The Structure of the Space of Possible Minds

Aaron Sloman
School of Computer Science, The University of Birmingham
(Written when the author was at the University of Sussex, and published in
The Mind and the Machine: philosophical aspects of Artificial Intelligence,
Ed. S. Torrance, Ellis Horwood, 1984, pp 35-42.)

Describing this structure is an interdisciplinary task I commend to philosophers. My aim for now is not to do it -- that's a long term project -- but to describe the task. This requires combined efforts from several disciplines including, besides philosophy: psychology, linguistics, artificial intelligence, ethology and social anthropology.

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. Machines too are not all alike, even when made on the same production line, for identical computers can have very different characteristics if fed different programs. Besides all these *existing* animals and artefacts, we can also talk about theoretically *possible* systems.

One common approach to this space of possible `behaving systems', to coin a neutral phase, is to seek a single sharp division, between those with minds, consciousness, souls, thoughts, or whatever, and those without. Where to draw the line then becomes a major problem, with protagonists of the uniqueness of man, or of living things, or champions of machine mentality, all disputing the location of the boundary, all offering different criteria for allocating things to one side or the other.

The passion accompanying such debates suggests that more than a search for truth motivates the disputants. To a dispassionate observer such debates can seem sterile.

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, feeling 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.

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 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 my being conscious, they are not themselves consciousness. Consciousness is not anything material.

Yet any other attempt to identify a referent for `mind', `consciousness', `pain' etc. has, until recently, looked like an attempt to populate the world with mysterious, inaccessible metaphysically unjustified entities.

What is different now is that Computing Science has provided us with the concept of *a virtual machine,* within which computational states and processes can occur. A virtual machine has much in common with the kind of formal system studied by mathematicians or logicians. It is an abstract structure which can undergo various changes of state. A virtual machine can be embodied in a physical machine without *being* that machine. The same virtual machine can be embodied in different physical machines. Different virtual machines can be embodied in the same physical machine. Different virtual machines can have very different abilities. Work in Artificial Intelligence has shown that some virtual machines can produce behaviour which previously had been associated only with minds of living things, such as producing or understanding language, solving problems, making and executing plans, learning new strategies, playing games. By studying the space of possible virtual machines we can replace sterile old boundary drawing disputes with a new, more fruitful, more objective investigation.

First we must abandon the idea that there is one major boundary between things with and without minds. Instead, informed by the variety of types of computational mechanisms already explored, we must acknowledge that there are *many* discontinuities, or divisions within the space of possible systems: the space is not a continuum, nor is it a dichotomy.

Secondly, 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 emboldened by the computational analogy we can see that some doings are external, and some internal: operations within a virtual machine. It is even quite possible for the internal processes to be too rich to be revealed by external behaviour, so that in an important sense external observers cannot know exactly what is going on. For instance, a computer program may be able to print out `tracing' information reporting some of its internal `states, but the attempt to trace the internal processes which produce trace printing can lead to an infinite regress. A more interesting example is a computing system with television camera performing complex and detailed analyses on large arrays of visual data, but with limited capacity

`output channels' so that any attempt to report current visual processing will inevitably get further and further behind. Here perhaps is the root of the sense of a rich but inaccessible inner experience which has been the source of so much philosophical argument.

TWO LEVELS OF EXPLORATION

We can attempt a two level exploration of the space of possible minds, one descriptive the other explanatory, though with some overlap between them.

The *descriptive* task is to survey and classify the kinds of things different sorts of minds (or if you prefer 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 recognizing a face, solving a problem, appreciating a poem. Different sorts of minds 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.

These explorations can be expected to reveal a very richly structured space - not one-dimensional, like a spectrum, not any kind of continuum. There will be not two but many extremes. For instance one extreme will be simple servomechanisms like thermostats or mechanical speed governors on engines. Another kind of extreme may be exemplified by the simplest organisms.

EXAMPLES OF DIVISIONS IN THE SPACE

Among the important divisions between different sorts of virtual machines are the following.

- 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 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 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' conditional instructions of the form `If ever X occurs do Y'. This seems to be crucial to many features of human and animal intelligence. When combined with the ability of some sub-processes to interrupt or modify others, we find the beginning of an explanation of certain characteristic features of emotional states.

- 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 generalize 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 achievable 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' 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 passively react to what the environment offers, like a paralysed person.

- Some perceptual mechanisms essentially only recognize 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. Thus some systems can represent only observable or measurable properties and relations between things, whereas others can construct hypotheses which go beyond that given. In particular, some 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 computational systems can construct formulae of predicate calculus and perform logical inferences. Other systems lack this ability.

- 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 Al programs, 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 action 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 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 both sides of each of these distinctions. 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 are being

explored by computer scientists - distinctions between sorts of languages, sorts of operating systems, sorts of algorithms, sorts of data-structures. Eventually we should have a far clearer grasp of the structure of this space, with some sort of global, generative, description of its contents.

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. 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 Al. A tentative list of such characteristics in need of explanation follows:

SALIENT FEATURES OF THE HUMAN MIND

(The order is not significant)

- Generality, 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.

  `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, producing, using, acting in relation to, predicting, etc.

- Being able to co-ordinate and control a variety of sensors and manipulators in achieving a task involving physical movement or manipulation.
- Coping with messy, ill-defined problems and situations, and incomplete or uncertain information; and degrading *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, producing less precise descriptions, etc.
- Various forms of development, learning, or self-improvement, including:
- increases in speed of performance, complexity of tasks managed; qualitative extensions to new domains, new kinds of abilities, etc.

Important special cases include the creation of new domains, and the novel combination of information about several different domains to solve a problem. 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, including reasoning with non-logical representations, e.g. maps, diagrams, models.

- Being able to answer hypothetical questions about `What would happen if . . .?' in order to make plans, make predictions, formulate and test generalizations.
- Using insight and understanding rather than brute force or blind and mechanical execution of rules, to solve problems, achieve goals, etc.
- Being able to communicate and co-operate with other intelligent agents, or take their beliefs, intentions, etc. into account.
- Coping with a multiplicity of 'motivators', e.g. goals, tastes, preferences, ethical principles, constraints, etc. which may not all be totally consistent in all possible circumstances.
- 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. It also includes the ability to behave sensibly when there is no time to collect or analyse all possibly relevant evidence or perform relevant inferences.
- 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.
- The ability to generate, or appreciate, aesthetic objects.
- The ability to experience bodily sensations.
- The ability to enjoy or dislike experiences, to be amused, angry, excited, irritated, hopeful, disgusted, etc.

Although there is no artificial computing system which combines more than a few fragmentary versions of these features, and there is no chance of combining all in the foreseeable future, work in AI suggests that provided suitable hardware and software architectures are used, most or all of these features can be explained in computational terms. (This is by no means established, however). There is still a lot more to be done to discover precisely what sorts of computational and representational 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 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, 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 emotion concepts will over-simplify the explanatory task. More generally, a philosophical stance is needed to criticize conceptual confusions and invalid arguments, and to assess the significance of all the other work. For example, does a computational model of mind really degrade us, as some suggest, or does it reveal unsuspected richness and diversity?

By mapping 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.

So much for exhortation. The hard work remains to be done.