**To appear in Proceedings of the 1992
Royal Institute of Philosophy Conference
'Philosophy and the Cognitive Sciences',
(eds) C. Hookway and D.Peterson,
Cambridge University Press**

THE MIND AS A CONTROL SYSTEM

Aaron Sloman

School of Computer Science

The University of Birmingham

Abstract

Many people who favour the design-based approach to the study of mind, including the author previously, have thought of the mind as a computational system, though they don't all agree regarding the forms of computation required for mentality. Because of ambiguities in the notion of 'computation' and also because it tends to be too closely linked to the concept of an algorithm, it is suggested in this paper that we should rather construe the mind (or an agent with a mind) as a control system involving many interacting control loops of various kinds, most of them implemented in high level virtual machines, and many of them hierarchically organised. (Some of the sub-processes are clearly computational in character, though not necessarily all.) A number of implications are drawn out, including the implication that there are many informational substates, some incorporating factual information, some control information, using diverse forms of representation. The notion of *architecture,* i.e. functional differentiation into interacting components, is explained, and the conjecture put forward that in order to account for the main characteristics of the human mind it is more important to get the architecture right than to get the mechanisms right (e.g. symbolic *vs* neural mechanisms). Architecture dominates mechanism

1 Introduction

This is not a scholarly research paper, but a 'position paper' outlining an approach to the study of mind which has been gradually evolving (at least in my mind) since about 1969 when I first become acquainted with work in Artificial Intelligence through Max Clowes. I shall try to show why it is more fruitful to construe the mind as a control system than as a computational system (although computation can play a role in control mechanisms).

During the 1970s and most of the 1980s I was convinced that the best way to think of the human mind was as a computational system, a view that I elaborated in my book *The Computer Revolution in Philosophy* published in 1978. (Though I did point out that there were many aspects of human intelligence whose explanation and simulation were still a very long way off.)

At that time I thought I knew exactly what I meant by 'computational' but during the late 1980s, while trying to write a second book (still unfinished), I gradually became aware that I was confused between two concepts. On the one hand there is a very precisely definable technical concept of computation, such as is studied in mathematical computer science (which is essentially concerned with *syntactic* relations between sequences of structures,

e.g. formally definable states of a machine or sets of symbols), and on the other hand there is a more intuitive, less well-defined concept such as people use when they ask what computation a part of the brain performs, or when they think of a computer as essentially a machine that *does* things under the *control* of one or more programs. The second concept is used when we talk about analog computers, for these involve continuous variation of voltages, currents, and the like, and so there are no sequences of states.

Attempting to resolve the confusion revealed that there were not merely two but several different notions of computation that might be referred to in claiming that the mind is a computational system. Many of the arguments for and against the so-called 'Strong AI Thesis' muddle up these different concepts and are therefore at cross purposes, arguing for not inconsistent positions, despite the passion in the conflicts, as I've tried to show in (Sloman 1992), which demonstrates that there are at least eight different interpretations of the thesis, some obviously true, some obviously false, and some still open to investigation.

Eventually I realised that the non-technical concept of computation was too general, too ill-defined, and too unconstrained to have explanatory power: whereas the essentially syntactic technical concept was too narrow: there was no convincing reason to believe that being a certain sort of computation in that sense was either necessary or sufficient for the replication of human-like mentality, no matter which computation it was.

Being entirely computational in the technical sense could not be necessary for mentality because the technical notion requires all processes to be discrete whereas there is no good reason why continuous mechanisms and processes should not play a significant part in the way a mind works, along with discrete processes.

Being a computation in the technical sense could not be sufficient for production of mental states either. On the contrary, a static sequence of formulae written on sheets of paper could satisfy the narrow *technical* definition of 'computation' whereas a mind is essentially something that involves processes that interact causally with one another.

To see that causation is not part of the technical concept of computation, consider that the limit theorems showing that certain sorts of computations cannot exist merely show that certain sequences of formula, or sequences of ordered structures (machine states) cannot exist, e.g. sequences of Turing machine states that generate non-computable decimal numbers. The famous proofs produced by Gödel, Turing, Tarski and others do not need to make assumptions about causal powers of machines in order to derive non-computability results. Similarly complexity results concerning the number of steps required for certain computations, or the number of co-existing memory locations do not need to make any assumptions about causation. Neither would adding any assumptions about computation as involving causation make any difference to those results. Even the definition of a Turing machine requires only that it has a sequence of states that conform to the machine's transition table: there is no requirement that this conformity be *caused* or *controlled* by anything, not even any mechanism implementing the transition table. All the mathematical proofs about properties and limitations of Turing machines and other computers depend only on the formal or syntactic relations between sequences of states. There is not even a requirement that the states occur in a temporal sequence. The proofs would apply equally to static, coexisting, sequences of marks on paper that were isomorphic to the succession of states in time. The proofs can even apply to sequences of states encoded as Gödel numbers that exist neither in space nor in time, but are purely abstract. This argument is elaborated in Sloman (1992), as part of a demonstration that there is an interpretation of the Strong AI thesis in which it is trivially false and not worth arguing about. This version of the thesis, I

suspect, is the one that Searle thinks he has refuted (Searle 1980), though I don't think any researchers in AI actually believe it. There are other, more interesting versions that are left untouched by the 'Chinese Room' argument.

Unfortunately, the broader, more intuitive concept of computation seems to be incapable of being defined with sufficient precision to form the basis for an interesting, non-circular, conjecture about the nature of mind. For example, if it turns out that in this intuitive sense *everything* is a computer (as I conjectured, perhaps foolishly, in (Sloman 1978)), then saying that a mind is a computer says nothing about what distinguishes minds (or the brains that implement them) from other behaving systems, such as clouds or falling rocks.

I conclude that, although concepts and techniques from computer science have played a powerful catalytic role in expanding our ideas about mental mechanisms, it is a mistake to try to link the notion of mentality too closely to the notion of computation. In fact, doing so generates apparently endless and largely fruitless debates between people talking at cross purposes without realising it.

Instead, all that is needed for a scientific study of the mind is the assumption that there is a class of *mechanisms* that can be shown to be capable of producing all the known phenomena. There is no need for researchers in AI, cognitive science or philosophy to make restrictive assumptions about such mechanisms, such as that they must be purely computational, especially when that claim is highly ambiguous. Rather we should try to characterise suitable classes of mechanisms at the highest level of generality and then expand with as much detail as is needed for our purposes, making no prior commitments that are not entailed by the requirements for the particular mechanisms proposed. We may then discover that different sorts of mechanisms are capable of producing different sorts of minds, and that could be a significant contribution to an area of biology that until now appears not to have produced any hard theories: the evolution of mind and behaviour.

2 How can we make progress?

When trying to find a general starting point for a theory about the nature of minds there are many options. Some philosophers start from the notion of 'rationality', or from a small number of familiar aspects of human mentality, such as beliefs and desires, or something common to several of them, often referred to as 'intentionality.' I suggest that it would be more fruitful to step back to the very general notion of a mechanism that interacts with a changing environment, including parts of itself, in a way that is determined by (a) the changeable internal state of the mechanism, (b) the state of the environment and (c) the history of previous interactions (through which the internal state gets changed). This is a deeply *causal* concept, the concept of a *control system*. So I am proposing that we revive some old ideas and elaborate on the not particularly novel thesis that the mind is essentially a control system. But this is still too general, for the notion of such a control system covers many physical objects (both naturally occurring or manufactured) that clearly lack minds. By adding extra constraints to this general concept we may be able to home in on a set of interesting special cases, more or less like human beings or other animals.

The purposes for which mental phenomena are studied and explained will vary from one discipline to another. In the case of AI, the ultimate requirement is to produce working models with human-like mental properties, whether in order to provide detailed scientific explanations or in order to solve practical problems. For psychologists the goal may be to model very specific details of human performance, including details that differ from one individual to another, or from one experimental situation to another. For engineering

applications of AI, the goal will be to produce working systems that perform very specific classes of tasks in well-specified environments. In the case of philosophy it will normally suffice to explore the *general* nature of the mechanisms underlying mental phenomena down to a level that makes clear how those mechanisms are capable of accounting for the peculiar features of machines that can think, feel, take decisions, and so on.

That is the goal of this paper, though in other contexts it would be preferable to expand to a lower level of detail and even show how to produce a working system, in a manner that would satisfy the needs of both applied AI and detailed psychological modelling.

Since there are many kinds of control systems, I shall have to say what's *special* about a mind. I shall also try to indicate where computation fits into this framework. I'll start by summarising some alternative approaches with which this approach can be contrasted.

3 Philosophical approaches to mind

Philosophers generally try to study the mind by using conceptual and logical approaches, with subtasks such as the following:

- Analyse ordinary concepts to define notions like 'mind', 'consciousness', 'pleasure', 'pain', etc.
- Attempt to produce arguments ('transcendental deductions' Kant called them) showing that certain things are absolutely necessary for some aspect of mind or other.
- Produce metaphysical theories about what kinds of things need to exist in order to make minds possible (e.g. different kinds of stuff, special kinds of causal relationships, etc.)

Further common philosophical questions include whether all mental phenomena can be reduced to some subset (e.g. whether all mental states can be defined in terms of collections of beliefs and desires), whether certain descriptions of mental phenomena are names of 'natural kinds', which phenomena can be assessed as rational or irrational, and whether it is possible to know the contents of another person's mind.

It is very hard to discuss or evaluate such analyses and theories, e.g. because

- Ordinary concepts are full of imprecision and indeterminacy limiting their technical usefulness. So questions posed in terms of them may lack determinate answers.
- The theories usually have a level of generality and imprecision that makes it very hard to assess their implications or evaluate them. Acceptance or rejection appears often to be a matter of personal taste or prejudice, or philosophical fashion.
- It is hard to distinguish substantive questions with true or false answers from questions that are to be answered by taking more or less arbitrary terminological decisions (e.g. where are the boundaries between emotions, moods, attitudes, or between animals that are and animals that are not conscious?)
- Very often the philosophical issues are posed in terms of a small subset of the known phenomena of mind (e.g. conscious thought processes expressible in words) whereas any theory of what minds are and how they work should encompass far more richness, including indescribably rich experiences (like watching a waterfall), and phenomena exhibited only in young children, people with brain damage, and in some cases other animals.
- The variety found in animals of various sorts, human infants, brain damaged people, etc. suggests that there are few or no absolutely necessary conditions for the existence of mental capabilities, only a collection of different designs with different properties.

- Philosophers often make false assumptions about what sorts of mechanisms can or cannot exist because they have not been trained as software engineers and therefore know only about limited classes of mechanisms and have only very crude conceptions of possible computational mechanisms. In particular, they tend to be ignorant of the way in which the concept of a 'virtual machine' has extended our ideas. (A virtual machine is created in a physical machine by programs such as 'interpreters' that make it possible to specify higher level machines that have totally different properties from the physical machine. In a machine running text-processing software such as I am now using, there are letters, numerals, words, sentences, paragraphs, diagrams, chapters, etc., and there are mechanisms for operating on these things, e.g. by inserting, deleting, or re-ordering these objects. However these textual entities are not physical entities and do not exist in the physical computer, which remains the same machine when the word-processing software is replaced by some other software, e.g. a circuit design package.)

These are among the features of philosophical discussion that often provoke exasperated impatience among non-philosophers, e.g. scientists interested in the study of mind who encounter phenomena that are ignored by philosophers, including other animals, people with brain damage and sophisticated machines.

The real determinants of the mind are not conceptual requirements such as rationality, but biological and engineering design requirements, concerned with issues like speed, flexibility, appropriateness to the environment, coping with limited resources, information retention capabilities, etc. We'll get further if we concentrate more on how it is possible for a machine to match its internal and external processes to the fine structure of a fast-moving environment, and less on what it is to be rational or conscious. Properties such as rationality and intentionality will then emerge if we get our designs right. 'Consciousness' will probably turn out to be a concept that's too ill-defined to be of any use: it will instead be replaced by a collection of systematically generated concepts derived from theoretical analysis of what different control systems can do.

4 Philosophers as designers

For the reasons given above, my preferred approach to many philosophical questions is to treat them from the standpoint of an engineer trying to design something more or less like a human being, but without assuming that there's going to be only *one* possible design, or that there are any absolutely necessary conditions to be satisfied, or even that the notion of what is to be designed is precisely specified in advance. I call this the 'design-based' approach (defined more fully in (Sloman 1993)).

This is closely related to what Dennett described as the 'design stance' (Dennett 1978). It requires us to specify our theories from the standpoint of how things work: how perception works, how motives are generated, how decisions are taken, how learning occurs, and so on. Moreover, it requires us to specify these designs with sufficient clarity and precision that a future engineer might be able to expand them into a working instantiation. Since this is very difficult to do, we may, for a while, only be able to *approximate* the task, or achieve it only for *fragments* of mental processes, which is what has happened in AI so far.

But the design stance does not require unique solutions to design problems. We must keep an open mind as to whether there are alternative designs with interestingly varied properties: abandoning Kant's idea of a 'transcendental deduction' proving that certain features are necessary. Instead we can explore the structure of 'design space' to find out what sorts of behaving systems are possible, and how they differ.

Adopting this stance teaches us that our ordinary concepts are inadequate to cope with the full variety of kinds of systems and kinds of capabilities, states, or behaviour that can emerge from exploratory studies of alternative designs in various kinds of environments, just as they are inadequate for categorising the full variety of forms of mind found in biological organisms, including microbes, insects, rodents, chimps and human beings. If we don't yet know what mechanisms there may be, nor what processes they can produce, we can't expect our language to be able to describe and accurately distinguish all the interestingly different cases that can occur, any more than ordinary concepts can provide a basis for saying when a foetus becomes a human being or when someone with severe brain damage is no longer a human being. Our concepts did not evolve to be capable of dealing with such cases.

We should assess theories in terms of their ability to support designs that actually work, as opposed to merely satisfying rationality requirements, fitting introspection, or 'sounding convincing' to willing believers.

In true philosophical spirit we can let our designs, and our theorising, range over the full space of possibilities instead of being constrained to consider only designs for systems that already exist: this exploration of possible alternatives is essential for clarifying our concepts and deepening our understanding of existing systems.

This is very close to the approach of AI, especially broad-minded versions of AI that make no assumptions regarding mechanisms to be used. Both computational and non-computational mechanisms may be relevant, though it's not obvious that there's a sharp distinction.

5 Key ideas, and some implications

I'll now try to list some of the key ideas driving the design-based study of mind.

- A mind is a well-designed, sophisticated, self-modifying control system, with functional requirements such as speed, flexibility, adaptability, generality, precision and autonomous generation of goals. It is able to operate in a richly structured, only partly accessible, fast-changing environment in which some active entities are also minds.
- This idea, that a mind is a control system meeting complex, detailed and stringent engineering requirements, when developed in full detail, has profound implications for several theoretical and scientific disciplines concerned with the study of aspects of the human mind, such as philosophy, psychology and linguistics: it suggests the form that explanatory theories have to take, and it has implications regarding evaluation of theories. For instance, it is not enough for a theory to be consistent with observed behaviour: Additional possible criteria can be explored such as (a) that the design should use 'low level' mechanisms like those found in brains, or (b) that the design should be capable of having been produced by an evolutionary process, or (c) that it must be a *good* design.
- By exploring different criteria of goodness for designs we can replace the impoverished philosophical criteria for agency, such as rationality or consciousness with a host of different sorts of requirements, including speed and flexibility, and explore their consequences.
- All this has practical implications, for education, counselling, and the design of usable interactive systems: for it is only when you understand how something *works* that you can understand ways in which it can go wrong, or design good strategies for dealing with it.

- The view that a mind is a control system is not provable or refutable: it defines an approach to the study of mind. In particular, it is not possible to argue against those who believe minds include a 'magical' element inaccessible except through introspection and inexplicable by scientific (mechanistic) theories of mind: that sort of belief is not rationally discussable. I shall simply ignore it here, though I think it can sometimes be overcome by a long sequence of personal philosophical 'tutorials', partly analogous to therapy.
- An *intelligent* control system will differ in important ways from the kinds of control systems hitherto studied by mathematicians and engineers. For instance, much of the control is concerned with how information is processed, rather than with how physical factors, such as force or speed, are varied. This point is developed below.
- By surveying types of control systems, their properties, the kinds of states they can have, we may expect to generate a 'rational reconstruction' of concepts currently used for describing mental states and processes, analogous to the way the periodic table of chemical elements led to a rational reconstruction of pre-scientific concepts of kinds of stuff. In both cases, primitive but usable collections of pre-theoretic concepts evolve gradually into more systematic families of concepts corresponding to configurations of states and processes compatible with a deep theory.
- In particular, the idea of a mind as control system leads to a new analysis of the concept of 'representation': a representation is part of a control state: and different kinds of representations play different roles in control mechanisms. (There are many different kinds of representations, useful for different purposes, each with its own syntax, semantics, and manipulation mechanisms.)
- Some AI work has concentrated excessively on representations (formalisms) and algorithms required for particular tasks, such as planning, reasoning, visual perception or language understanding. We also need to consider global *architectures* combining several different functions and we need to explore varieties of mechanisms within which such architectures can be implemented. This means considering not only what information is used, how it is represented and how it is transformed, but also what the important functional components of the system are, and what their causal powers and functional roles are in the system. Much of the functionality can be circular: the function of A is partly to modify the behaviour of B, and the function of B is partly to modify the behaviour of A. (Beliefs and desires are related in this circular fashion, which is why purely behavioural analyses of mental states fail.)
- From the standpoint outlined here, some debates about the relative merits of connectionist mechanisms and symbol-processing mechanisms appear trivial, for they are concerned with 'low level' details, whereas it is more important to understand the global architectures capable of supporting mind-like properties. I suspect that we shall find that, as in many control systems, architecture dominates mechanism: that is changing the low level implementation details will make only a marginal difference to the capabilities of the system at least in normal circumstances.
- In an intelligent control system most of the important processes are likely to be found in abstract or 'virtual' machines, whose main features are not physical properties and physical behaviour, though they are *implemented* in terms of lower level physical machines. The virtual machines manipulate complex information structures (such as networks of symbols) rather than physical objects and their physical properties. E.g. a word-processor manipulates words, paragraphs, etc., though these cannot be found in the underlying physical machine. Similarly, although they interact causally, the components of

a virtual machine do not interact via physical causes, such as forces, voltages, pressures, magnetic fields, even though they are implemented in terms of machines that do.

This last point, I believe, is the most important contribution of computer science to the philosophical study of mind, rather than the concept of a program or algorithm that generates behaviour, though much discussion of the relevance of computation has focused on the latter.

6 What distinguishes mind-like control systems?

Suppose we think of a mind as: an incredibly complex, self-monitoring, self-modifying control system, implemented at least in part as a collection of interacting virtual machines. This raises the following question, already hinted at: How is it like and how is it unlike other control systems? For instance, there is a large body of mathematics concerning control systems, usually represented as a set of measurable quantities and a set of differential equations stating how those quantities change over time. Does that help us understand how minds work? I believe the answer turns out to be: 'not much'! This is for the following reasons:

- The most important changes and processes in a mind don't map onto numeric variation: many of the architectural changes and the processes that occur within components of the architecture are structural, not quantitative. For example, they may involve the creation and modification of structures like trees and networks, for instance parse-trees representing perceptual structures. By contrast the typical components of an unintelligent control mechanism will be concerned with varying some measurable quantity, and even if there are many quantities sensed or modified, and many links between them, the variety of causal roles is limited to what can be expressed by sets of partial differential equations linking changing numerical measures. This does not allow for processes like creation of a parse-tree when a sentence is analysed or creation of a structural description when a retinal image is interpreted. (This point is rather subtle: I am not denying that mechanisms of the required type can be virtual machines that are *implemented* in mechanisms of the wrong type: a pattern that pervades computer science and software engineering.)
- The architecture of a human mind is so rich: there's enormous functional differentiation within each individual. Not only are there many different components to a working human-like mind, they have very different functional roles, including analysing and interpreting sensory input, generating new motives, creating and executing plans, creating representations of possible futures, storing information for future use, forming new generalisations, and many more. These differences of function are not easily captured by standard mathematical formalisms for representing changing systems.
- The architecture of an intelligent, human-like system is not static, it develops over time. A child gradually develops new combinations of capabilities concerned with cognitive skills, and also motivational and emotional control. A fixed set of differential equations can't model a changing architecture. Even if the architecture at a particular time could be expressed by such a set of equations, something more would be needed to represent the change from one set of equations to another, and that's not something differential equations can do (though it can be done by symbol manipulating programs). This point is not unique to mind-like control systems: it is commonplace in biological systems, where seeds change into trees, caterpillars change into moths and every embryo has a rapidly changing structure. The architectural changes in a human mind are more subtle and far harder to detect than structural changes in organisms, especially if they are changes in virtual machine structures without any simple physical correlates.

All of this implies that:

- Causal influences are not all expressible as transmission of measurable quantities like force, current, etc. Some involve transmission of structured ‘messages’ and instructions between sub-components. Some processes build new structures. Some of the causal interactions occur in the virtual machines that are supervenient on physical machines. (I am aware that some philosophers believe that supervenient processes cannot interact causally. It would take too long to refute that here: what happens in software systems is a concrete refutation.)
- New kinds of mathematics are needed to cope with this, although there has been some progress already, for example in the mathematical study of formal languages, proof systems, parsing mechanisms, and transformations of datastructures.
- Most of the control systems previously studied by mathematicians and engineers do not involve operations in virtual machines: they are physical machines with physical processes that are controlled by other physical processes. Thus the basic processes are expressible as physical laws relating physical quantities. In a mind the processes occur in a variety of virtual machines whose laws differ greatly. The processes involved in creating a 3-D interpretation from a 2-D image, and the processes involved in deriving a new plan of action from a set of beliefs and goals are very different from each other and from the way changes in temperature can produce changes in pressure or volume.

The currently fashionable ideas from dynamical systems theory are unlikely to prove rich enough to fill the need. I shall try to explain why in the next section.

7 The need for new concepts

We need new thinking tools to help us grasp all this complexity. We lack good ‘global ideas’ to help us think about the architecture of the whole mind: how the bits studied by AI fit together. A key idea is that a control system has independently variable causally interacting sub-states. By looking at ways in which complex systems can be analysed into components with their own changing states with different characteristics, we can begin to describe global architectures. For this purpose we should not think of a behaving system as having single ‘atomic’ total state that changes over time, as is common in physics and engineering. Rather we need the notion of a ‘molecular’ state, which is made of several different states that can change separately.

- atomic state: The whole system state is thought of as indivisible, and the system moves from one state to another through a ‘state space’ sometimes referred to as a ‘phase space’. The total system has a single ‘trajectory’ in state space. There may be ‘attractors’, that is regions of state space which, once reached, cannot be left.

The idea of a complete system as having an atomic state with a ‘trajectory’ in phase space is an old idea in physics, but it may not be the most useful way to think about a system that is made of many interacting subsystems. For example a typical modern computer can be thought of as having a state represented by a vector giving the bit-values of all the locations in its memory and in its registers, and all processes in the computer can be thought of in terms of the trajectory of that state-vector in the machine’s state space. However, in practice this has not proved a useful way for software engineers to think about the behaviour of the computer. Rather it is generally more useful to think of various persisting sub-components (strings, arrays, trees, networks, databases, stored programs) as having their own changing states which interact with one another.

So it is often more useful to consider separate subsystems as having their own states, especially when the architecture changes, so that the set of subsystems, and substates, is not static but new ones can be created and old ones removed. This leads to the following notion:

- molecular state with sub-states: The instantaneous state of a complete system sometimes includes many coexisting, independently variable, interacting, states of different kinds, which change in different ways, under external or internal influences. These may be states of subsystems with very different functional roles. The number of relevant interacting substates may change over time, as a result of their interactions.

A physicist or engineer who represents a complex system by a vector of measurements representing a point in a high dimensional 'phase space' and treats all change as motion of the point is using the *atomic* notion of a state. By contrast, if the system is thought of as having many different components and the processes of change in those components are studied separately, the concept of state is then *molecular*. Of course, if the atomic state is represented by a vector there are independently variable components: the components of the vector. But in a molecular state the number of components and their connections can vary over time, and some of the components will themselves have complex molecular states; whereas for atomic states the number of dimensions of a phase space is fixed, and the components of the vectors are numerical values rather than complex structures. Thus the molecular conception of state allows the state of a system to be hierarchically structured, with changing structures at several levels in the hierarchy.

Within the molecular approach we can identify a variety of functional sub-divisions between sub-states and sub-mechanisms, and investigate different kinds of functional and causal interactions. For example, we can describe part of the system as a long term information store, another part as a short-term buffer for incoming information, another as concerned with interpreting sensory input, another as drawing implications from previously acquired information, another as storing goals waiting to be processed, and so on. The notion of a global atomic state with a single trajectory is particularly unhelpful where the various components of the system function asynchronously and change their states at different rates, speeding up and slowing down independently of other subsystems.

Thus if dynamical systems theory is to be useful it will be at best a characterisation of relatively low level implementation details of some of the subsystems. It does not provide a useful framework for specifying how intelligent mind-like control systems differ from such things as weather systems.

8 Towards a taxonomy of interacting substates and causal links

In order to make progress with this approach to the study of mind we need to develop a collection of concepts for describing different substates of an intelligent control system. This will be an iterative process, starting with some initial concepts and then refining, discarding or extending them in the light of experience of attempts to survey an increasing variety of designs in increasing depth. In order to start the process, I'll use terms from ordinary language to bootstrap a new conceptual framework. In particular, the following types of substates seem to be important for mind-like control systems:

- Desire-like control states: these can be thought of as initiating processes, maintaining or modifying processes, and terminating processes, of various kinds. Some desire-like states create or modify other desire-like substates rather than directly generating or modifying behaviour. Speaking loosely we can say that causation 'flows away' from desire-like states

to produce changes elsewhere. George Kiss at the Open University pointed out to me that the concept of 'attractor' in dynamical systems theory, namely a region in phase space towards which a system tends and from which it does not emerge once having entered, is partially like a desire-like state. It seems unlikely to me that this notion of an attractor is general enough to play the role of desire-like control states in intelligent systems. There are several reasons for this, including the fact that some desire-like states appear to have a complex internal structure (e.g. wanting to find or build a house with a certain layout) that does not seem to be capable of being well represented by a region in phase space. Moreover desire-like states can themselves be the objects of internal manipulation, for instance when an agent suppresses a desire, or reasons about conflicting desires in deciding what to do. Of course, in principle a defender of the dynamical systems analysis could try to construe this as a higher level dynamical system with its own attractors operating on a lower level one. Whether this way of looking at things adds anything useful remains to be seen.

- Belief-like control states: these are more passive states produced and changed by causes 'flowing into' them. Only in combination with desire-like states will they tend to produce major new processes. (I do not know whether dynamical systems approach can give a convincing account of how belief-like and desire-like states interact.)
- Imagination-like control states: these are states that may be very similar in structure to belief-like states, but have a different causal basis and different causal effects. They may be constructed during processes of deciding what to do by exploring possible consequences of different actions. It is not clear how many animals can do this!
- Plan-like control states: these are states which have pre-determined sequences of events encoded in a form that is able, via an 'interpreter' mechanism to generate processes controlled by the plan. A stored computer program is a special case of this. Research on planning and acting systems in AI has unearthed a wide variety of forms and functions for such states. A more comprehensive, though still inadequate, list of control states apparently required for intelligent agents can be found in Beaudoin and Sloman (1993).

The concepts introduced here have been 'boostrapped' on our ordinary understanding of words like 'desire' and 'belief' in combination with hints at their significance from the design standpoint. This is an unsatisfactory intermediate state in our understanding, to be remedied later when we have a clearer specification of the functional differences between the different sorts of control states. (The definitions will necessarily be mutually recursive in a systems with many feedback loops, a point implicitly acknowledged by Gilbert Ryle in *The concept of mind* insofar as he rejected the kind of behaviourism that defined mental states purely in terms of external stimuli and behaviour.)

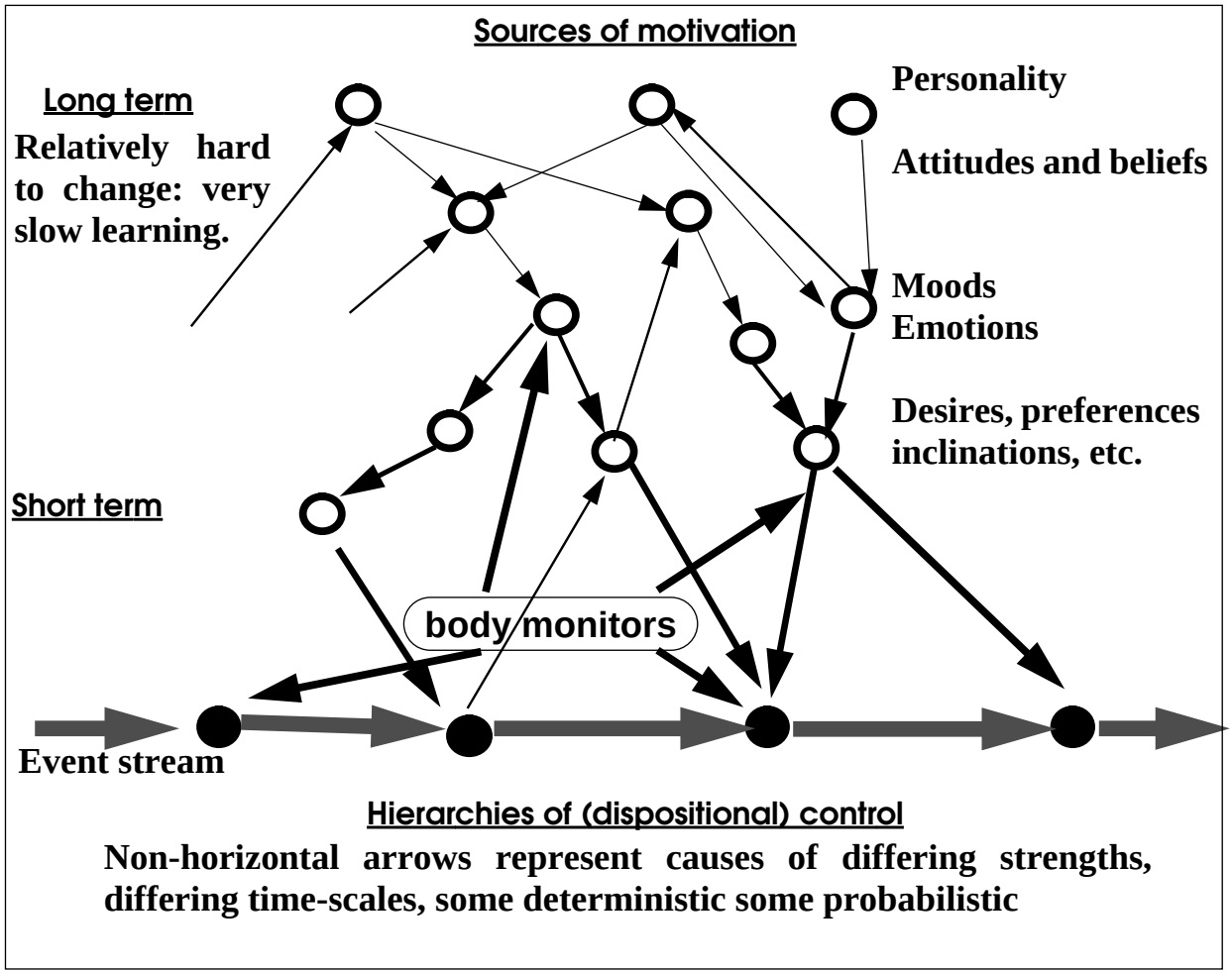
The control states listed above are not the only types of states to be found in intelligent agents: they merely indicate the *sorts* of things that might be found in a taxonomy of substates of an intelligent system. For complete specifications of control systems we would need more than a classification of states. We would also need to specify the relationships between states, such as:

- Kinds of variability: how subsystems can change makes a large difference to the kinds of roles they can play. Some physical states can change only by varying one or a few quantitative dimensions, e.g. voltage, temperature. Software developers are now accustomed to a much richer variety of types of change, apparently more suitable for the design of intelligent systems: AI in particular has explored systems making use of changes such as the creation or destruction or modification of symbolic structures in the forms of

propositions, trees, networks of symbols. Often what determines the suitability of a mechanism for a functional role is whether it can support the right *kind* of variability and at a suitable *speed*. It is difficult for physical mechanisms to change their structure quickly, so the full range of structural variability at high speed may be achievable only in virtual machines rather than physical machines.

- What 'flows' in causal channels: In many control systems the nature of the causal link between subsystems can be described in terms of flow of something (e.g. amount of liquid, amount of electric current) or some physical quantity like force, torque or voltage. By contrast in intelligent systems the causal links may often best be described in terms of a flow of information (including questions, requests, goals, instructions, rules and factual information). The information that flows may itself have a complex structure (like the grammatical structure of a sentence) rather than being a measurable quantity.
- Remoteness and proximity of causal links: Some causal links between substates are tight and direct links between substates of directly interacting submechanisms, for instance certain reflexes, whereas other links are *loose* and *indirect*, such as the causal links between input and output channels where the interaction is mediated by many other internal states. The causal connection between something like a preference for socialism and actual behaviour (e.g. political canvassing) would typically be extremely indirect and mediated by many mechanisms.
- Hierarchical control structures: Some internal control states (e.g. desire-like states) may produce behaviour of a specific kind fairly directly whereas others (e.g. high-level attitudes, ideals, and personality traits) work through a control hierarchy, for instance, by changing other desire-like states rather than directly triggering behaviour. The next figure gives an approximate indication of this. States that are at a high level in the hierarchy may be longer lasting, more resistant to change, more general in their effects, less direct in their effects, less specific in their control of details of behaviour (internal or external). For example a person's generosity is likely to be high level in this sense, unlike a desire to scratch an itch. Some control states are long term dispositions that are hard to change (e.g. personality, attitudes), others more episodic and transient (e.g. desires, beliefs, intentions, moods). Some of the control relationships are very direct: from sensory stimulation to action (as in innate or trained reflexes). Many of the high level states are complex, richly-structured, sub-states, e.g. political attitudes. Causal interactions involving these are both context-sensitive (dispositional) and (in some cases) probabilistic (propensities, tendencies), not deterministic. Engineers know about control hierarchies, but we need richer mechanisms than parameter adjustment

• .



• Time-sharing of causal channels: information channels may be shared between different subsystems, and between different purposes or tasks. This point is elaborated below.

These are merely some initial suggestions regarding the conceptual framework within which it may be useful to analyse control systems in general and intelligent control systems in particular. A lot more work needs to be done, including exploration of design requirements, specifications, designs and mechanisms, and analysis of trade-offs between different designs. All this work will drive further development of our concepts.

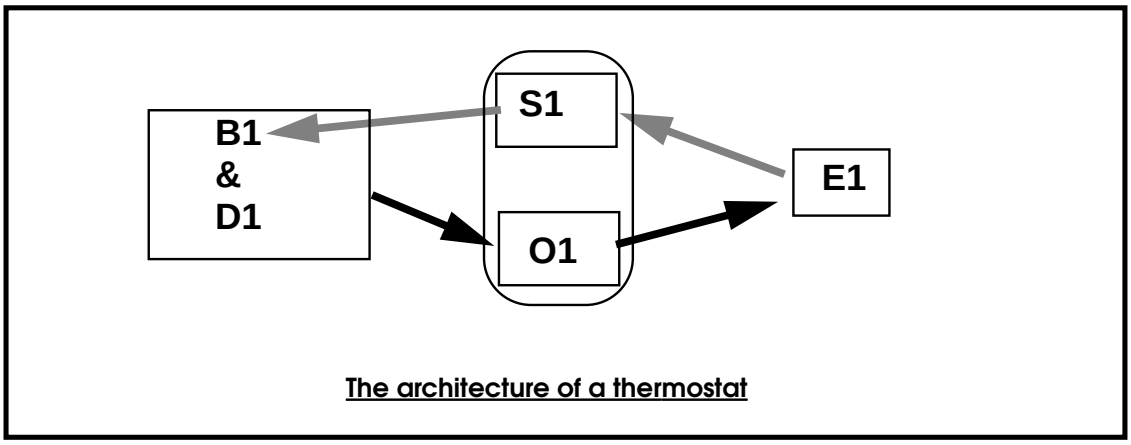
9 Control system architectures

Systems vary in their underlying *mechanisms* (e.g. chemical, neural, symbolic, digital, analog, etc.), and, more importantly, in their *architectures*. Within a complex architecture with many different components different (changeable) control substates may have different functional roles. There is a huge variety of possible architectures for control systems, depending on the number and variety of types of components, types of links, types of causal influences, types of variability of components, and the number and variety of higher level feedback loops implemented by the architecture. One way of beginning to understand the dimensions of variation in control system designs is to examine example systems to see how they can be changed to produce different systems with different capabilities. A full survey

would be many lifetimes' work, but we can get some idea of what is involved by looking at some special cases.

Thermostats provide a very simple illustration of the idea that a control system can include substates with different functional roles. A thermostat typically has two control states, one belief-like (B1) set by the temperature sensor and one desire-like (D1), set by the control knob.

- B1 tends to be modified by changes in a feature of the environment E1 (its temperature), using an appropriate sensor (S1), e.g. a bi-metallic strip.
- D1 tends, in combination with B1, to produce changes in E1, via an appropriate output channel (O1)) (I've omitted the heater or cooler.) This is a particularly simple feedback control loop: The states (D1 and B1) both admit one-dimensional continuous variation. D1 is changed by 'users', e.g. via a knob or slider, not shown in this loop.



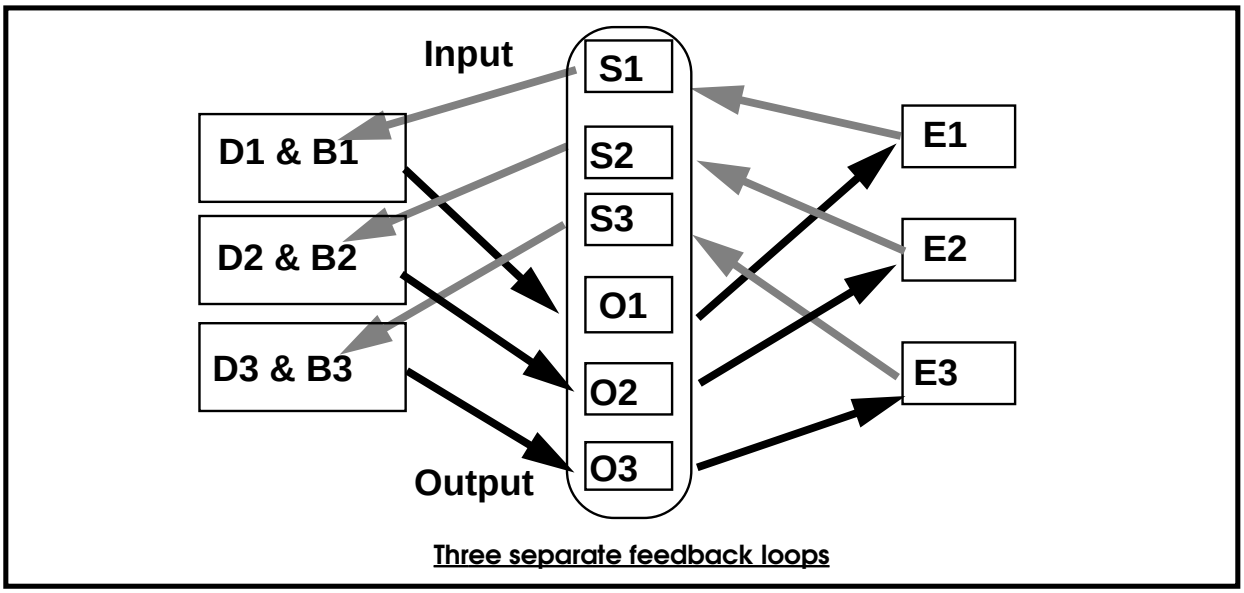
The architecture of a thermostat

Arguing whether a thermostat *really* has desires is silly: the point is that it has different coexisting substates with different functional roles, and the terms 'belief-like' and 'desire-like' are merely provisional labels for those differences, until we have a better collection of theory-based concepts. More complex control systems have a far greater variety of coexisting substates. We need to understand that variety. Thermostats are but a simple limiting case. In particular they have no mechanisms for changing their own desire-like states, and there is no way in which their belief-like states can include errors which they can detect, unlike a computer which, for example, can create a structure in one part of its memory summarising the state of another part: the summary can get out of date and the computer may need to check from time to time by examining the second portion of memory, and updating the summary description if necessary. By contrast the thermostat includes a device that directly registers temperature: There is no check. A more subtle type of thermostat could learn to predict changes in temperature. It would check its predictions and modify the prediction algorithm from time to time, as neural nets and other AI learning systems do.

Moving through design-space we find architectures that differ from the thermostat in the kinds of sub-states, the number and variety of sub-states, the functional differentiation of sub-states, and the kinds of causal influences on substates, such as whether the machine can change its own desire-like states.

Systems with more complex architectures can simultaneously control several different aspects of the environment. For example, the next figure represents a system involving three independently variable states of the environment, E1, E2, E3, sensed using sensors S1, S2,

S3, and altered using output channels: O1, O2, O3. The sensors are causally linked to belief-like internal states, B1, B2, B3, and the behaviour is produced under the influence of these and three desire-like internal states D1, D2, D3. Essentially this is just a collection of three independent feedback loops, and, as such, is not as interesting as an architecture in which there is more interaction between control subsystems.

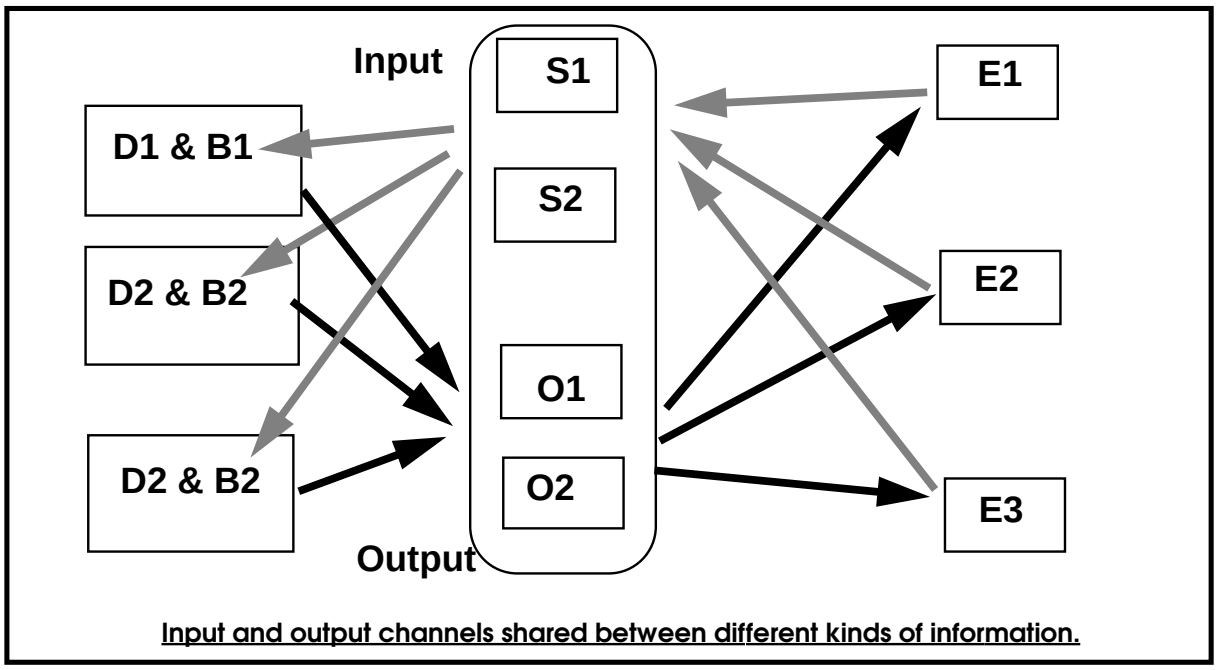


Three separate feedback loops

The architecture can be more complicated in various ways: e.g. sharing channels, using multiple layers of input or output processing, self monitoring, self-modification, etc. Some of these complications will now be illustrated.

An interesting constraint that can force internal architectural complexity occurs in many biological systems and some engineering systems: Instead of having separate sensors (Si) and output channels (Oi) for each environmental property, belief-like and desire-like state (Ei, Bi, Di) a complex system might share a collection of Si and Oi between different sets of Ei, Bi, Di, as shown in the next diagram. The sharing may be either simultaneous (with data

relevant to two tasks superimposed) or successive.



Input and output channels shared between different kinds of information.

Examples of shared input and output channels are:

- Sharing two eyes (S1, S2) between a collection of beliefs about different bits of the environment
- Sharing two hands (O1, O2) between different desires relating to the state of the environment, for instance pushing a door open whilst carrying a bulky object.
- Sharing large numbers of retinal cells and millions of visual pathways between processes of perception of several different objects simultaneously visible in the environment.
- Sharing millions of motor pathways among a smaller collection of tasks involving manipulating objects in the environment.
- Time-sharing input or output channels between different perceptual processes or different actions done in sequence. For instance first looking in one direction then in another direction, or first carrying one thing then another.

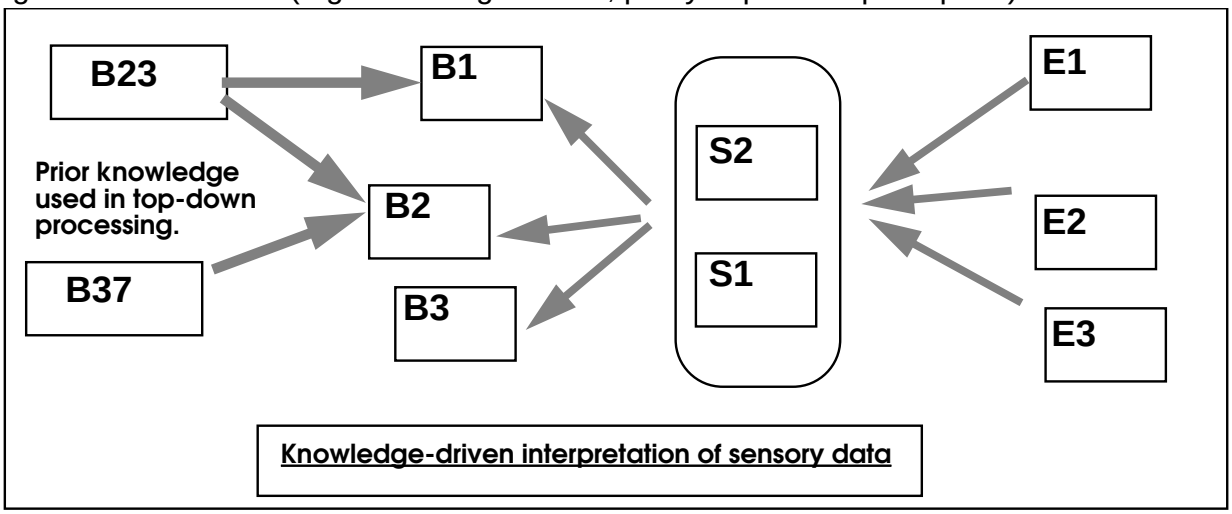
The need to decode information distributed over multiple sensory input channels, and the need to be able to compose control signals to produce coordinated contractions in many muscles in order to achieve a common goal are both requirements that lead to quite complex internal information processing. These may have been among the factors driving the evolutionary development of animal brains. Some of these points will now be illustrated in a little more detail.

10 Multi-layered bi-directional sensory processing

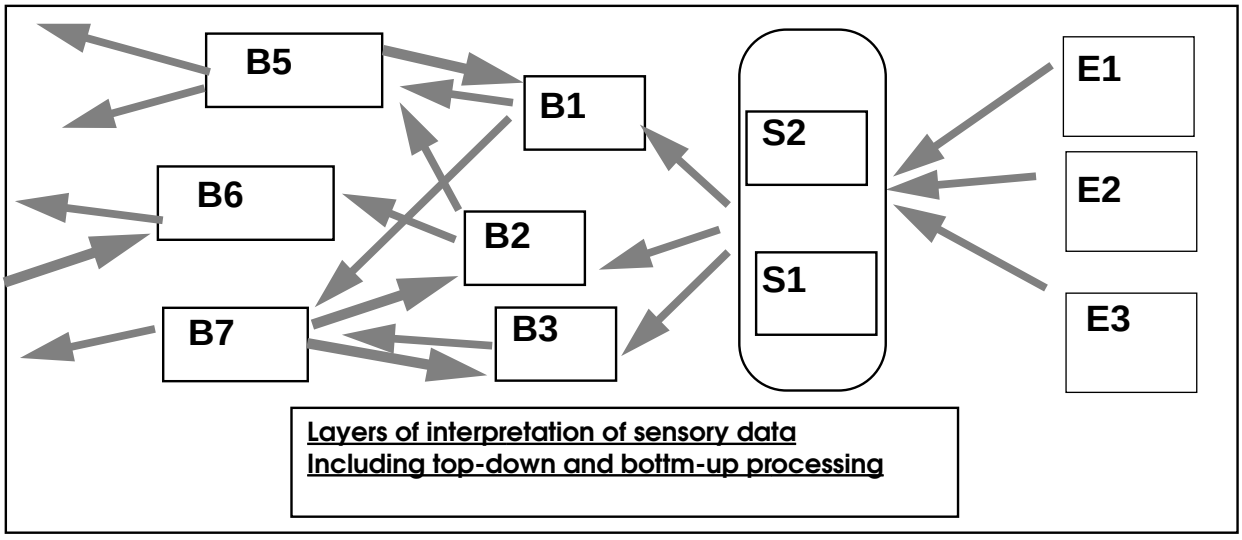
Production of belief-like states from sensory information can be more complicated than the examples presented so far.

- Sharing input channels between different Ei and Bi necessitates *interpretation* processes, to extract information relevant to different Bi from sensory 'arrays.' Often this requires specialised knowledge to play a role: general principles do not suffice for disambiguation. E.g. getting 3-D structure from 2-D visual arrays is a mathematically indeterminate

problem, yet human brains solve it very rapidly. For this reason, and for the sake of speed, or coping with noisy signals, some or all of the Bi may be produced or modified on the basis not only of *incoming* information, but also using *previously stored* particular or general information (e.g. knowledge-driven, partly 'top-down' perception).



- Many layers of interpretation may be needed: Sometimes it is impossible to extract information about the environment in one step. Different intermediate processes may be required, each producing different kinds of data, which may then be combined in the process of arriving at a single high level interpretation. For different purposes, different depths of processing of incoming information may be required, as shown in the next figure. (E.g. phonemes, words, phrases, meanings, theories.)



- In some cases the multi-layered processes may also include 'top-down' flow, with partial results in intermediate information stores used partly to control further processing at lower levels (nearer the sensory periphery). This is an example of an *internal* feedback control loop.
- Different layers of interpretation may use different forms of information storage: retino-topic, analogical, histograms, 'structural descriptions' (e.g. trees, networks), labels for recognised complexes, etc. Whether there is any single good general purpose shape representation is an unsolved problem in AI. It may be that very specific mechanisms are

required for creating visual percepts at different levels of interpretation (see figure in the next section).

- Different intermediate 'databases' may be used for different purposes. (E.g. some intermediate visual information stores are used, unconsciously, for posture control as well as contributing to perception and recognition of objects in the environment.) These different uses may need different 'inference' mechanisms as well as different representational systems.
- Some of the Bi may be stored for future use, or may modify previous long term information stores. Some Bi will be generalisations derived from many particular Bi. Some may be highly tuned specialisations derived from more general forms.
- Internal self monitoring is possible: some control loops involve only *internal* processes and substates, like a thermostat whose E1 is part of the internal virtual machine, not a property of the physical environment. An example is a computer operating system that keeps track of how much swapping and paging it does, or which builds internal summaries of some of its own internal structures, which it can also change, like building an index to a database. The development of internal self monitoring and self control submechanisms may be one of the factors that ultimately produced what we think of as human (self) consciousness, though this is a very muddled and ill-defined notion.
- Time-sharing of input channels may require inputs received *at different times* to be integrated for certain of the Bi. (E.g. looking at different parts of a house in order to grasp its structure). This requires temporary information stores that can continue to hold information after the sensory input has ended. There may be different information stores at different levels of processing, with different time delays.

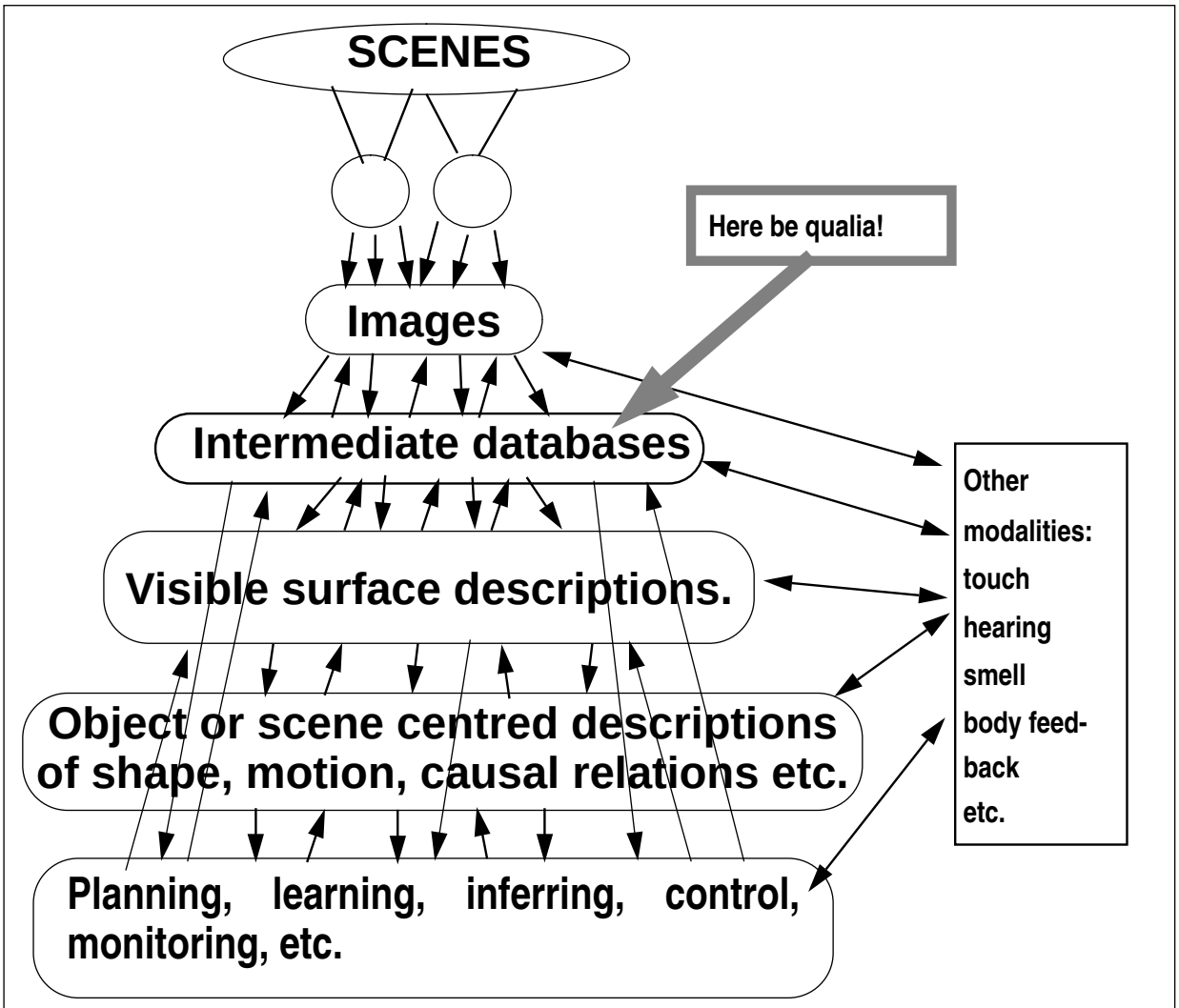
All of these points have implications for the architecture (the global design) of a perceiving agent. But we still understand very little about what the full requirements are for human-like perceptual processing, nor what kinds of designs are capable of meeting those requirements, nor what the trade-offs are between different solutions.

11 An example: perceptual architectures

Perception does not merely label things. Visual functions also include providing explanations ('that's how the clock works'), controlling actions (e.g. fine-grained control of movements) and many inner reflexes (e.g. being reminded, finding something or someone beautiful or repulsive). The sort of architecture that seems to be required in a visual system with these diverse capabilities illustrates many of the points already made. For example, a visual system typically (though not necessarily) includes retinal images (which are strictly themselves only rapidly changing samples of the less rapidly changing available 'optic array', as J.J.Gibson (1979) put it). In addition there appear to be requirements for several very different intermediate databases of information derived from a combination of retinal input and, when appropriate, other information. Examples of such intermediate databases are:

- Edge-maps, texture-maps, colour maps, intensity maps, optical flow maps, etc.
- Histograms of various sorts (Hough transforms)
- Databases of edges, lines, regions, binocular disparities, specularities (highlights), colour, etc.
- Groupings into larger 2-D substructures, recognizable 2-D objects, descriptions of their relationships (e.g. near to or overlapping in the visual field, relative size, etc.)

- Databases of 3-D shape fragments inferred from:
  - intensity and colour variation
  - optical flow
  - texture
  - stereo (binocular disparities)
  - edge contour information
- Groupings into larger 3-D substructures (e.g. surfaces, corners, limbs, eyes)
- Descriptions of 3-D shapes of visible objects, and their spatial, causal and functional relationships in the scene, and processes involving them:
  - spatial (inside, next to, touching...)
  - causal (pushing, pulling, pressing, twisting)
  - functional (part of, holding up, keeping shut, guiding)
  - intentional (walking towards, picking up, etc.)
- Names of types of things that have been recognized: e.g. a particular combination of 3-D surfaces, edges, corners, etc. may be recognized as a table, and another as a chair. Some recognition may be based on 2-D structures. Some names will label recognized actions, e.g. a pirouette, opening a door, pouring a liquid, etc.



A partial view of a visual architecture

The diagram above is an attempt to illustrate all this architectural richness in a visual system, albeit in a very sketchy fashion.

In human beings some, but not all, of the intermediate perceptual information stores are accessible to internal self-monitoring processes, e.g. for the purpose of reporting how things look (as opposed to how they are), or painting scenes, or controlling actions on the basis of visible relationships in the 2-D visual field. I believe that this is the source of the kinds of experiences that make some philosophers wish to talk about 'qualia'. From this viewpoint, qualia, rather than being hard to accommodate in mechanistic or functional terms, exist as an *inevitable* consequence of perceptual design requirements. Of course, there are philosophers who add additional requirements to qualia that make them incapable of being explained in this way: but I suspect that those additional requirements also make qualia figments of such philosophers' imaginations. Not pure figments, since such philosophical tendencies are a result of the existence of real qualia of the sort described here.

Vision, or at least human-like vision, is not just a recognition or labelling process: creation and mapping of structures is also involved, and this requires architectures and mechanisms with sufficient flexibility to cope with the rapidly changing structures that occur as we move around in the environment. I've tried to elaborate on all this in Sloman (1989) arguing that contrary to views associated with Marr, vision should not be construed simply as being a system for producing information about shape and motion from retinal input. There are other sources of information that play a role in vision, there are other uses to which partial results of visual processing can be put (e.g. posture control, attention control), and there are richer descriptions that the visual system itself can produce (e.g. when a face *looks* happy, sad, dejected, beautiful, intelligent, etc.)

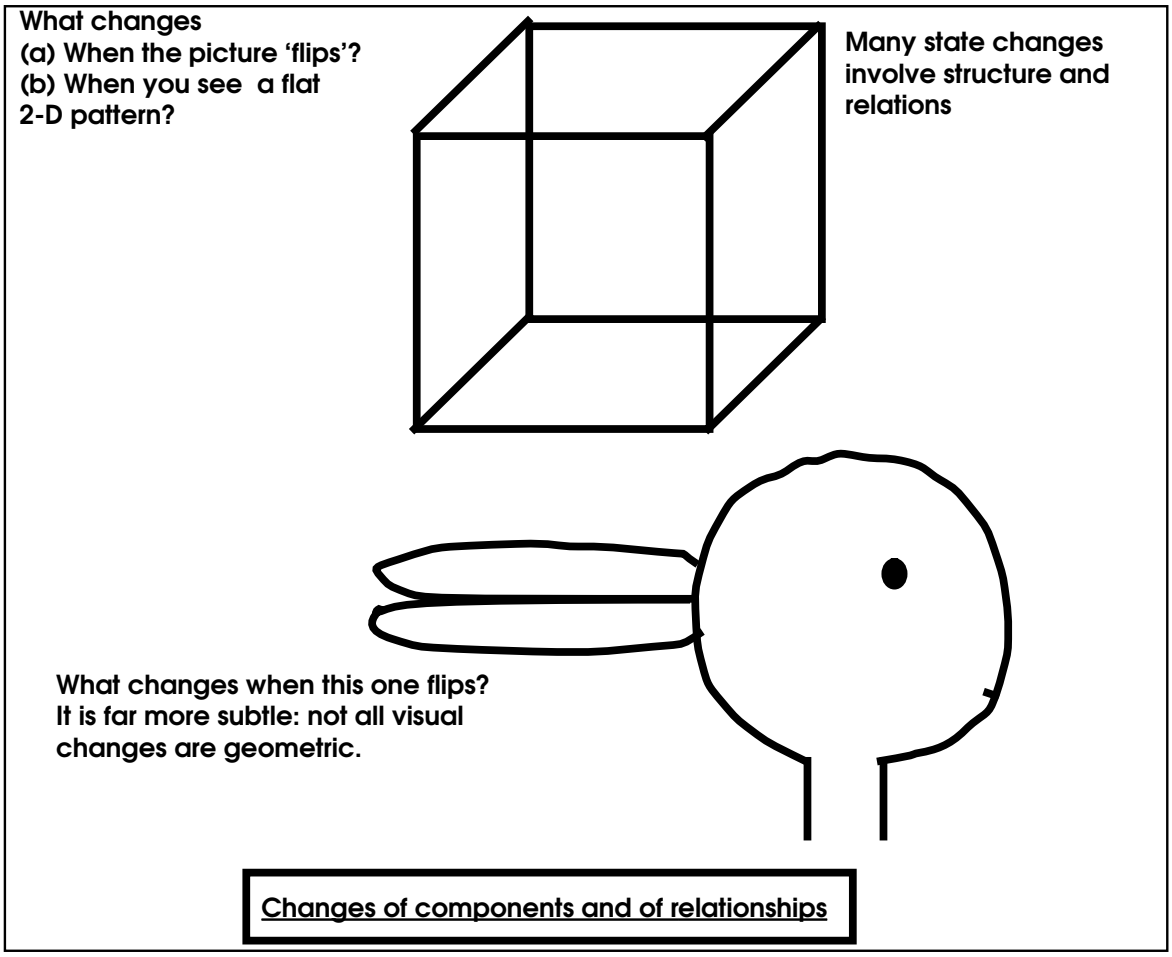
The internal information structures produced by a perceptual system depend not only on the nature of the environment (E1, E2, etc.) but also on the agent's needs, purposes, etc. (the Di) and conceptual apparatus. Because of this, different kinds of organisms, or even two people with different information stores, can look at the same scene and see different things. Many representational problems are still unsolved, including, for instance the problem of how arbitrary shapes are represented internally. Clues to human information structures and processes come from analysing examples in great detail, such as examples of things we can see, how they affect us, and what we can do as a result. I believe that every aspect of human experience is amenable to this kind of functional analysis, and that supposed counter-examples are put forward only because many philosophers do not have sufficient design creativity: most of them are not good cognitive engineers!

12 Kinds of variability in perceived structures

Different mechanisms (or parts of one mechanism) provide different kinds of variation. A temperature sensor requires only *linear* (continuous?) variation. A house-perceiver or sentence-understander needs *structural* variation. The next diagram illustrates some of the ways visual percepts can change in structure. The changes may be purely geometric or they may be more abstract and subtle, as when the duck-rabbit flips. Exactly what sorts of internal variability are required for different sub-mechanisms is still not understood, nor which mechanisms are capable of supporting which kinds of variability.

For example, it may be that variations during construction of a plan of action, variations during visual perception of a continuously moving object, and variations when wondering what conclusions can be drawn from some puzzling evidence all require very different internal structural changes, and that different sorts of sub-mechanisms are therefore

required.

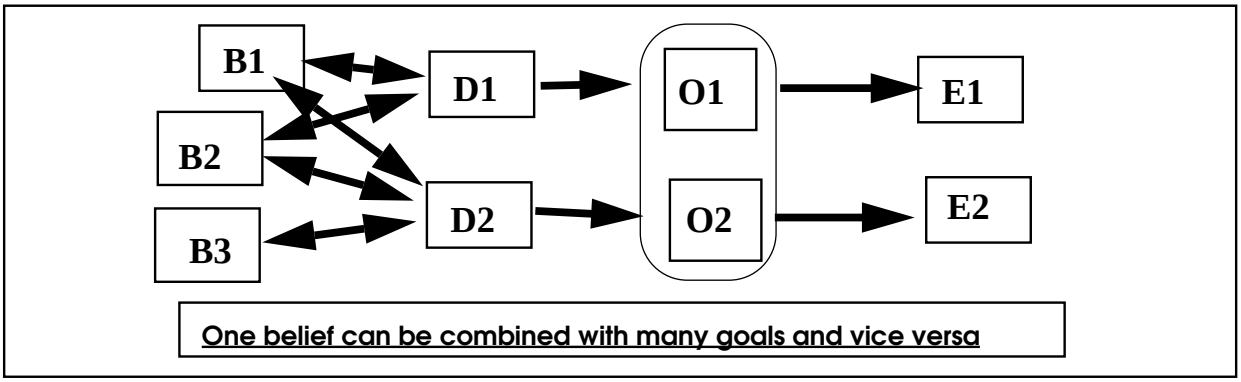


The kind of variability needed in Bi and Di states depends on both the environment (e.g. does it contain things with different structures, things with changing structures, etc.?) and the requirements and abilities of the agent. Compare the needs of a fly and of a person. Do flies need to see structures (e.g. for mating)? Do they deliberately create or modify structures? Rivers don't. There is lots more work to be done analysing the design requirements for various organisms in terms of their functional requirements in coping with the environment and with each other. This is one way in which to provide a conceptual framework for investigating the evolution of mind-like capabilities of different degrees of sophistication.

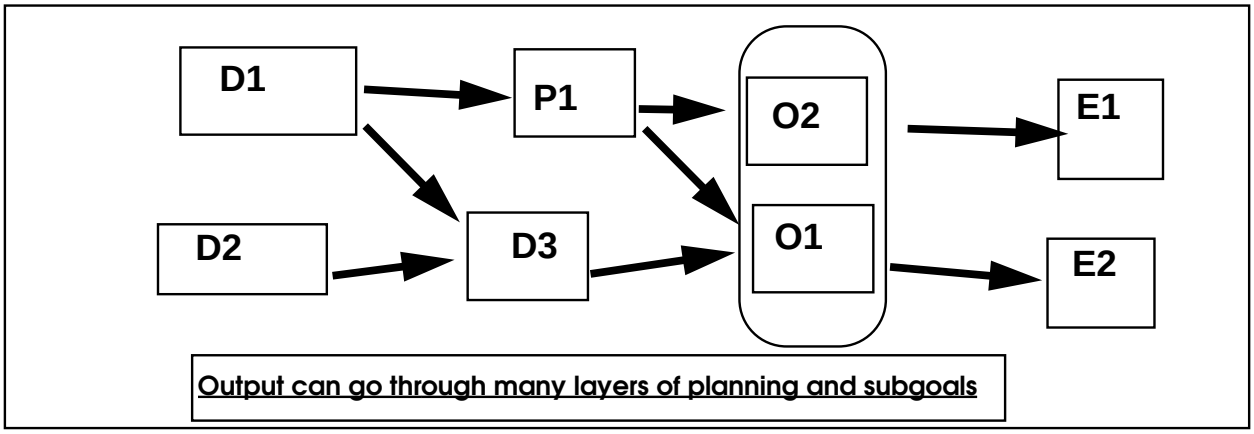
13 Architectural variety regarding desire-like sub-states

There are various ways in which the generation of outputs from desire-like states (in combination with belief-like states) may be more complicated than the examples shown so far. Some of these complications are analogous to the complications previously discussed in relation to processing of incoming information to create or modify belief-like states. In particular there can be shared output channels as well as shared input channels, and just as sensory interpretation processes may have multiple intermediate states, so can the output processes that generate behaviour.

- Information sharing: Particular Di may use several different Bi in producing output signals (e.g. using many facts in deciding whether and how to achieve one goal), and particular Bi can be used by many Di (e.g. using knowledge about cars both to help you drive, and help you avoid being run over)



- Causal links between Di and Oi may be indirect, via several layers of causation e.g.
  (a) going via planning mechanisms, and using different sub-goals to achieve a single goal
  (b) translating high-level to low-level instructions.



- Just as there is internal monitoring so can there be internal behaviour. Some Di change internal states, e.g. other Di and Bi. So some control is *self* control: e.g. making yourself concentrate on something. In that case some of the Ei are internal. (The mind is part of the environment, for itself) Desires themselves may be produced by deeper or higher level desire-like states (e.g. general attitudes, preferences, etc.) interacting with various Bi to produce new motives. So motivation can involve *hierarchies* of dispositions. (See earlier diagram of hierarchical control states.)
- Some Di are long term *dispositions* to produce various changes: they don't actually *do* anything until certain conditions arise. E.g. personality traits, and attitudes like racial prejudice. (Compare the previous comments on hierarchies of dispositional control states.)
- Some 'higher level' control states will not be concerned with particular goals or desires, but with principles or preferences for selecting between conflicting Di.
- Different intermediate Di-controlled sub-states in 'output' pathways may use different forms of information storage and transmission. (Compare layers of interpretation of

inputs.)
- E.g. having a thought, shaping a sentence, generating a syntactic form, selecting words, intonation patterns, stress patterns, volume, etc. may all require different intermediate data representations. Compare dancing, sculpting, assembling a clock.

- The Di need not determine *instantaneous* output: they may require *temporally extended* actions. This requires
  (a) Di states with rich internal structure (e.g. stored plans, with suitable temporary memory mechanisms)
  (b) 'output channels' with considerable sophistication (e.g. program-execution mechanism for 'translating' static plans into behaviour in time, rule-following mechanisms, etc.)
- In a system that is required to control continuous physical movement, it is likely that some of the output signals are not discrete instructions to 'motors' to perform complete steps. Instead there may be continuously varying output signals, such as a voltage or torque, whilst the effects of the behaviour thus produced are monitored continuously and the results used to modify the output: i.e. there are some continuous feedback control loops. An example would be the fine-grained control of motion of a violin bow so as to produce a sustained beautiful tone. Other cases may include a mixture of continuous and discrete monitoring and control, e.g. looking where you are walking, to make sure you are still on the intended route to your destination. A discrete high level signal could be an instruction to turn left at a certain corner. At lower levels control might still be continuous.
- The global control architecture itself may need to change as a result of learning. E.g. number and variety of Bi and Di (and other types of control sub-states) change over time, and new causal linkages develop:
  - A child eventually learns not to let the latest powerful motive dominate. What architectural changes enable the developing child to compare different motives, assess short and long term benefits?
- Some of the structures, and structural changes produced by the control processes, like changes in the Bi, may occur only in high level *virtual* machines.

14 What sorts of underlying mechanisms are needed?

The discussion so far is neutral as to what physical mechanisms are used to implement the various kinds of substates and causal linkages. They might be neural mechanisms or some other kind. As in circuit design, the global properties of the architecture are more important than which particular mechanisms are used, when the overall design is right.

'Architecture dominates mechanism'

The detailed mechanisms make only marginal differences as long as they support the design features required for reasons given earlier, such as:

- sufficient structural variability
- sufficient architectural richness
  - number of independently variable components
  - functional differentiation of components
  - variety of causal linkages
- sufficient speed of operation
- sufficiently smooth performance for controlling physical movement.

As we've argued above, 'virtual' machines in computers seem to have some of the required features, including rich structural variability and the ability to change structures very quickly. It may be that brains can also do this, though if they do it will also most likely involve another virtual mechanism, for it is not possible for networks of nerve cells to change their structures rapidly. In computers the virtual machine structures are usually implemented in terms of changing configurations of bit patterns in memory. Perhaps in brains it is done via changing configurations of activation patterns of neurones. In computers the same mechanisms are used for both short term and long term changes (except where long term changes are copied into a slower less volatile memory medium such as magnetic disks and tapes). In brains it seems likely that different mechanisms are used for long term and short term changes. For example in some neural net models the long term changes require changing 'weights' on excitatory and inhibitory links between neurones, and getting these changes to occur seems to require much longer 'training' processes than the changing patterns of activation produced by new neural inputs. (However, there are well known remembering tricks that produce 'one shot' long term learning.) It seems very likely that there are other kinds of important processes used in brains including chemical processes.

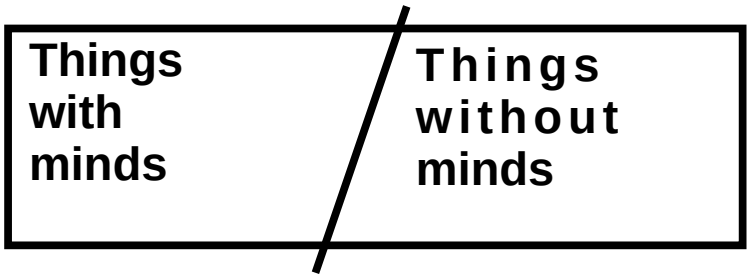
Whatever the actual biological implementation mechanisms may be it is at least theoretically possible that the very same functional architectures are capable of being implemented in different low-level mechanisms. It is equally possible that this is ruled out in our physical world because some of the processes require tight coupling between high level and low level machines, and it could turn out that in our universe the only way to achieve this is to use a particular type of brain-like implementation. E.g. it could turn out that, in our universe, *only* a mixture of electrical pathways and chemical soup could provide the right combination of fine-grained control, structural variability and global control. I have no reason to believe that there is such a restriction on possible implementations: I merely point out that it is a possibility that should not be ruled out at this stage.

But we don't know enough about requirements, nor about available mechanisms, to really say yet which infrastructure could and which couldn't work These are issues still requiring research (not philosophical pontificating!).

15 The shape of design space

I've suggested earlier that it is not enough to produce a single design: in order to understand the costs and benefits of particular designs we need to explore alternative designs in order to understand how they differ in the kinds of behaviours they support and their implementation requirements. Within the framework of such a design-based theory we may be better able to formulate sensible questions about how behavioural capabilities evolved in biological organisms, and instead of being faced with unanswerable questions such as 'Which animals are and which are not conscious?' we can hope to use new technical concepts for classifying natural and artificial behaving systems.

Many people feel that their concepts are so clear and precise that they can be used to produce a sharp division in the world. That is there is a major dichotomy like this:

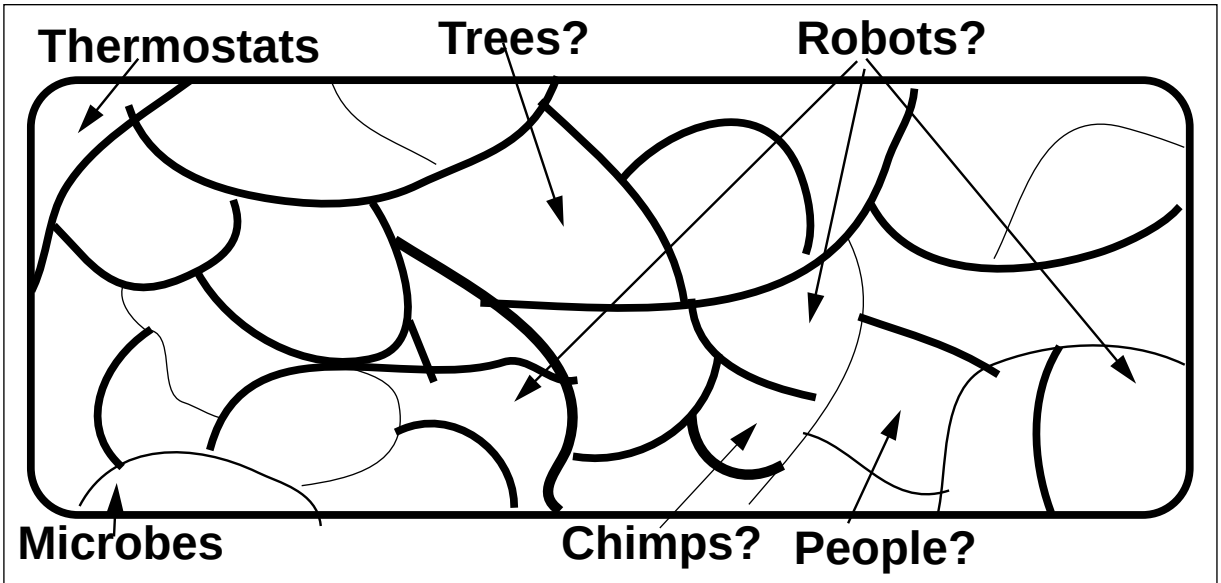


Unfortunately when they attempt to decide where the dividing line actually is they generally find it so hard to provide one, especially one on which everyone will agree, that many of them then jump to the conclusion that the space is a smooth continuum with no natural division, so that it's a purely a matter of convenience where the line should be drawn. So they think of design space like this:

This is a deep mistake: any software designer will appreciate that there are *many* important discontinuities in designs. For instance a multi-branch conditional instruction in a typical programming language can have 10 branches or 11 branches but cannot have 10.5 or 10.25 or 10.125 branches. Each condition-action pair is either present or not present.

Similarly, a machine can have skids for moving over the ground or it can have wheels, but there is no continuous set of transformations that will gradually transform a skidded vehicle to a wheeled vehicle: eventually there will be a discontinuity when the system changes from being made of one piece to being made of pieces that can move against each other (like an axle in a hole). If we think of biological organisms as forming a continuum then we fail to notice that there is a very important research task to be done, namely to explore the *many* design discontinuities in order to understand where they occur, what difference it makes to an organism whether it is on one side or the other of the discontinuity, and what kinds of evolutionary pressures might have supported the discontinuous jump. (Notice that none of this is an argument in support of a creationist metaphysics: it is a direct consequence of Darwinian theory that since acquired characteristics cannot be inherited there can only be a finite number of designs occurring between any two points in time, and therefore there must be many discontinuous changes, even if many of them are small discontinuities, such as going from N to N+1 components where N is already large.)

However it could well turn out that some of the discontinuities were of major significance. So we should keep an open mind and, for the time being assume that design space includes a large number of discontinuities of varying significance, some far more important than others. We could picture it something like this.



The shape of design space?

This picture is still too simple: e.g. it is single-layered, whereas different maps may required for different levels of design. There are still many design options and trade-offs that we don't yet understand. We need a whole family of new concepts, based on a theory of *design architectures and mechanisms*, to help us understand the relation between structure and capability (form and function).

16 Towards a general theory of attention

One implication of the kind of architecture sketched above is that there are typically multiple causal channels between sub-mechanisms. Thus any event or process occurring at one part of the system may have different effects elsewhere depending on which interactions are allowed to happen. This implies a need for many kinds of internal control in order to determine which causal channels are allowed to operate: which kinds of information are allowed to go to which sub-systems, and what is done with them. One example is deciding which subset of current sensory input should be processed and how it should be processed. Another example is deciding which current goals should be acted on and how they should be acted on.

Within this framework we can construe different kinds of attention in terms of different ways patterns of activity can be selected. The selection may involve changing which information is analysed, how it is analysed (i.e. which procedures are applied), and selecting where the results should go. Another example would be selecting which goals to think about or act on, and, for selected goals, choosing between alternative issues to address, e.g. choosing between working out whether to adopt or reject the goal, working out how urgent or important it is, selecting or creating a plan for achieving it, etc.

Some selections will be based solely on what is desirable to the system or serves its needs. However, sometimes two or more activities that are both desirable cannot both be pursued because they are incompatible, such as requiring the agent to be in two places at once, or looking in two directions at once or requiring more simultaneous internal processing

than the agent is capable of. The precise reasons why human thought processes are resource limited is not clear, but resource limited they certainly are. So the control of attention is important, and allowing control to be lost and attention to be diverted can sometimes be disastrous. The architecture should therefore include mechanisms that have the ability to filter out attention distractors.

These remarks are typical of the problems that arise when one adopts the design stance that would not normally occur to philosophers who don't do so. Their significance is that they point to the need for mechanisms in realistic, resource-limited, agents in terms of which mental states and processes can be defined that would be totally irrelevant to idealised agents that had unlimited processing capabilities and storage space. Thus insofar as it is part of the job of philosophers to analyse concepts that we use for describing the mental states and processes of *real* agents, and not just hypothetical imaginary ideal agents, philosophers need to adopt the design stance.

This can be illustrated with the example of a certain kind of emotional state. I have tried to show elsewhere (Sloman and Croucher 1981, Sloman1987, Beaudoin and Sloman 1993) that certain kinds of resource-limited systems can get into states that have properties closely related to familiar aspects of certain emotional states, namely those in which there is a partial loss of control of our own thought processes. Such capabilities would not be the product of specific mechanisms for producing those states, but would be *emergent* properties of sophisticated resource-limited control systems, just as saltiness emerges when chlorine and sodium combine, and 'thrashing' can emerge in an overloaded computer operating system. Our vocabulary for describing such emergent global states will improve with increased understanding of the underlying mechanisms.

There are many shallow views about emotional states, including the view that they are essentially concerned with experience of physiological processes. If that were true then anaesthetising the body would be a way to remove grief over the death of a loved one.

A deeper analysis shows, I believe, that the what is important to the grieving mother (and those who are close to her) is that she can't help thinking back about the lost child, and what she might have done to prevent the death, and what would have happened if the child had lived on, etc. There may also be physiological processes and corresponding sensory feedback but in the case of grief they are of secondary importance. The socially and personally important aspects of grief are closer to control states of a sophisticated information processing system.

Several AI groups are now beginning to explore these issues. But there is much that we still don't understand about design requirements relating to the sources of motivation and the kinds of processes that can occur in a system with its own motivational substates.

17 Further implications

Although the ideas sketched here do not constitute a full blown theory, but merely indicate the outlines of a research programme, I believe they have many deep implications for old philosophical problems about the nature of mind, the relations between mind and body, and the analysis of mental concepts. I shall conclude by drawing attention to an arbitrarily selected subset of these implications.

It is often said that a machine could never have any goals of its own: all of its goals would essentially be goals of the programmer or the 'user.' However, consider a machine that has the kind of hierarchy of dispositional control states described previously, analogous to very

general traits, more specific but still general attitudes, preferences, and specific desire-like states. Now suppose that it also includes 'learning' mechanisms such that the states at all levels in the hierarchy are capable of being modified as a result of a long history of interaction with the environment, including other agents. After a long period of interacting with other agents and modifying itself at different levels in the control hierarchy such a machine might respond to a new situation by generating a particular goal. The processes producing that goal could not be attributed entirely to the designer. In fact, there will be such a multiplicity of causes that there may not be any candidate for 'ownership' of the new goal other than the machine itself. This, it seems to me, is no different from the situation with regard to human motives which likewise come from a rich and complex interplay of genetic mechanisms, parental influences and short and long term, direct and indirect effects of interaction with the individual's environment, including absorption of a culture.

Issues concerning 'freedom of the will' get solved or dissolved by analysing types and degrees of autonomy within systems so designed, so that the free/unfree dichotomy disappears. (Compare Dennett 1984, Sloman 1978)

Exploration of important discontinuities in design-space could lead to the formulation of important new questions about when and how these discontinuities occurred in biological evolution. For example, it could turn out that the development of a hierarchy of dispositional control states was a major change from simpler mechanisms permitting only one control loop to be active at a time. Another discontinuity might have been the development of the ability to defer some goals and re-invoke them later on: that requires a more complex storage architecture than a system that always has only one 'adopted' goal at a time. Perhaps the ability to cope with rapid *structural* variation in information stores was another major evolutionary advance in biological control systems, probably requiring the use of virtual machines.

One implication of the claim that there's not just one major discontinuity, but a large collection of different discontinuities of varying significance is that many of our concepts that are normally used as if there were a dichotomy cannot be used to formulate meaningful questions of the form 'Which organisms have X and which organisms don't?', 'How did X evolve?' 'What is the biological function of X?' This point can be made about a variety of substitutes for X, e.g. 'consciousness', 'intelligence', 'intentionality', 'rationality', 'emotions' and others.

However, a systematic exploration of the possibilities in design space could lead us to replace the supposed monolithic concepts with collections of different concepts corresponding to different combinations of capabilities. Detailed analysis of the functional differentiation of substates and the varieties of process that are possible could produce a revised vocabulary for kinds of mental states and process. Thus, instead of the one ill-defined concept 'consciousness' we might find it useful to define a collection of theoretically justified precisely defined concepts C1, C2, C3... Cn, which can be used to ask scientifically answerable questions of the above forms.

This evolution of a new conceptual framework for talking about mental states and processes could be compared with the way early notions of kinds of stuff were replaced by modern scientific concepts as a result of the development of the atomic theory of matter.

18 The richness and inaccessibility of internal states and processes

One feature of the kind of architecture outlined here is that there are large numbers of active internal causal pathways, with many internal feedback loops. This makes the whole system

inherently unstable: internal states are constantly in flux, even without external stimulation. Most of the 'behaviour' of such a machine would then be internal (including changes within virtual machines). Moreover, since most of the causal relationships between external stimuli and subsequent behaviour in such a system would be mediated by *internal* states, and since these states are in a state of flux, the chance of finding interesting correlations between external stimuli and responses would be very low, making the task of experimental psychology almost hopelessly difficult.

For similar reasons, there would not necessarily be any close correspondence between internal control states such as the Bi and Di, and external circumstances and behaviour. So, for such a system, inferring inner states from behaviour with any reliability is nearly impossible. Moreover, if many of the important control states are states in virtual machines there won't be much hope of checking them out by opening up the machine and observing the internal physical states either. This provides a kind of scientific justification for philosophical scepticism about other minds.

Thus, even if design-based studies lead to the development of a new systematic collection of concepts for classifying types of mental states and processes it may be very difficult to apply those concepts to particular cases. This could be put in the form of a paradox: by taking the design stance seriously we can produce reasons why the design stance is almost impossible to apply to the understanding of particular individuals which we have not designed ourselves.

If some of the internal processes are 'self-monitoring' processes that produce explicit summary descriptions of what's going on (inner percepts?) these could give the agent the impression of full awareness of his own internal states. But if the self-monitoring processes are selective and geared to producing only information that is of practical use to the system, then it will no more give complete and accurate information about internal states and processes than external perceptual processes give full and accurate information about the structure of matter. Thus the impression of perfect self-knowledge will be an illusion. Nevertheless the fact that all this happens could be what explains the strong temptation to talk about 'qualia' felt by many philosophers. I have previously drawn attention to the special case of this where internal monitoring processes can access intermediate visual databases.

More generally, a host of notions involving sentience, self-monitoring capabilities, high-level control of internal and external processes including attention, and the ability to direct attention internally, including attending to 'qualia', could all be accounted for by a suitable information-processing control system.

19 Potential practical implications

The new conceptual framework could be of great practical importance in connection with improving the human lot. Human mental processes often seem to go wrong, for example multiple personalities, emotional disorders, learning disabilities. This is not at all surprising in such a complex system. In fact it is hard to understand how coherent control of such a system is possible at all, and why it doesn't go wrong more often. When things do go wrong, you can't hope to be much good at helping (therapy, counselling, training) without knowing the underlying design principles. Otherwise it's a hit and miss affair. (I.e. craft, not science or engineering. But some 'craft' skills are highly effective, even if we don't know why!)

When we have a good design-based theory of how complex human-like systems work it could lead us to many new insights concerning ways in which they can go wrong. This could, for example, help us to design improved teaching and learning strategies, and strategies for

helping people with emotional and other problems. If we acquire a better understanding of mechanisms underlying learning, motivation, emotions, etc. then perhaps we can vastly improve procedures in education, psychotherapy, counselling, and teaching psychologists about how minds work (as opposed to teaching them how to do experiments and apply statistics).

20 Intentionality and semantics

An issue that I have not yet addressed, but which exercises many philosophers, is how semantics can get into the system. What features of the design of a system make it possible for a machine to use one object to represent another? Which organisms are capable of having intentional states in which they somehow refer to objects, and why can't other organisms do it?

By now readers will be aware that such questions are based on the unjustified assumption that we have a precisely defined concept which generates a dichotomous division. This is an illusion, just like all the other illusions that bedevil philosophical discussions about mind. It's an illusion because our ability to represent or think about things is not a monolithic ability which is either entirely absent or all present in every other organism or machine. Rather it's a complex collection of (ill-understood) capabilities different subsets of which may be present in different designs.

One group of relevant capabilities involves the availability of sub-mechanisms with sufficiently varied control states for particular representational purposes. The kinds of variability in the mechanisms required for intermediate visual perception are likely to be quite different from the sorts of variability required for comparing two routes, or thinking about what to do next week. There are probably far more organisms that share with us the former mechanisms than share the latter. We can label the structural richness requirement a *syntactic* requirement.

Another group of requirements involves *functional* diversity of uses of the representing structures. Humans can have states in which they perceive things, wonder about things (e.g. is someone in the next room?), desire things (e.g. wanting a person to accept one's marriage proposal) or plan sequences of actions. Being able to put information structures to all these diverse uses requires an architecture that supports differentiation of roles of sub-mechanisms. Some organisms will have only a small subset of that diversity in common with us, others a larger set. A bird may be capable of perceiving that there are peanuts in a dispenser in the garden, but be incapable of wondering whether there are peanuts in the dispenser or forming the intention to get peanuts into the dispenser. (Of course, I am speaking loosely in saying what it can see: its conceptual apparatus may store information in a form that is not translatable into English. It's hard enough to translate other human languages into English!)

What exactly are the syntactic and functional requirements for full human-like intentionality, i.e. representational capability? I don't yet know: that's another problem on which there's work to be done, though I've started listing some of the requirements in previous papers (Sloman 1985, 1986). One thing that's clear is that any adequate theory of how X can use Y to refer to Z is going to have to cope with far more varied syntactic forms than philosophers and logicians normally consider: besides sentential or propositional forms there will be all the kinds of representing structures that are used in intermediate stages of sensory processing. Thus an adequate theory of semantics must account for the use of pictorial structures and possibly also more abstract representational structures such as

patterns of weights or patterns of activation in a neural net.

What convinces me that the problems of filling in the story are not insuperable is the fact that there are clearly primitive semantic capabilities in even the simplest computers, for they can use bit patterns to refer to locations in their memories, or to represent instructions, and they can use more complex 'virtual' structures to represent all sorts of things about their own internal states, including instructions to be obeyed, descriptions of some of their memory contents, and records of their previous behaviour. A machine can even refer to a non-existent portion of its memory if it constructs an 'address' that goes beyond the size of its memory. With more complex architectures they will have richer, more diverse semantic capabilities.

Being able to refer to things outside itself, or even to non-existent things like the person wrongly supposed to be in the next room or the action planned for tomorrow which never materialises, requires the machine to have a systematic and *generative* way of relating internal states to external actual and possible entities, events, processes, etc. Although this may seem difficult in theory, in practice fragmentary versions of such capabilities are already possessed by robots, plant control systems and other computing systems that act semi-autonomously in the world (Sloman 1985,1986). Of course, they don't yet have either the syntactic richness or the functional variety of human representational capabilities, but the question how to extend their capabilities is to be treated as an engineering design problem. Instead of proving that something is or is not possible, philosophical engineers, or design-oriented philosophers, should expect to find a range of options with different strengths and weaknesses.

Anyone who tries to prove that it is impossible to create a machine with semantic capabilities risks joining the ranks of those who 'knew' that the earth was flat, that action at a distance was impossible, that space satisfied Euclidean axioms, that no uncaused events can occur, or that a deity created the universe a few thousand years ago.

Acknowledgment

I am grateful for comments and criticisms of earlier versions of this paper and related papers, made by colleagues and students in the Cognitive Science Research Centre, the University of Birmingham.

REFERENCES

This is not a comprehensive bibliography, merely a list of items expanding on points made in the paper.

L.P.Beaudoin and A.Sloman (1993), 'A study of motive processing and attention', in A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds) *Prospects for Artificial Intelligence*, IOS Press, Amsterdam.

D.C.Dennett (1978) *Brainstorms*: *Philosophical Essays on Mind and Psychology,* Harvester Press, Hassocks.

D.C. Dennett (1984) *Elbow Room: The Varieties of Free Will Worth Wanting,* Clarendon Press, Oxford

D.C. Dennett (1991) *Consciousness Explained*, Allen Lane, the Penquin Press.

J.J.Gibson (1979) *The Ecological Approach to Visual Perception,* Lawrence Earlbaum Associates, (reprinted 1986)

G.Ryle (1949) *The concept of mind*, Hutchinson.

J.R.Searle (1980) 'Minds Brains and Programs' in *The Behavioral and Brain Sciences*, 3,3.

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

A.Sloman and M.Croucher (1981) 'Why robots will have emotions', in *Proceedings 7th International Joint Conference on Artificial Intelligence* Vancouver, 1981, also available as Cognitive Science Research Paper 176, Sussex University.

A.Sloman (1985) 'What enables a machine to understand?' in *Proceedings 9th International Joint Conference on AI*, pp 995-1001, Los Angeles, August 1985. (Also Sussex University Cognitive Science Research Paper 053)

A.Sloman (1986) 'Reference without causal links' in *Proceedings 7th European Conference on Artificial Intelligence*, published as J.B.H. du Boulay, D.Hogg, L.Steels (eds) Advances in Artificial Intelligence - II North Holland, 369-381, 1987. (Also Sussex University Cognitive Science Research Paper 047)

A.Sloman (1993), 'Prospects for AI as the general science of intelligence', in A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds) *Prospects for Artificial Intelligence*, IOS Press, Amsterdam