the well known puzzle of the noise you only notice when it stops: were you conscious of it before, and if you were not how could you hear it stopping? If someone says:

‘‘I’ve been conscious for the last three months that you are hoping to take over my job’’
this does not mean that the speaker has been dwelling constantly on the threat. He is merely talking about information that has been *available* to his thinking and decision making processes, whether or not he actually made use of the information.

These and other examples indicate that the ordinary concept ‘‘conscious’’, and a host of related concepts (notice, aware, attend, study, see, feel, think, consider, careful, etc, etc) are very complex and subtle (Sloman, 1978, Ch 4). In particular, there is no reason to suppose that there is a well understood sense of ‘‘consciousness’’ in which consciousness is something you either definitely have or definitely lack. The fact that we have the *noun* ‘‘consciousness’’ can be a trap which leads us to think there is some *thing* to which it refers.

Consciousness is often confused with various kinds of self-consciousness. It is unlikely, for example, that a cat is self conscious in the sense of being able to reflect on and reason about its own states, abilities, prospects, character, etc. However, there is no doubt that it is conscious of the mouse that it chases, or the person it rubs itself against. Moreover, an animal which uses a paw to catch or manipulate something, must be conscious of the relationship between its paw and the object, in order to judge the appropriate movements to make. And insofar as it treats the paw differently from the other object, it can be said in some sense to be aware that the paw its its own. Here then is a tiny fragment of self-consciousness. No doubt many animals are self-conscious in this sense even if they are not able to pass the ethologist’s test of being able to respond to themselves in a mirror.

My point is that it is an error to think of familiar mental concepts as marking states or properties or processes which are clearly distinguishable from other states, etc. The world is too rich and varied for our ordinary concepts. Though complex and very subtle (Sloman 1978, chapter 4), our ordinary concepts are probably too ill-defined and unrefined to provide an adequate basis for a systematic study of minds. We need a better system of concepts if we are to proceed beyond banalities, to explanatory insights concerning the different sorts of states and processes which characterise minds. (This is not to say that we should ignore common concepts. As J.L. Austin said, they may not be the last word, but they are certainly a good starting point.)

3. A survey of possible types of minds?

One way of trying to make progress is to attempt a survey (a generative taxonomy perhaps) of different possible types of mind. This can help to undermine preconceptions about what minds have to be like, showing where we have been thought-prisoners of concepts which evolved not for the purpose of providing deep explanations, but for the purposes of everyday interactions.

We start from the assumption that if minds are anything other than totally unintelligible mysteries, they must be mechanisms. (Inside every intelligent ghost there must be a machine.)

The question is what sort of mechanism? Whether the mechanism is physical or spiritual (whatever that means) is not important. Both alternatives lack explanatory power. We want an explanation capable of generating the fine structure of the phenomena of mind: the ability to perceive, to learn, to decide, to have feelings and emotions, to dream, to find things funny, etc.

To say what the mechanism is made of, is not to say how such states and processes can occur. For that we need an account of the architecture and laws of the mechanism: how the parts are related to one another, how their properties and relationships can change. I.e. it’s the structure not the stuff that counts.

But not just the static structure: the principles by which the structures can change -- i.e. the rules or constraints governing the processes which can occur in the system -- are crucial. A system of rules also has a structure: a computational architecture. There is an infinite variety of possible architectures, and as yet we understand very little about that variety. We have begun to explore only

the simplest possibilities. We need to work towards a generative specification of mental possibilities, and their limits.

Two fallacies are to be avoided:

(a) The set of possible minds forms a continuum with simple things like amoebas and thermostats at one extreme and human minds (or more complex minds) at the other.

(b) There is no comparison between the simplest mechanisms or organisms and human minds: there is a total discrepancy of kind.

(a) Is a fallacy because the space of possible computational systems (and therefore the space of possible mental mechanisms) is not a continuum. It is full of important discontinuities. Computer scientists have only just begun to understand some of these discontinuities. There is a lot more yet to be learnt.

(b) Is a fallacy, because, despite the discontinuities, there may be a collection of common principles on the basis of which all the different sorts of computational systems are generated. For instance, all involve some kind of storage, construction, or manipulation of internal symbols, and all involve the possibility of treating such symbols as categorical or conditional instructions. Differences concern the number and variety of sub-databases, the types of representations used, the types of inference procedures and other internal symbolic processes used, the extent to which there is internal monitoring of internal processes, the extent to which structures and procedures can modify themselves during the course of interacting with the rest of the world, the extent to which different sub-processes run in parallel, with the ability to interrupt other processes, and so on. We don't yet know enough to have an overview of the space of possible computational systems. E.g. the fact (if it is a fact) that they can all be represented in Turing machines is not much more informative than the fact that all the plays and poems of Shakespeare can be expressed using 26 letters of the alphabet and a few punctuation signs.

The second fallacy is closely connected with the assumption that our everyday concepts can be used for drawing precise boundaries. We assume that among the set of possible systems there must be a sharp boundary between those which are conscious and those which are not, between those which really have experiences, or emotions, and those which do not. In reality the space of possibilities is barely understood and our ordinary concepts did not develop for the purpose of drawing absolute boundaries. Instead of assuming there are boundaries we should map out the terrain, and find out how the different sorts of systems are similar to one another, and how they differ. Wittgenstein somewhere (alas I have forgotten where) compared this with sterile arguments about whether one is still ''REALLY'' playing chess if one player plays without the queen. Apart from tournaments, the only interesting question is what difference the missing queen makes to the game, to the balance of power, to the variety of possible states and strategies. Whether it is still chess is a pseudo question with no right or wrong answer. The same applies to attempts to argue that machines, or some of the simpler animals, cannot really be conscious. The interesting question is: what are the similarities and differences, and what are their implications?

If someone wants to say that the human mind really is totally different from any other sort of mind, or that animal minds are totally different from artificially constructed minds, then either he must specify the differences, and we can argue with him, or else he must rely on the claim that one sort of mind is essentially mysterious and unintelligible, in which case no argument is possible. Someone who takes the latter position can always go on claiming, no matter how rich and powerful our theories, ''But you have still left out something which I experience which I cannot describe, but which I know you know from your experience too''. No doubt some really intelligent, philosophically inclined, robots would also tend to argue like that.

More important than attempts to refute positions which are defined so as to be irrefutable is to get on with the task of understanding the structure of the space of possible mechanisms. In particular, we need a survey of some of the important discontinuities, and an analysis of the most general common principles. (Some are listed below.)

4. What sort of mechanism?

I said earlier it's the structure not the stuff that counts. How can we describe the structures, and the types of internal processes which can occur, in human minds?

To think about symbolic structures and process-rules we need a suitable language, a set of concepts. Until recently the set of concepts available has been restricted to ''common sense'' concepts supplemented by the concepts of science and technology. E.g. Freud made considerable use of hydraulic analogies. Common sense makes use of many analogies from mechanics (''I was pulled in two different directions'', ''I was balanced on a knife edge''. etc.) Piaget used analogies deriving from biological studies.

Recently the concepts of computing science have considerably extended our ability to discuss mental processes. We need no longer wave our hands and hope to convince colleagues with fine turns of phrase and consistency with common experience. We can actually demonstrate the power of our theories by using them to design *working* minds, or fragments of minds. And like all designers of complex systems we shall learn from our failures.

Computing concepts are essentially derived from a small number of basic assumptions about the nature of a symbol manipulating system:

(a) It has a memory able to contain a very large number of independently variable symbols. This means that the set of possible states it can be in, and the set of possible transitions between states is astronomical, or even bigger! E.g. if there are just twenty symbols, which can take one of two forms, e.g. "A" or "B", then the number of possible states is over a million. If there are millions of symbols, as in modern computers, the number of possible states becomes unimaginably large. This gives enormous scope for individual variability, in physically similar machines.

(b) The symbols can be stored, searched for, compared, removed, altered, and new ones can be constructed. This means that such a system can include a large and *changing* repository of information.

(c) The symbols can be interpreted as *instructions* i.e. they can control behaviour, both internal and external. This means that the system can generate behaviour, and be self-controlled. It also means that it can change its own programs.

(d) Some of the instructions are conditional, where the conditions may be internal or external. This allows for adaptable and intelligent behaviour, and learning which is very much dependent on environmental influences, so that initially identical systems can diverge rapidly, with positive feedback.

(e) Some of the symbols can be made to correspond (according to suitable rules) with information flowing into the system via cameras and other sensors, and can be used by conditional instructions. This means that the system can treat its symbols as representing *beliefs* about the world.

(f) Besides the lowest level symbolic instructions which directly cause processes to occur, the use of symbols with meaning allows instructions, like assertions, to refer to an external world. Instructions can then be of the form:

''Make it the case that .....''

In other words, a computing system can be *goal-directed.*

(g) Instead of being directly executed, instructions themselves can be examined, compared and analysed. In particular, goals can be compared, inconsistencies detected, priorities decided on. Goals may even be rejected, in the light of higher level goals. Further the process of acting on goals can lead to the formation of new sub-goals. (We'll see that the need to be able to cope with multiple goals can lead to the development of mechanisms which can produce emotional states.)

It follows from all this that different computational systems with the same initial configuration can be placed in slightly different environments and over time develop in quite different directions, so that their common origins are totally beyond recognition. In particular, such a system may be able completely to obliterate all the instructions that it started off with as it gradually builds up new programs in the light of experience. It then ceases to be doing what it was originally programmed to do. This does not apply to every computational system -- only those which make use of all the possibilities listed above. But our specification is still too general. It covers a very wide range of mechanisms, with many discontinuities between the simplest and the most complex ones we know.

5. What sort of computing mechanism are we?

So far we have answered the question ''What sort of mechanism?'' by saying that a mind is a computational mechanism: a system which acquires, builds, stores, and uses symbols. How can we constrain our study of possible systems of this general kind: given that there are so many sorts, and we know so little about their scope and limitations? One way to proceed, is to pretend to be God, or evolution, designing something like human beings. We can then specify the design task, and try to determine how far this constrains the possible design solutions. In particular, we can distinguish design considerations arising out of the following:

(a) The nature of the environment (e.g. its complexity, variety, unpredictability, and mixture of friendliness and unfriendliness). These impose constraints on the types of perceptual systems required, the kinds of belief representations, the kinds of planning and executing mechanisms, the kinds of learning mechanisms, etc.

(b) The fact that the creature to be designed will inhabit and have to control a fairly complex, yet fragile, body, with changing needs. For instance, this implies a need for body monitors able to feed information to decision makers. It also implies a need for very complex subsystems able to control and co-ordinate independently moving parts.

(c) The need to be part of a social system capable of adapting to changes of many kinds. This requires individuals to be able to acquire new forms of knowledge (e.g. new concepts, new languages) and to modify some of their rules of behaviour to cope with changing social needs. If we are designing something that will work together with others like it, we may have to provide an ability to have and act on unselfish motives.

(d) The need to be able to cope with relatively helpless young. This implies that the adults need to be provided with motives or motive-generators which are essentially unselfish. The young will need to have forms of behaviour which cause these motives to be acted on.

(e) The need to cope with a relatively large number of changing goals, principles, ideals, preferences, likes, dislikes, not all mutually commensurable, not all simultaneously satisfiable. This implies a need for motive-comparators, including strategies for deciding between incommensurable alternatives! It also requires the ability to take decisions concerned with long-term as well as short term activities and ends, and the ability to ignore

or suppress some motives or needs in the light of others.

These design considerations suggest the need for certain sorts of computational architectures. For instance, there will be a complex set of not necessarily mutually consistent, nor mutually commensurable, motives (desires, wants, fears, etc.), and this leads to a requirement for mechanisms for comparing and choosing between different motives, possibly using rules of thumb developed on a trial and error basis. (E.g. if you often find that choosing A rather than B leads to failure on task A, whereas trying to do B enables you to achieve A as a side-effect, then a good rule is to choose B, even if it does not have greater a priori merit.)

The processes of comparison may be arbitrarily complex, based on knowledge of many kinds, possibly acquired over a long time. Sometimes just deciding between goals may itself have to be a complex goal directed activity, with information-gathering sub-goals.

Further, if there have to be different subsystems concerned with controlling different processes (for instance subsystems controlling the movements of different parts of the body, or monitoring and dealing with different sorts of needs), then since their goals and subgoals will sometimes be inconsistent, there is a need for some mechanism which can resolve conflicts, and control the global behaviour of the system. I have argued in Chapters 6 and 10 of Sloman (1978) that some aspects of our concept of consciousness can be related to this idea of a sort of centralised arbitrator, or administrator, controlling a number of relatively independent parallel subsystems. Not all the internal information and processes may be accessible or controllable by this ‘‘governing’’ process. Different kinds of inaccessibility may account for different ways in which we may fail to be conscious of something. (Of course this cannot be a complete account of the nature of human consciousness: that depends in many ways on the detailed structure and functioning of our bodies, and on the environment, and what we have absorbed from the culture, including concepts, tastes, personality traits, etc.)

In addition, since the world is not completely predictable, and may sometimes be unexpectedly friendly or unexpectedly unfriendly, there is a need for various monitoring systems which are able to interrupt and redirect or modify other processes. In fact, the issues are very complex. In general, the decision whether to interrupt or modify ongoing processes because of new information may require arbitrarily complex processes of reasoning or problem-solving. This means that in order to decide whether to interrupt, it may be necessary to interrupt anyway. This could be fatal. So it is desirable to develop mechanisms which take decisions of that sort on a relatively automatic, crude ‘‘rule of thumb’’ basis, perhaps using criteria which are partly learned from experience and which are adequate most of the time, even if not always. The detailed development of this suggestion would require a very lengthy discussion of a rather complex collection of interacting sub-processes. We’d also need to see how in certain circumstances it may be necessary for new developments not only to interrupt things in order to cause new decisions to be considered, but sometimes to cause direct and immediate evasive or predatory action. Some opportunities and dangers leave no time for consideration: only reflex action will do.

Full discussion of these ideas also requires elaborating the notions of ‘‘motive’’ ‘‘goal’’ ‘‘preference’’ and related notions. In particular, there are various sorts of higher level motives which are used in the evaluation and generation of motives - I call them motive-comparators and motive-generators. In principle there can be comparators and generators for these also.

All this provides a framework for explaining various kinds of emotional states in terms of the interactions between a variety of active and dormant motives, motive comparators, motive generators, and a collection of beliefs. The essential idea is that emotions arise out of the mechanism which allows relatively unintelligent sub-mechanisms automatically to gain attention, even though that may sometimes turn out to be incompatible with higher-level goals. Thus emotions have an involuntary aspect: they are hard to suppress. The classification of different sorts

of emotions arises out of the different patterns of relationships between motives and the processes they can generate.

The existence of the monitoring processes with their ability to modify other processes may, for example, account for the ways in which all sorts of apparently extraneous aspects of a situation can interfere with communication and learning. In particular, if there is some unfulfilled or threatened goal or motive which constantly interrupts central planning and reasoning procedures, and diverts attention away from some task, then various forms of learning and reasoning may be seriously hampered.

It can also be argued that there is no way to build a super-intelligent robot, which also copes with a complex set of different sorts of motives, in a partly unpredictable world, without giving that robot mechanisms which are capable of producing emotional states, as a result of performing the cognitive tasks for which they are required. That is, the possibility of having emotions may be a by-product of being able to cope with a complex and partly unpredictable world in an intelligent way. (This does not mean that every intelligent robot will necessarily be emotional, only that it will have the ability -- and abilities are not always exercised.)

More mundane (?) aspects of the design of a mind are concerned with the problems of accounting for memory, visual perception, the understanding of language, problem-solving, concept formation, the development of skills, the execution of plans. Attempting to give computers these abilities teaches us that they involve far more complex processes than we would otherwise have realized. When our programs are simple, it turns out that they don’t do what we had hoped to explain. It follows that tasks performed even by very young children, and many other animals, are extremely complex, requiring very powerful computational abilities. Insofar as we don’t really understand these abilities, much of our educational practice is based on total ignorance of what is really going on in the mind of a child. (The child is equally ignorant, since most of what goes on is below the level of consciousness, for instance the recognition and interpretation of grammatical structures in sentences we hear.)

6. A space-time trade-off

One of the aspects of mind which flows from the need often to take decisions fairly quickly, e.g. before opportunities disappear, concerns the trade-off between space and time. It seems that the human brain is made from relatively slow computational units, although there are very many of them. This means that if recognising dangerous situations, or working out what to do, requires long chains of reasoning from general principles, then, before decisions are taken, death or other disasters may ensue. One way of coping with this problem is to store results of reasoning in some form which makes it possible to access them quickly and use them blindly when they are needed later, without repeating the complex inference processes, i.e. without making use of any ability to understand the rules. This will be specially useful in cases where the process of inference involves a lot of trial and error searching for a successful chain of reasoning. A simple example is the need to store tables of addition and multiplication instead of always working out sums from the most basic principles. This trading of space for time may be a pervasive feature of the way human minds (and perhaps other animal minds) work, over many areas, including language learning, visual perception, many kinds of problem-solving and planning.

All this implies that an intelligent system needs to be partly rigid and rule-bound.

7. Conclusion

We understand very little about mental mechanisms and the kinds of processes they can generate. Work in Artificial Intelligence is making some progress towards the design of systems which exhibit some of the properties of human minds, including some learning abilities. For the time being it is probably wise for psychotherapists, teachers, and all who attempt to study and control mental processes of others, to admit humbly that we don't really know much about what we are doing or why it succeeds or fails.

[See End Notes Below]

Further Reading:

[1] Margaret Boden *Artificial Intelligence and Natural Man,* Harvester Press, 1977.
[2] Douglas Hofstadter, *Godel Escher Bach, An Eternal Golden Braid,* Penguin Books, 1977
[3] Daniel Dennett, *Brainstorms,* Harvester Press, 1978.
[4] Aaron Sloman *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind, Harvester Press, 1978*

More technical:

[5] Patrick Winston *Artificial Intelligence,* Addison Wesley, 1977
[6] R. Wilensky, in *Cognitive Science, 1982.*
(Note added 8 Aug 2012: I cannot find a more detailed reference, and cannot remember what was referred to. Suggestions welcome.)

The opposing viewpoint:

[7] H.L. Dreyfus, *What Computers Can't Do,* (2nd Edition) Harper and Row, 1979.

End Notes, Added 8 Aug 2012

Section numbers have been added, and some minor errors corrected.

I cannot recall what provoked the writing of this paper -- e.g. whether it was merely a response to an invitation from the editors to contribute to the book.

Related developments

There were some important ideas about forms of representation and types of information processing (e.g. in vision) that I had written about in [4, Chs 7 & 9] and in other papers not mentioned here, and further developments of these ideas in papers written after this one, for example ideas about architectural and representational requirements for visual servoing, the relative unimportance of object recognition compared with the importance of perception of possibilities, invariants and a wide variety of types of affordance (proto-affordances, action affordances, vicarious affordances, epistemic affordances, deliberative affordances and others), and the variety of control functions that need to be simultaneously active in a human-like mind. Papers and presentations developing these ideas can be found on the Cognition and Affect Web sites:

http://tinyurl.com/BhamCog http://tinyurl.com/BhamCog/talks

After this paper had been written, fashions grew focusing on a subset of consequences of embodiment and interactions with physical environments, and ideas about mental processes partly implemented in the environment became popular (''extended minds'' -- compare [4, Chs 6 & 7]). Proposers and followers of these fashions tended (a) to notice only a small subset of the relevant phenomena (the ones that supported their ideas) and (b) to ignore the requirements satisfied by some of the computational modelling techniques that they objected to, e.g. requirements for formation of multi-step plans, requirements for geometrical reasoning, and requirements for constructing explanatory theories that go beyond the immediately perceivable environment.