This paper first appeared in **AISB Quarterly** 72, pp 8--14, 1990.
It is based on notes for a Rockefeller-funded workshop on consciousness organised by Daniel Dennett in March 1990, while he was working on *Consciousness Explained*.

REFORMATTED 31 Oct 2015; 6 Nov 2017
Note: I now argue that "conscious/consciousness" is a polymorphous concept exhibiting parametric polymorphism, in
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/family-resemblance-vs-polymorphism.html

NOTES ON CONSCIOUSNESS(*)

Aaron Sloman
School of Cognitive and Computing Sciences University of Sussex
[Now at the University of Birmingham]

(*) This paper was originally a set of notes for a discussion group on consciousness - held in Italy during March 1990 - rather than a paper for publication. The author gratefully acknowledges suggestions made by the editor of *AISB Quarterly,* Steve Torrance, for clarifying its structure.

"CONSCIOUSNESS"

{1} The noun "consciousness" as used by most academics (philosophers, psychologists, biologists...) does not refer to anything in particular.

So you can't sensibly ask how it evolved, or which organisms do and which don't have it.

Some people imagine they can identify consciousness as "What I've got now". Thinking you can identify what you are talking about by focusing your attention on it is as futile as a Newtonian attempting to identify an enduring portion of space by focusing his attention on it.

You can identify a portion of space by its relationships to other things, but whether this is or isn't the same bit of space as one identified earlier will depend on WHICH other things you choose: the relationships change over time, but don't all change in unison. Similarly, you can identify a mental state or process by its relationship to other things (e.g. the environment, other mental states or processes, behavioural capabilities, etc), but then whether the same state can or cannot occur in other organisms or machines will depend on WHICH relationships you have chosen -- and there is no uniquely "correct" set of relationships. This may become clearer below.

{2} There's nothing wrong with the ordinary (non-technical) uses of the word "consciousness", and a host of related words and phrases "attend", "awake", "aware", "consider", "detect", "discern", "enjoy", "experience", "feel", "imagine", "notice", "remember", "see" "self-conscious", "smell", "suffer", "taste", "think", and hundreds more. But they allude to a large number of different kinds of states and processes, serving different biological (or cognitive) functions, and using different sorts of capabilities (or mechanisms).

"That fly detected my approaching hand and got away".
"My dog is aware that I am watching him".
"While sleep-walking Fred noticed that the door was open and shut it".
"He has been conscious for a few minutes, but is still a bit dopey"
"I've been aware for months that you don't like me".
"He felt very self-conscious coming into the room".
"In my dream I felt frightened" (Did I or did I not have consciousness then?)

It's just bad philosophy to assume that there's some common "thing" underlying all these states, that you've either got or haven't got (consciousness, intelligence, intentionality, etc.)

ESSENCES

{3} Consider chess: there are many possible games with rules that differ slightly from international chess rules. Are they still chess? Which ones have and which don't have the "essence" of chess?

If I play chess with a child, but start with only one rook in order to even up the game, is it still chess? Suppose there were no internationally agreed rules, and slightly different rules were used in different countries - e.g. one country builds a time-limit into each move, another limits only the average time, another rejects the rule that 50 moves without a capture entails a draw, etc. Would some countries then play REAL chess and others not?

Suppose further that there are many many variants in the games, and that at one extreme the game labelled "chess" is played without pawns, at another there's only a king and eight pawns per side, etc. Is there a line to be drawn between those versions that have the "essence" of chess and those that don't?

Is there a chess/not-chess dichotomy? "Essentialism" is the view that there must be.

{4} There are several false moves you can make in considering that sort of question.

E.g.:

(a) Essentialism: "There must be some games that have the real essence of chess, whereas others merely share certain superficial features with it".

NO: The only reason for believing that there is a unique essence is the existence of a word or phrase that we understand. But that does not imply that it has a well defined referent with well defined criteria for identity. All words, even technical terms of science and engineering, are partially semantically indeterminate in a variety of ways and that is one of the things that makes the evolution of language possible: in Physics the words "mass", "position", "time", etc. evolved from Newtonian to Einsteinian concepts without suddenly completely changing their meanings.

(b) Conventionalist Anti-Essentialism: "(i) There's a continuum of cases, and therefore (ii) where you draw a line is arbitrary, or simply a matter of convenience."

NO (i): It's not a continuum! There is a DISCRETE set of possible games, and it isn't true that between any two of them there is always some well-defined intermediate game, as would be required by a continuum.

NO(ii): There are MANY non-arbitrary lines to be drawn. By analysing the properties of the different games you can find interesting objective differences between them. E.g. some, but not all, have the property that the game tree cannot be searched exhaustively in the life of the universe. Some, but not all, have the property that there's a strategy for white that guarantees that you don't lose (like tic-tac-toe). Some have more subtle-properties like providing opportunities for deceptive tactics, that give the opponent a false idea about your sub-goals, whereas others may not be rich enough for that. Etc.

So the question where to draw the line between those cases that are "real" chess and those that are not, just isn't worth asking: "essentialism" is just a form of self delusion ("We know exactly what we mean"), whereas the conventionalist alternative misses a host of important distinctions, by seeing only a continuous blur.

Anti-Essentialist multiplicity: What IS worth while is trying to describe all the many interesting distinctions between different cases, and analysing the implications of the differences. This may require stretching old terminology and inventing new terminology to mark some of the previously unnoticed similarities and distinctions. The process is never ending.

{5} The principle of ANTI-ESSENTIALIST MULTIPLICITY: when studying a complex state, process, capability, etc. don't expect to find any essence defining a dichotomy between things with and without it, and don't expect an undifferentiated continuum of cases either.

Look not for one (or none), but for MANY, interesting and important distinctions.

{6} What goes for the set of chess-like board-games also goes for the SPACE OF POSSIBLE DESIGNS for biological or artificial mechanisms with interesting internal and external behavioural capabilities. Again there are false essentialist and anti-essentialist moves: (a) "There must be a dividing line between animals (or machines) with and those without consciousness (intentionality, mind, or whatever)." (b) "There's a continuum of cases and we can draw an arbitrary line wherever it suits us."

NO: there isn't any major dichotomy between systems with and systems without consciousness. There are many, many, many interesting distinctions, including some illustrated by differences between organisms with different designs. (Except that mostly we don't know what the differences are, because as yet biologists know little about how biological processes like perception, learning, planning, motor control, etc., actually work: how does a squirrel run along a branch without falling off? How does a spider build its web? How does a bird fly through the branches to its nest without crashing into anything on the way? How does a young chimp learn to imitate behaviour it observes in others?)

NO: the space is not a continuum: there are many kinds of important discontinuity. E.g. some designs include the use of controlling programs, and between two programs that differ (e.g. by the removal of a simple instruction) there's not necessarily some intermediate program, let alone a continuum of intermediate programs. Where a difference is only quantitative (e.g. speed) there may be a continuum of cases. Mostly there isn't.

NO: deciding how to divide up the space is NOT just a matter of arbitrary decision, or social convention, or deciding what is convenient for us (e.g. for purposes of explanation and prediction).

RATHER: the design differences produce objective properties that are significant in relation to such things as

- whether and how the organism or machine can cope in a particular sort of environment,

- whether it will or will not be able to compete successfully (always, or sometimes, or with a certain frequency) against systems with certain other designs,

- whether it will or will not be able to perform certain sorts of tasks,

- how fast it can perform tasks,

- how performance degrades as conditions become less favourable or the time available decreases,

- what it can learn (skills, facts, behavioural ideals, etc. etc.),

- whether and in what ways it can adapt to changing circumstances or new tasks (e.g. generated by old goals in new conditions),

- whether it can form plans in advance of executing complex actions,

- whether it can contemplate and evaluate alternative possible actions,

- what kinds of social relationships it is capable of entering into,

- and so on.

In accordance with the principle of anti-essentialist multiplicity we can summarise so far: although there are things that clearly do have minds and some that clearly don’t, expect neither a sharp dichotomy nor smooth variation between cases.

Instead look for many interesting discontinuities in the space of possible designs, and try to understand their implications.

VIRTUAL MACHINES AND STRUCTURES

{7} The concept of design here is not just a matter of physical architecture. We’ve learnt from computer systems engineering that a VIRTUAL machine is often the most important thing to think about when designing certain systems. "Virtual memory" systems that continually swap bits of programs between core memory and disk are a special case. Even without virtual memory a computer can implement different virtual machines for programs written in different languages (e.g. LISP and FORTRAN)

A program written in a high level programming language can manipulate lists, strings, numbers, arrays, trees, networks, etc. These are not physical entities: you will not find them if you open up your VAX (or whatever). They are virtual objects in virtual machines; and running programs are virtual machines that create, compare, store, modify, etc. these virtual objects. The same virtual machine can be *implemented* in very different ways in different physical machines.

{8} Even bits are not really physical things e.g. states of a set of binary switches. Rather they are virtual objects implemented in different ways in different systems. E.g. the bits in a bit-string may even have a richer structure than the set of physical switches (transistors) used for their implementation, when data-compression techniques are used for storage.

Similarly, sparse arrays can have components that don't have any independent physical realisation. A huge array most of whose cells have the same contents may be implemented using procedures that record only which cells don't have the "default" value, saving enormous amounts of space. To other programs this "sparse" array may look exactly like one that is encoded explicitly (on a slightly slower computer).

A logical data-base in a computer can contain "theorems" whose existence depends on a derivational capability. It may not even be a fixed capability: the resources available for the derivation can depend on what else is going on in the system at the time. So the data-base is a virtual structure with fuzzy edges.

{9} The idea of a "Physical symbol system" (originally promoted by Herbert Simon and Allen Newell) has been a great red herring in recent years. What we need to think about in understanding how we (and possibly other organisms and artificially intelligent systems) work is not physical symbols, but "virtual" structures, virtual mechanisms, virtual architectures.

Different virtual structures and mechanisms have different properties, fitting them for different roles in complex designs for behaving systems. (And they don't form a continuum!)

The properties of virtual structures are defined by the (virtual) operations available for manipulating them. Thus what makes a virtual structure in the machine a tree is a collection of procedures for accessing and updating nodes of the tree, for selecting branches, for comparing trees, for inserting a new tree as a sub-tree of an old one, etc. The structure thus defined will determine what uses are possible.

{10} Virtual structures treated as having logical (propositional) syntax and subject to logical transformations (like variable instantiation, or Modus Ponens) have different uses from structures that are treated more like maps, diagrams, etc. Which of these an organism a (virtual) machine is able to create and manipulate will affect its abilities to take in and use information about the environment, to make plans, to communicate, etc.

E.g. some kinds of structures have "generative power" lacked by others. If all available structures are composed of N symbols from a pre-determined finite alphabet of K characters then there will be exactly K^N structures all of the same form. If the vector contains N items selected from a continuum there will be infinitely many structures all of the same form. However, if new substructures can be hung off old ones (as in tree-manipulating virtual machines), then the number of possibilities is potentially infinite, and so is the variety of forms.

The kind of structural variability in a design is very important: organisms with the latter type of virtual machine may have a potential for coping with novelty that others lack.

{11} Many of the design differences concern trade-offs of various kinds - e.g. between speed and storage requirements, between flexibility and accuracy, between speed and reliability, between generality and power, between temporary storage for problem solving and long term storage for facts and the like.

{12} Among the important discontinuities in design space are those concerned with the types of structures and transformations of structures available to the system for embodying percepts, beliefs, desires, plans, skills, strategies, etc. This may be one of the important differences between a cricket and a chimpanzee, and perhaps even between a mouse and a chimpanzee.

{13} The virtual structures that a system is capable of creating and manipulating are not determined only by its enduring physical structure: two identical computers running different programs may implement different virtual machines, as can one computer at different times. A time-shared computer can interleave states in which different virtual machines are running.

Similarly one biological organism may develop new virtual structures and mechanisms in its life-time. Some of the differences between adults of two species may also be differences between infants and adults of one species. (I think Piaget was trying to understand this kind of development with totally inadequate conceptual apparatus.)

{14} Of course in our physical world, different ways of implementing a particular virtual machine will produce behavioural differences. E.g. certain physical designs may be incapable of performing certain kinds of tasks fast enough to solve real-world planning or recognition problems, or with sufficient long term reliability, or they may fail in particular kinds of stressful environments: so physical implementation details can be important as well as virtual machine design. But they are important in different ways.

NEURAL VERSUS SYMBOLIC

{15} The controversies between those who are for and those who are against connectionist(PDP) approaches to AI seem to me to be mostly misguided.

It is very likely that several different kinds of mechanisms are needed for different sub-tasks. Some people think that the use of one or other of these two in itself makes a big difference to whether truly intelligent, conscious, or whatever systems can be implemented (or explained). However, what is far more important is the global architecture of the system, i.e. what kinds of different sub-systems there are and how they interact with one another and with the environment: implementation details for the sub-systems matter less.

In addition there are muddles about distributed vs non-distributed representations, symbolic vs non-symbolic processing, etc. E.g. connection weights can implement virtual symbols representing probabilities, and distributed representations are commonplace in logic: where a host of theorems may be implicitly distributed across a set of axioms. Similarly AI chess-programs can use information about diagonals implicitly distributed over x and y co-ordinates.

{16} In a particular global architectural design, it can turn out that for many sub-systems either a PDP or more conventional approach can be adopted and this will make differences to such things as speed, flexibility, generality, etc. but not to the applicability of most of the global descriptions of mental states and processes (e.g. perceiving, thinking, feeling, learning, understanding, etc.) these depend on the global architecture, not sub-system implementation details.

{17} One way of thinking about the global design of a relatively high level virtual machine is to ask how many co-existing, independently variable, causally interacting, states it supports.

For instance a thermostat is designed to have two independently variable causally interacting states: one corresponding to a measure of the ambient temperature (often embodied in the degree of curvature of a bi-metallic strip), and one corresponding to the required temperature.

The former is "belief-like" in its function the latter "desire-like", but because it has such a simple set of possible states and possible interactions between states, the thermostat provides only very primitive instantiations of these concepts.

Many AI programs have a much richer functional differentiation between co-existing substates, and still don't come anywhere near the sophistication and complexity of the high level virtual machines running on humans or even chimpanzees. For instance, in many cases their "desire-like" states are all generated from a single top-level goal (solve that problem, win that game), whereas animals have multiple, independent, sources of motivation, that are often mutually incompatible.

Again, many AI systems are incapable of receiving information from the environment except when the execution of a plan includes a perceptual test, whereas humans and other animals have perceptual processes running in parallel and asynchronously with others, and are capable of generating a variety of types of new sub-processes, including interrupting other goal-driven processes. (This is one source of emotional states.) (Real-time control systems also have this feature.)

SELF-MONITORING

{18} One important design discontinuity concerns whether the system has an INTERNAL self-monitoring capability, for high level virtual machine states. I.e. can it generate and use "belief-like" states about its own states? (This may be an important aspect of what people find intriguing about consciousness -- insofar as it includes self-awareness.)

There are plenty of discontinuities between different types of self-monitoring capabilities. I suspect that one of the important differences between humans and other animals, and also between adult humans and young children concerns the availability of such internal self-monitoring.

There may also be cultural differences, insofar as, for example, the ability to perceive and classify one's own internal states will in part depend on one's conceptual capabilities, and these are partly culturally determined. Post-Freudian self-consciousness differs from pre-Freudian. Similarly a trained phonetician (phonologist) will hear things in his own speech that are inaudible to others.

{19} Different kinds of self-monitoring capabilities flow from different kinds of mechanisms and structures.

Many kinds of self-monitoring have been explored in software and hardware design, e.g. for implementing protection systems, for enabling operating systems to monitor and limit the powers of the user programs, for allowing systems to tune themselves by collecting information about their own behaviour and adjusting rules or parameters to improve performance according to some required criteria. Self-modifying neural nets introduce a new range of cases.

I suspect that at present we have only the foggiest notion of the kinds of self-monitoring capabilities that are possible with different architectural designs and different low-level implementations.

{20} Some kinds of self-monitoring are important for instantaneous control, e.g. in feedback systems, studied in control theory. Other kinds are important for long-term learning, and may require elaborate processes of credit/blame assignment. E.g. if a complex succession of actions enables you to do something better than or worse than you expected, how can you tell which parts of your plan, or which items of information used, are responsible for the discrepancy?

Sometimes there is no answer, whereas in other cases important kinds of learning can be based on finding the answer.

{21} It would be of some interest to see whether observable capabilities of different organisms, together with neurophysiological evidence, could be used as a basis for inferring the differences in global functional architecture. But I think this will be very difficult.

STRUCTURAL VARIABILITY

{22} The principle of REQUISITE STRUCTURAL VARIABILITY states that if an organism or machine needs to be able to distinguish and react differently to a range of circumstances, then it needs to be capable of being in different states whose variability matches the RELEVANT variability of the environment. If a system has only N possible internal states then it cannot distinguish N+1 different objects in the environment.

> **Note: Added 6 Nov 2017**
> The Principle of Requisite Variety is due to W. Ross Ashby, e,g see *An Introduction to Cybernetics,* (1956), Chapman & Hall
> http://pespmc1.vub.ac.be/ASHBBOOK.html
> See also this recorded presentation on the law by Stafford Beer (March 1990).
> https://www.youtube.com/watch?v=bDRudRhNgy4

This point needs some care: a marble's behaviour in rolling down different chutes differs according to the structure of the chute. The marble doesn't need to be able to have internal states that correspond to the differences in chute structure. This is because the behaviour is mostly under the control of the chute, not the marble and its state changes continuously under the influence of the changing physical forces exerted by the chute.

A child deciding which of two chutes, seen at a distance, will be more fun to slide down, needs coexisting states that discriminate the two structures -- but not necessarily in every physical detail. ("Cognitive friendliness" in the environment includes requiring relatively simple computations, and relatively simple storage requirements.)

An agent's need for significantly different internal states arises only when the environment leaves the agent a wide range of behavioural options and the agent itself has to make the selection. (Similarly when the behaviour is internal.)

Further, it is not just the NUMBER of different states that is important, but also the KIND OF VARIABILITY. Different kinds of variability provide different kinds of generalization capabilities, different kinds of learning, different kinds of control mechanisms, different kinds of inference mechanisms.

Some variability is one-dimensional (the temperature measurement in the thermostat), some multi-dimensional (e.g. a TV image digitizer with a large number of pixels measuring intensity or colour in different viewing directions). Some N-dimensional variability is continuous, some discrete.

Some variability is not N-dimensional, but STRUCTURAL, for instance in a parser that creates trees on reading in a sentence or a program, or an image interpreter that builds a structural description of the scene.

Structural variability implies changing possibilities for change! I.e. when the structure is more complex there are usually more ways of changing it (adding bits, removing bits, swapping bits, etc.), whereas at every point in an N-dimensional vector space the same possibilities for change exist: i.e. any subset of the N components can go up or down.

{23} Sometimes one kind of variability has to be mapped into another. A large space with little structure can be carved up so as to implement a smaller space with a richer structure. E.g. standard AI programs map a rich range of different kinds of variability into variability in a huge N dimensional space where each dimension admits only two values: e.g. different configurations of trees, networks etc. all get mapped into N-dimensional bit vectors, where N is the number of bits in the address space.

What higher level virtual structures correspond to substates in any lower level virtual machine, or even a physical substate, depends on how other parts of the system relate to that substate. Thus there is no direct correlation between physical portions of a computer and the virtual structures they implement. Similarly we should not expect direct and general correlations between neuronal substates and mental substates of human beings. It’s TOTAL machines that are mapped.

{24} One thing that computers and brains have in common that isn’t found in most previously studied mechanisms is the existence of very large numbers of independently variable units, whose combined state controls the fine structure of the system’s behaviour. Because 2 to the power N is an astronomically huge number for large N, even binary units can provide an unimaginably large number of possible states for the system, where there are lots of the units.

A major difference between brains and computers is that if there’s a single processor that changes the units, even if it does so to 16 or 32 of them at a time, then certain state changes cannot be done instantaneously, but have to go through intermediate states.

E.g. if going from state A to state B requires 1 million bits to be flipped, and only 32 can be flipped at a time then at least 31,248 intermediate states have to be traversed. This may or may not matter, depending on other features of the design. For instance there could be a problem if an interrupt occurred during the transition. By contrast, if different units can change their state simultaneously, e.g. on receiving broadcast messages, then it may be possible to have complex transitions with far fewer intermediate states.

On the other hand the mechanisms that co-ordinate such parallel state transitions may impose constraints on possible transitions that mean that a brain-like system has a more restricted set of possible state trajectories.

This could be an advantage if the restrictions correspond to the range of processes required for dealing with the range of situations the environment can provide. It could be a disadvantage if it limited the ability to learn to deal with significantly new types of situations.

{25} Similarly, a brain-like system may be more limited than a computer in the range of kinds of functional differentiation into co-existing interacting (virtual) sub-systems that it supports, but it may be able to support those sub-systems more efficiently and more robustly than a single very fast and very large computer used for the same purpose, which will be very vulnerable to errors or variability in the performance of the CPU as it scurries around changing large numbers of different things.

In particular, the need to go through large numbers of spurious intermediate states in order to go from state A to state B may make it possible for unexpected events to turn up in the interim (e.g. because new information has arrived through sensors) during states in which the system is incapable of dealing with them properly. The standard treatment of this problem is to have "critical regions" in programs during which interrupts etc are suppressed. But it may be impossible to anticipate and make provision for all such possible unwanted interactions.

{26} This discussion is pointing towards a conception of human minds as virtual machines implemented, possibly via several intermediate virtual machines, in a particular kind of physical machine: a brain.

However, what has been said above implies that there may be OTHER physical implementations of similar high level virtual machines (i.e. systems with similar kinds of decomposition into co-existing, independently variable, causally interacting substates).

Not all implementations will have EXACTLY the same properties, especially if computers are used instead of brains. Moreover, even among biological systems using brains we can expect to find a variety of different virtual machines, with different properties. In any case, no two people are EXACTLY the same.

So there’s a potentially huge space of possible designs (and implementations), and it provides us with a wonderful opportunity for collaborative research involving philosophers, psychologists, ethologists, social scientists, computer scientists, neurophysiologists, and engineers. Part of the task is to devise a good conceptual taxonomy for different kinds of designs by analysing some of the important discontinuities in the space and their implications, and seeing how they map on to both differences found among biological organisms and differences between biological and artificial agents waiting to be designed.

{27} I’ve put a lot of stress on mental states, processes and capabilities being essentially the states, processes and capabilities of a high level virtual machine in a complex system with sufficient internal functional differentiation and variability. However, what we normally describe as mental states etc. (e.g. "He saw the lion coming", "He longed for his mother") are rarely descriptions of purely internal states of one individual. Rather there are inextricable references to other bits of the world (the lion, his mother, etc.), and it is harder than you might think to avoid this. "There’s a red splodge in his visual field" implicitly relates his state to red things in the world.

Thus the mental states as we describe them in ordinary language are not just mental states, but also relations with the environment. This means that there will never be a simple correspondence between such ordinary descriptions and the kinds of (independently variable) internal states

discussed above.

COMPUTATION

{28} AI used to be thought of as committed to the idea that mental processes are essentially computations. I now think that the notion of computation is not able to carry the explanatory weight we expected of it. Instead we need to survey many different kinds of mechanisms to see what kinds of roles they can play within the sorts of designs referred to above. Computational mechanisms have played an important catalytic role in helping us seem beyond previous conceptions of mechanisms ("mere" machines).

Whether the mechanisms used in creating or explaining intelligent systems are or are not computational will be of less importance than their other features.

{29} A survey of discontinuities in the space of possible designs would generate questions for distinguishing types of mechanisms, including the following:

1. How many independently variable (physical or virtual) substates can the machine support?

2. What forms of variation do the substates admit? (Continuous or discrete? N-dimensional vector space or structural variability? Does it allow logical structures with substitution of variables? Can one substate be stored in another, then later retrieved? etc.)

3. What kinds of causal interactions does it support?
   (Can internal variation of one substate be finely controlled "online" by another substate? Can the addition of a new substate to trigger new processes? What kinds of internal feedback loops does it support? Are most of the changes controlled by the environment or is the majority of processing internally generated? Are all changes synchronised (e.g. by a clock) or can substates change at different speeds? Can some states be "belief-like", i.e. largely under the control of causal input from the environment, and others "desire-like", i.e. able to generate external actions when the environment does not "fit" them? Does it allow records of events to be added to a long term store for future use?)

4. How fast can states change and causes propagate, relative to the kinds of changes that can occur in the environment?

5. What is the global architecture of the system: what kinds of separable, but interacting, sub-systems can be distinguished, performing different functions within the whole?

6. Are the different substates spatially separable, i.e. embedded in different sub-structures (like computer datastructures) or superimposed in a distributed form (like superimposed wave forms or neural states)?

7. Is there always cross-talk between the different substates, so that changing one can alter another or can they be causally isolated?

And so on.....

SUBJECTIVITY, FREEDOM

{30} "Subjective experience" - mostly this is just a muddle (as if there could be an "objective" kind of experience). Anything that has sensors has subjectivity of a kind (what it is capable of sensing will be determined in part by the nature of the sensors and the amount and kind of internal variability, and in part by its "viewpoint").

Some of the talk about subjective experience is based on a dim recognition of the role of internal monitoring, discussed above.

{31} What philosophers call "raw feels" - the directly experienced contents of such states as having a pain, feeling thirsty, seeing a colour, enjoying an icecream, etc. can all be accounted for in the above framework provided that we accept that there is no such thing as an intrinsically identifiable mental state or process or content, only states (processes, etc) with networks of relationships to other states, processes etc.

Working out the precise networks of relationships involved in particular mental states (feeling a toothache) is a non-trivial task.

{32} Questions about whether computers can have "Real" pains, or understanding, (e.g. Searle's questions) are generally based on the Essentialist fallacy that there is some well-defined dichotomy between things which do and things which don't have pains, or understanding. The question evaporates when we do a detailed exploration of the full range of design discontinuities: because that shows up the limitations of ordinary concepts.

{33} Freedom: I've never been able to understand why so many philosophers are so concerned about this. As long as my decisions and actions flow from my beliefs, preferences, desires, ideals, hopes, fears, etc. (with the appropriate causal connections) I can't see what other kind of freedom is required, or possible. A machine could also have that kind of freedom.

Reflexes, instinctual reactions, external threats, physical force are all capable of limiting it. But simply being physically determined, or even programmed, if programs underly all the relevant kinds of mental states, does not limit freedom.

Of course, I have problems when my beliefs, preferences, desires, ideals, etc. are not all mutually consistent.

Only a detailed mechanist analysis can show why we are not "mere machines". ("mere" = fitting the late 19th Century concept.)

{34} Implications of autonomy. There's no space for a justification here, but I claim that a study of design issues for an intelligent system with multiple independent sources of motivation and limited resources in a complex, changing, and only partly knowable world shows that mechanisms (for dealing with, and in some cases filtering, interrupts) are required that are capable, as a side effect, of producing emotional states, just as the design of certain kinds of operating systems can produce "thrashing". So, no special emotional subsystem is necessary. (Of course, some animals may have one - but it's not necessary for having emotions.)

CONCLUDING REMARKS

{35} Magic seems to be the only alternative to the position sketched here. If human thought, feelings, sensations, decisions, emotions, moods etc. are amenable to any kind of scientific analysis and explanation then it must be because we are made of appropriate kinds of (virtual) machines: Every intelligent ghost must contain a machine.

Alternatively, there may be inexplicable, magical, phenomena in the universe. So far I have seen no evidence for this, only evidence for our currently limited ability to understand and explain, and evidence that the frontiers are constantly being pushed forward.

Hubert Dreyfus once compared this with the claim "We're getting nearer to the moon" made by people climbing trees. Suppose there were two tribes, one that kept trying to climb taller and taller trees, and one that merely scrabbled around on the ground. Which would you bet on as more likely to get to the moon one day?

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham
Converted from 'groff' source on: 27 Dec 2007