THE EVOLUTION OF WHAT?

Aaron Sloman

School of Computer Science & Cognitive Science Research Centre

Email: A.Sloman@cs.bham.ac.uk

WWW: http://www.cs.bham.ac.uk/~axs/

(Last changed: 1998)

Abstract

There is now a huge amount of interest in consciousness among scientists as well as philosophers, yet there is so much confusion and ambiguity in all the claims and counter-claims that it is hard to tell whether any progress is being made. This "position paper" suggests that we can make progress by temporarily putting to one side questions about what consciousness is or which animals or machines have it or how it evolved. Instead we should focus on questions about the sorts of architectures that are possible for behaving systems and ask what sorts of capabilities, states and processes, might be supported by different sorts of architectures. We can then ask which organisms and machines have which sorts of architectures. This combines the standpoint of philosopher, biologist and engineer.

If we can find a general theory of the variety of possible architectures (a characterisation of "design-space") and the variety of environments, tasks and roles to which such architectures are well suited (a characterisation of "niche-space") we may be able to use such a theory as a basis for formulating new more precisely defined concepts with which to articulate less ambiguous questions about the space of possible minds.

For instance our initially ill-defined concept ("consciousness") might split into a collection of more precisely defined concepts which can be used to ask unambiguous questions with definite answers.

As a first step this paper explores a collection of conjectures regarding architectures and their evolution. In particular we explore architectures involving a combination of coexisting architectural levels including: (a) reactive mechanisms which evolved very early, (b) deliberative mechanisms which evolved later in response to pressures on information processing resources and (c) meta-management mechanisms that can explicitly inspect evaluate and modify some of the contents of various "internal" information structures.

It is conjectured that in response to the needs of these layers, perceptual and action subsystems also developed layers, and also that an "alarm" system which initially existed only within the reactive layer may have become increasingly sophisticated and extensive as its inputs and outputs were linked to the newer layers.

Processes involving the meta-management layer in the architecture could explain the origin of the notion of "qualia". Processes involving the "alarm" mechanism and mechanisms concerned with resource limits in the second and third layers gives us an explanation of three main forms of emotion, helping to account for some of the ambiguities which have bedevilled the study of emotion. Theoretical and practical benefits may come from further work based on this design-based approach to consciousness.

A deeper longer term implication is the possibility of a new science investigating laws governing possible trajectories in design-space and niche-space, as these form parts of high order feedback loops in the biosphere.

Contents

NOTE: 25 Jan 2019
The Turing-inspired Meta-Morphogenesis project has added a lot of relevant material concerning evolution of information processing, since 2012
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

1 Consciousness is back in fashion, but...

Consciousness has become a fashionable topic in the past few years. There are workshops, conferences, journals, books, international email lists and no doubt many other manifestations of a revival of respectability. Despite this, there is evidence of widespread confusion and much argumentation at cross-purposes. This paper attempts first to identify some of the conceptual confusions, partly by articulating questions on which there appear to be irreconcilable differences of opinion, and partly by indicating sources of the confusion.

A strategy is outlined for replacing the muddled questions with new ones on which real progress can be made. I'll try to show that this is not a case of turning from deep and difficult questions to shallower and easier ones, as often happens when scientists propose new answers to old philosophical problems, like the proverbial drunkard looking for his lost keys only underneath the street lamp.

Notice that I am not claiming that we shouldn't talk about consciousness in scientific or philosophical contexts, but that we need to find a new way of talking about it instead of assuming that our pre-analytical concepts suffice to identify topics for investigation. The new strategy is based on the notion of types of *architectures* capable of supporting various types of mental states and processes. This provides a framework for enriching and refining our ordinary concepts with a collection of far more precisely defined concepts.

Unfortunately it is very difficult at present to put forward such a theory without triggering a host of preconceptions associated with shallow notions of representation, computation, functionalism and reductionism.

That is because although I'll be referring to mechanisms, representations, and functional relationships, I need to do so in such a way as to avoid commitments to a range of false assumptions, and it may be hard for readers familiar with different ways of thinking about these things to avoid drawing wrong conclusions from what I say. For example, I use a notion of representation which is not committed to any particular type of embodiment or any particular form of representation, or any particular type of syntax, or any particular type of semantic correspondence. That's because[1] information can be embedded not only in physical symbols but also in "virtual machines" of various kinds, and there are widely different types of syntax, widely different types of functional roles for representations, within a wide range of types of information users, human, animal and artificial.

For this and other reasons this paper has a bootstrapping problem: I cannot explain the main theory without using unusual variants of familiar concepts, and I cannot explain those concepts except in the context of the whole theory. All I can hope is that open-minded readers will find that the illustrative examples and the increasingly detailed presentation of the theoretical ideas will gradually enable them to grasp the concepts required. This will enable them to distinguish between relevant counter arguments and counter evidence and irrelevant ones directed at superficially similar theories.

I'll start by briefly summarising some of the apparently unanswerable questions and irresolvable disagreements. Instead of concluding that the whole subject is vacuous and not worth discussing at all, or that discussion is premature[2] I'll adopt a more constructive stance. This involves sketching, in several stages, the explanatory architecture generating a new conceptual framework, while explaining some of its implications, e.g. providing a basis for distinguishing concepts like "motivation", "mood", "attitude", and different sorts of emotion concepts linked to functional roles of different "layers" in the architecture. Various more ore less closely related notions of consciousness can be based on different architectures, and also on different subsets of mechanisms within a multi-functional architecture.

An organism whose (virtual machine) architecture is not static but develops after birth may therefore be capable of having different sorts of consciousness at different developmental stages. This may help us clarify questions about whether a new born infant is conscious or even whether a foetus can feel pain.

The architecture proposed for humans is linked to evolution insofar as it is conjectured that different aspects of the architecture evolved at different times. This raises questions about what sorts of pressures are likely to have influenced evolution, since many different architectures can in principle be functionally equivalent as regards the behaviour they can produce in a range of environments.

By drawing attention to some tradeoffs which are not always noticed in discussions of these issues (especially in some recent attacks on "good old fashioned AI") I hope to lend some credibility to the architectural ideas proposed, though these are tentative arguments open to refutation by further investigations, in the context of the broader study of trajectories in "design-space" and "niche-space".

Most of the ideas presented here are not new: though perhaps the particular combination is. For example, there are many similarities to the writings of Daniel Dennett.[3] However there are also some important differences which I'll spell out in a later section. In particular, whereas he tries to rule out any coherent notion of qualia I'll try to show how such a notion arises naturally within the framework of the architecture proposed. In particular, it is to be expected that robots designed with that sort of architecture will re-discover all the deep and puzzling questions about consciousness that have troubled human philosophers. And many ~~of the philoso~~phical robots will fall into the same confusions, for the same reasons.

[1] A point I have tried to elaborate previously, e.g. (Sloman 1971; 1978; 1985; 1994; 1996b)

[2] As I previously claimed in (Sloman 1990)

[3] E.g. (Dennett 1978; 1991; 1996)

2 Conflicting views on consciousness

I shall now outline some of the difficulties in talking about consciousness, which suggest that when we get involved in philosophical or scientific debates we may be deeply confused about what we are referring to, even if in non-academic contexts we use the notions without any problems (e.g. asking if an accident victim has regained consciousness yet.)

There are many disagreements over consciousness which at first seem to be disagreements on matters of fact, but after much debate appear to be irresolvable. This suggests that there may be some unclarity as to which questions are being asked. I'll briefly survey some examples in this section, before proposing a way forward.

2.1 Example disputes

One of the most familiar disagreements concerns which other animals have consciousness and what would count as evidence that they have it. Is a single-celled organism which flinches when touched conscious at all? Where should we draw the line between animals with and without consciousness? There does not seem to be any clear division, which leads some philosophers and scientists to say that consciousness is simply a matter of degree and perhaps even a bacterium has it to a small degree. Others think it cannot be a matter of degree: it is present or absent, though its contents can vary in kind and amount. It is not at all clear how such disagreements could ever be resolved.

There are many who believe consciousness can survive the destruction of physical bodies, and even relish (or fear) the prospect of an afterlife, while others say it can exist only in living organisms (and possibly robots of the future), so that it must end with bodily death – though they may disagree on what counts as death, a matter which could be of some concern to organ donors.

An issue on which there is ethical as well as factual disagreement concerns whether an unborn foetus has some sort of consciousness. Some regard reactions to noises and other external stimuli as indications that the unborn child has some sort of consciousness. Others argue that there cannot be full consciousness until the brain is fully formed, which happens only some time after birth. Both sides cite evidence about the reactions and brain structures of unborn or newly born infants. However, it is not at all clear what kind of evidence could possibly settle the question unambiguously.

This can also lead to disagreements over whether a newborn infant which opens its eyes and looks around is really conscious or whether it merely gives the impression of consciousness to admiring adults. Similar questions arise over whether the unborn or recently born child feels pain, or merely produces expressions of pain as part of a biological mechanism for triggering remedial actions in adults. Adoring parents generally have no doubt that their child, when awake, is conscious and experiencing its environment. A cautious scientist might regard the question as unsettled. Later I shall suggest a way of reformulating the question so that we can hope to find answers.

Even one individual may answer questions differently at different times. Most people would tend to regard it as obvious that consciousness is absent in unconscious states, and would regard the main difference between being awake and asleep as involving the presence or absence of consciousness. At other times some of them also feel inclined to say that they are conscious when dreaming, although fast asleep. Sleepwalking is another state in which our normal concepts seem to fall into disarray: the sleepwalker is asleep and therefore unconscious, yet sees the door and is therefore conscious.

There are those who say consciousness, or at least some of its main features, have been explained (see for example (Baars 1988; Dennett 1991)), those who say it cannot be explained, for it is not reducible to anything else, e.g. (Chalmers 1996)).

Some scientists assume that it has a biological function, while others say cannot, because all biological functions are adequately performed by physical processes in the brain and body and therefore consciousness, which is clearly not a physical phenomenon, cannot add anything and therefore cannot have a biological function.

On the first view, the evolution of consciousness is explained at least to the extent that organisms with consciousness have some biological advantage over similar organisms without it. On the second view, consciousness cannot confer any biological advantage and therefore could not have evolved through natural selection, and is therefore biologically inexplicable.

There are controversies about the relation of consciousness to other things. For instance, it is often claimed that having an emotion (e.g. joy, sadness, anger) involves consciousness of having it, whereas it is also often claimed that someone is angry, or afraid, or infatuated, or jealous without being aware of that fact, even if his friends are. Novelists and playwrights often use such scenarios, although they did not invent the phenomenon.

There are some who say a machine could have consciousness, while others deny that possibility, and among the former there are conflicting theories about whether consciousness can or cannot be implemented in computational systems or whether it requires analog circuits or hitherto unknown mechanisms, perhaps mechanisms which will be discovered by brain scientists, or physicists of the future.

There are clearly irreconcileable differences regarding the relevance of some arguments regarding consciousness, as shown by all the critical commentaries on (Searle 1980) and Searle's response. Searle and many of his admirers regard his "Chinese room" argument as a conclusive refutation of the claim that something like human consciousness (e.g. the sort involved in understanding Chinese text) can arise simply out of the implementation of some sort of computer program. His opponents disagree not only with him, but also among themselves: some argue that as long as the machine running the program is embodied in something like a robot actually sensing and acting on its environment there must be conscious understanding, while others argue that understanding (and consciousness) can be produced whether or not the system is causally embedded in a physical environment.

More generally there are disagreements on the question whether some sort of "disconnected" information processor (not necessarily computational) might be conscious or whether causal embedding in a body located in an environment that can be sensed and acted on is essential for the existence of consciousness or any sort of mental states with semantic content.[4]

Another issue on which entrenched positions seem to be unshiftable is whether consciousness (and more generally mental states with semantic content) can exist only in connection with an organism that has a biological evolutionary history, with mechanisms that proved their usefulness in a process of natural selection. A positive answer would rule out the possibility of machines with minds.[5]

Some people (whether aware of these arguments or not) claim to have built machines that already have consciousness (perhaps in a rudimentary form) while others argue that those machines have nothing remotely like consciousness: e.g. all they can do is classify input patterns, or follow rules mechanically, doing only the sort of thing the designer intended them to do, possibly with minor variations due to primitive forms of learning.

Some say that computers running trainable neural nets have a type of consciousness because they are not directly programmed but construct their own categories, whereas others respond that they are no more conscious than any other type of self-modifying computer program. There are those who say that if robots are ever built with "internal" (virtual machine) processes that have all the causal and functional properties of our

[4] See, for example, Dennett's brilliant and entertaining "brain in a vat" scenario in the final chapter of (Dennett 1978).

[5] The issue is discussed, for example in (Young 1994).

mental states and processes, then those robots will by definition have consciousness, since our consciousness is nothing more than a collection of internal functional capabilities and resulting states and processes.

Others argue that no matter which causal powers and functional relationships are replicated in a robot, it is still a possibility that it will lack consciousness and simply be a "zombie". I shall return to that issue later, since people understand "functional" in different ways. E.g. some take "functional" to refer to *external* behavioural capabilities whereas others talk about functions involving externally unobservable *mental* processes similar to processes occurring in what computer scientists and software engineers call "virtual machines": this leads to the idea of virtual machine functionalism, a notion which will become clearer later.

A more subtle form of disagreement concerns the question whether we have any mental processes of which we are not conscious and which are essentially similar in kind to those involved in consciousness. For example one familiar type of conscious process is following a rule, whether memorised (as a novice car driver remembers the sequence of actions to perform when starting to drive), or read off a set of instructions, e.g. for installing a new software package. It is often suggested that very similar processes can also happen unconsciously. For example, very young children somehow pick up the rules of the language in their environment, without apparently being at all conscious of what they are doing, at least not for the first few years. Some people infer from this that unconscious mental processes must be involved both in the learning of those rules, triggered perhaps by what the child hears consciously, and also in subsequently using the rules (including over-using them, as when a child uses "hitted" for the past tense of "hit", though he has never heard an adult say that).

Fluent adult speakers of a language implicitly use many linguistic rules both in producing and in understanding sentences, yet, unless they have taken courses in theoretical linguistics they are generally totally unconscious of any of the rules. This suggests that speech comprehension and production both involve many unconscious processes in which linguistic rules are used. Others argue that the fact that we can *describe* the linguistic products in terms of rules does not imply that there is anything like rule-following actually going on unconsciously: it could simply be a vast collection of physical processes which happen to produce the same result, but without anything in those processes corresponding to an encoding of the rules or any mechanism for interpreting such an encoding.

Unfortunately this leads into another morass of ill-defined questions regarding whether there is any absolute difference between processes which do involving rule-following and processes which do not, and a closely related debate about whether brains use representations or not, a distinction I have criticised at length in (Sloman 1996c), since instead of a distinction between systems with and systems without representations we need to consider many dimensions in which the types of information used by systems can differ, including how the information is stored, how it is manipulated, how it is used, and what it is used for.

Often the arguments on one side look compelling to their proponents and seem irrelevant or just wrong to their opponents, without there being any way of settling the issue: a very reliable indicator of conceptual confusion and arguments that are at cross purposes because the disputants cannot tell when they are talking about different things.

The danger in all these discussions is starting from the assumption that we know what questions we are asking, when we may simply be (unwittingly) deceiving ourselves.

2.2 One concept or many?

Underlying all this confusion, and partly explaining it, is confusion and disagreement about the definability and uniqueness of the concept of "consciousness".

Some people think it is blindingly obvious what words like "consciousness" and "subjective experience" refer to because we all have direct awareness of the referent, whereas others argue that such ostensive definitions fail to identify anything unique (just as thinking about two simultaneous experiences does not define "simultaneous").

Some think the concepts are riddled with ambiguity and confusion and inherently unsuitable for use in serious scientific debate or research: this was for a long time a common view among scientific psychologists. Others accept that there is considerable ambiguity and confusion at present, while allowing that there is a subject matter suitable for scientific investigation, and that progress is possible following conceptual clarification.

Dennett, at times (e.g. (Dennett 1978)) argues as if there is no *factual* question whether humans, animals or robots actually have consciousness or any mental states. Rather it is more like a *practical* question, e.g. whether it is convenient in the long term to adopt "the intentional" stance in thinking and talking about them. Others (Nagel 1981) argue that what it is like to be a bat is a matter of fact, though not something that can be investigated by any of the methods of science. David Chalmers has even argued in (Chalmers 1996) that there might be a law of nature linking consciousness to certain physical configurations.

When there is so much disagreement both concerning which things are or can be conscious, and also higher order disagreements regarding what sort of evidence or argument could relevant to answer the question, then something is probably wrong with the concepts used to pose the question. The best way to demonstrate this is to provide a new more powerful collection of concepts, which is the aim of this essay.

If the ideas presented below are correct, it will turn out that many such discussions, including some written by distinguished scientists and philosophers, are based on deep conceptual confusions, due in part to the radical ambiguity of the word "consciousness" (some revealed for example in (Block 1995) and the commentaries on his paper), so that there is no clearly definable "it" of which all the questions can be asked, such as: How did it evolve? Does it have a function? Do animals have it? Could robots have it? Can it be absent if all the behavioural manifestations of it are present? and so on.

If there is so much ambiguity and confusion would it not be best simply to ignore all this discussion that apparently leads nowhere, and get on with real work, such as finding out how brains work, finding out the details of human sensory, cognitive, motivational and emotional processes and their development, and exploring the detailed similarities and differences between different sorts of animals?

Perhaps, but some of the people who claim to be studying "it" are actually doing really interesting work finding out about various capabilities of human and animal brains and how those capabilities are explained, and how they can go wrong. For instance, there are really important questions about how different streams of visual information entering via two eyes merge into a single percept of a 3-D environment, and why it is that sometimes different visual information presented to two eyes leads to binocular rivalry so that only one is experienced at a time. And there are many important questions about the effects of drugs and brain damage or disease, where the effects are not only visible to observers but change the experience of the afflicted person.

I'll argue below that there is a core collection of important ideas which we can reconstruct within a powerful explanatory framework. Then we can ask new deeper questions, to guide both empirical and theoretical research, and perhaps help us with important practical problems, such as how best to help emotionally disturbed people or how to devise educational programmes that maximise learning opportunities.

Instead of either rejecting consciousness as a confused topic unfit for scientific study or assuming that the questions as originally posed are important and have right and wrong answers, we can look for a way of reconceptualising the phenomena and the problems so that we can make new progress and put aside our irresolvable disputes, when we understand how we have been arguing (to some extent) at cross purposes.

But we must do this in such a way as to avoid the charge that we are simply ignoring the interesting phenomena that raised all the questions in the first place, or that we are simply choosing to formulate questions which our current methods of investigation can answer (searching only under the street lamp).

3 Towards a multidisciplinary solution

Although something very new is needed, it can build on work in progress. Human consciousness, whatever it is or is not, includes a host of specific abilities including seeing, hearing, solving problems, taking decisions, making plans, learning, having desires and emotions, and many more. Instead of assuming we all know what consciousness is and then asking which animals have "it", how "it" evolved, what "its" function is, etc. perhaps we can make more progress by asking such questions about the component abilities.

Even if there isn't any unique, well defined, "it" worth talking about, there may still be a whole cluster of different but related things that are worth talking about, and which at least some of the discussants are referring to, even if they don't always realise exactly what they are referring to.

These phenomena are being studied not only by brain scientists and psychologists, but also by researchers in AI investigating how to give machines specific abilities modelled on human characteristics. Because of present limitations of computer speed and memory capacity, and above all limitations in our own engineering know-how and our understanding of the problems, only very simplified models have been built so far. Nevertheless there is progress and I shall try to show later how to build on it.

We need to learn how to integrate, within a unifying philosophical framework, the empirical research of biologists, psychologists, social scientists and brain scientists, and also the new design concepts and explanatory models of AI theorists, software engineers and electronic engineers. I believe that within such a framework the real progress in various kinds of research can be accommodated and futile questions and debates can be clearly distinguished from fruitful ones.

Meanwhile, anyone proposing a new theory of consciousness or a new definition of "consciousness" should reflect on how a sincere, intelligent, well-informed scientist or philosopher might raise objections to it. Almost certainly someone already has.

3.1 Empirical questions, design questions and conceptual questions

I shall outline some ways in which we can hope to avoid muddles and futile debates and make progress through collaborative, multi-disciplinary research. By doing this work we are sure to learn *something*, though we may not know what till much later!

This paper merely sketches a *research programme*, based on an outline theory which is put forward not as something for which there is already compelling evidence nor as a presumed final solution, but merely as a simplified example of a class of theories that we need to explore. In order to avoid tedious repetition of qualifying phrases, illustrative tentative theoretical ideas and conjectured implications will simply be described *as if* they were established.

Testing these ideas will be a long term collaborative process, involving empirical research, design and implementation of exploratory models, conceptual analysis and testing through both laboratory experiments and applications such as therapeutic techniques based on evolving theories.

We need to distinguish three different types of questions which are not always clearly separated: empirical questions, design questions and conceptual questions. Design questions are at the heart of our problems, since by exploring them we can make conceptual advances, which in turn will enable us to formulate new empirical questions replacing existing questions which look clear but are full of deep ambiguity and confusion.

3.2 Empirical questions (and conceptual confusions)

Empirical questions are about what exists in the world, what evolved when or how, what is correlated with what and what happens if you poke something, give it electric shocks or whatever. Empirical questions may be concerned with the states, properties, capabilities, etc. of *particular* objects or regions of the universe, or with *general* laws constraining possible combinations of states, properties, behaviours in classes of objects.

Deeper empirical questions seek to go beyond what can be directly observed and measured. However, that requires development of theories about hidden structures and processes. Only within the framework of such a theory which extends our ontology can the deeper empirical questions be formulated.[6].

Often a question which superficially appears to be empirical may be too confused, or based on concepts which are too ill-defined, for the question to have any answers, even if the questions "feel" clear and intelligible to us.[7] Prior to Einstein's devastating critique of ordinary concepts of simultaneity people thought they understood what it meant to talk about any two arbitrary events being simultaneous, and thought questions about which of two events occurred first were unambiguous. Einstein's special theory of relativity provided a new conceptual framework within which simultaneity and temporal ordering depended on a frame of reference, so that whether event E1 is be before, after, or simultaneous with event E2 depends on the frame of reference within which the measurements are being made.

At one time the question: "How fast is the earth moving through space (or through the aether)?" appeared to be a sensible empirical question. Michelson and Morley even designed an ingenious experiment to find out the answer. Instead it eventually turned out that the experiment tested a different question about the invariance of the speed of light as measured in different contexts. Experiments do not necessarily test or measure what the experimenters *think* they do, especially when they are concerned with such ill-defined pre-theoretical concepts as consciousness.

Likewise much empirical research in psychology and neuroscience laboratories is interesting and valuable, until the researchers start claiming that they have discovered something important about consciousness, as if consciousness were a well defined topic for research, like influenza, magnetism, or carbon.

If, as suggested throughout this paper, the concept of "consciousness" currently used by philosophers and scientists is ambiguous and muddled, the empirical questions and empirical claims (e.g. about neural correlates of consciousness) will also be muddled, even if the experiments themselves are interesting and tell us useful things about human or animal brains and human mental functions, for instance which bits of the brain react to which sorts of sensory stimulation, which sorts of sensory stimulation produce binocular rivalry, how the auditory system locates sounds in 3-D space, which capabilities are lost or modified by which forms of brain damage, which cognitive or emotional disorders are linked to particular chemical deficiencies, which visual capabilities remain in patients with blindsight due to cortical damage, and so on.

When empirical research is claimed to shed some light on consciousness, disagreements often ensue regarding either the objectivity of the experiments (e.g. if they involve introspection) or the relevance of the experiments (e.g. if they involve measurements of brain processes). E.g. there are always those who will say that a finding about the brain is not about consciousness as such, but about some biological capability which is closely related to consciousness. I shall try to show below how to avoid such irresolvable disputes. There is, however, no knock-down argument. Readers are simply invited to try out the new approach for a few years, to see whether it offers deeper insights in the long run!

3.3 Design questions and ontologies

Design questions are about which kinds of mechanisms and capabilities can be assembled in various ways in order to achieve various new kinds of capabilities. The assembly may or may not involve physical

[6] A distinction can be made concerning empirical questions about *form*, which are about what sorts of things can exist, what sorts of things are possible, and which laws limit those possibilities, and empirical questions about *content*. Many philosophies of science ignore the distinction or get it wrong, regarding form as concerned entirely with laws. See chapter 2 of (Sloman 1978)

[7] Consider the question: What is the time in London when it's noon on the moon?

construction. For instance, software engineering normally does not.

Engineers, whether concerned with large scale or with small scale systems, whether concerned with physical structures or with abstract software structures, all have to think about designs. So do biologists, brain scientists and psychologists attempting to make sense of existing systems. Some designs provide a basis for creating new systems, whereas others form part of the explanation of existing systems, for example when scientists explain how the design of the cornea and its associated system of muscles enables sharply focused images to be projected to the retina.

This does not imply that somebody designed the animal optical systems, merely that the animal eyes in question involve some general principles which can be deployed in different organisms or in machines, to achieve similar functions. Similarly the design of a bird's wing, like that of an aeroplane's wing, explains how forward motion can generate lift by forcing air to travel further along the upper surface than the lower surface. General principles of aerodynamics apply both to natural and to artificial designs.

Designs can involve assemblies based on very varied ontologies. At one time physical assemblies with mechanical or hydraulic relationships and linkages were the main type of design studied. There has also long been an interest in social assemblies, in human social systems and political organisations of various sorts. More and more complex electronic assemblies were investigated throughout the 20th century, including designs concerned with generation, transmission and transformation of energy, and also designs concerned with transmission of information.

During the last half of the century some of the most important designs have been concerned with storage, transformation and use of information, with or without transmission (except within a machine). Increasingly, design know-how has focused on ways of combining information manipulation mechanisms (some computational, some not) to create more and more complex systems for manipulating information. Some of these interact closely with a human user (e.g. a word-processor or interactive theorem prover), some only with other machines (e.g. an automated factory control system) and some mostly interact with themselves (e.g. a computer operating system which manages complex resources such as main memory, backing store memory, one or more CPUs, all used by a changing collection of competing computational processes). An excellent example involving Tandem Corporation's design for ultra-reliable computers can be found on page 72 of (Picard 1997).

All this has led to development of a fast growing collection of ideas and techniques relevant to the construction of "virtual" or "abstract" machines in which the components are not mechanical, hydraulic or electronic mechanisms such as can be observed and measured using instruments of physics, but more abstract mechanisms such as image analysers, word processors, parsers, compilers, planners, theorem provers, constraint analysers, schedulers, file-managers, etc.

That is one area in which ideas about "design-space", the space of possible designs for working systems, have expanded rapidly in recent years, though it is arguable that we are still only on the threshold of a vast uncharted territory.

Another area of design-space concerns biological mechanisms of many types involved in development, reproduction, resistance to disease, and many brain processes. Larger scale biological mechanisms involving symbiosis, competition and food and energy chains in ecosystems are also part of biological design-space.

Here too our knowledge at the end of the 20th century is fragmentary and in years to come will probably turn out to have covered only a miniscule fraction of the space of designs already implicit in biological organisms and ecosystems. There are signs of considerable overlap between regions of design-space concerned with biological systems and regions concerned with systems which manipulate information. For instance it is already commonplace to refer to DNA as storing information for transmission from one generation to the next, and immune systems seem to use a type of learning. Brains are clearly self-modifying information-processing control systems, although different brains instantiate different designs.

The study of designs for organisations and social systems is very old, going back at least as far as Plato's *Republic*, and is found also in more recent work on political theory, management theory, sociology and anthropology.

It is very likely that the set of concepts currently available for thinking about designs for information processing systems, whether biological, social or artificial, will turn out inadequate and will have to be extended or replaced as our understanding grows, continuing the sort of development which has been accelerating rapidly in the last fifty years.

The concept of "design-space" is fairly familiar. However, there is a closely related notion of "niche-space" which is not so familiar, but will be used below, though it is harder to characterise precisely.

The biologist's notion of an organism's niche is very closely related to the engineer's notion of a set of requirements for a machine or other designed system. Both are abstractions involving notions like tasks, constraints, costs, and an environment or context. To a first approximation we can think of a niche as defined by a description of the requirements which an organism or machine must satisfy in order to be successful in some way. The set of possible requirements specifications will form a space which we refer to as niche-space. Different niche-spaces may be separated out according to what counts as success, or according to other criteria.

More will be said about these two spaces later. For now, note that just as questions can exist which no questioner has asked, and mathematical theorems can exist which nobody has discovered, so designs do not need a designer and requirements do not need a requirer. They are abstractions which can have interesting features and relationships whether or not anyone thinks about them or is concerned about them, or whether they have actual instances or not.

For instance it may be a property of a particular type of design that it allows different inputs to be processed in parallel, and that it requires a particular kind of supporting hardware. The design has those features even though nobody has so far thought of the design, or produced an actual implementation. The design of a lift inducing wing had various properties long before the first flying animal evolved and certainly before any human designer thought about it.

A design does not suddenly acquire these properties when someone first thinks of it, any more than the square root of two suddenly acquired the property of being irrational when people began thinking about numbers. Designs, like numbers, are abstract entities which cannot come in and out of existence. Some readers will find this "platonistic" view unacceptable. There is no space for a full discussion here. I am not concerned with semantic quibbles about what "design", or "exists", really means in ordinary English. I have suggested elsewhere (Sloman 1992)), that arguments between platonists and anti-platonists are vacuous. The only important point for now is that neither designs nor niches have any necessary connection with human designers or requirers. We can talk about the design that is shared by two species of animals just as we can talk about the shape shared between two rocks. Different levels of abstraction can be identified: e.g. two rocks may be cube-shaped, or two rocks may be merely convex in shape. Likewise, two eyes may share designs at different levels of abstraction. Two birds may have eyes sharing a high level design with a mammal and a more specific design with each other. None of this presupposes that anyone has noticed the shape or produced the design.

We can also talk about a set of requirements which might or might not be satisfied by a certain class of designs, for instance controlling stock in a warehouse, or monitoring intensive care patients, or enabling novices to read and send email. Sets of such requirements, like designs, propositions, theorems are abstractions which have features and relationships to classes of designs, whether or not anyone thinks about them. A design which manipulates information by adding new options to a queue can meet a requirement for doing a breadth first search for a solution to a problem, whereas a design using a stack of options meets the requirement for a depth first search — whether or not anyone ever implements such an algorithm or requires such a search.

Much of the intellectual history of mankind can be seen as exploration of design-space and niche-space. However, only a tiny fraction of possible sets of requirements and possible designs will ever have been thought about at any stage of human history.

Our understanding of design-space and its (many and varied) relationships to niche-space (described below) is still very primitive, though our understanding is being extended by work in theoretical biology, artificial life, AI, the study of complex dynamical systems and all branches of engineering.

Unfortunately, we don't always realise how shallow our understanding of design-space is and so people are often tempted to presume they have a deep understanding and can make pronouncements about what certain classes of designs can and cannot do, e.g. pronouncements about the likelihood, or the impossibility, of computer-based systems, or systems based on classical physics, having consciousness. I shall try to outline a strategy for making progress based on knowledge and understanding, instead of prejudice and rhetoric.

I've argued elsewhere (Sloman 1978; 1993; 1995; 1997; 1999) that many old philosophical questions are now best discussed in relation to properties of designs. In particular if we see how a particular design, when implemented, explains a range of possible properties and behaviours for a working system, then we can use that as a basis for systematically generating a class of concepts for describing such a system and similar systems. This is partly analogous to the way in which a theory of the architecture of physical matter could be used to generate a taxonomy of types of elementary physical substances: the periodic table of the elements. I'll return to that analogy later.

3.4 Conceptual questions

Conceptual questions are commonplace both in philosophy (e.g. attempts to clarify what we mean by words like "good", "true", "possible", "experience", "design", "function") and also in deep science. Examples from science are Einstein's analysis of the concept of "simultaneity", biologists' attempts to clarify concepts like "function", "gene", "species", and the struggles of modern physicists to clarify concepts used in quantum mechanics.

Among the concepts requiring clarification in discussions of consciousness are not only the concepts used to describe the phenomena being explained, e.g. "consciousness", "experience", "awareness" and "feeling", but also the concepts that are used in formulating explanations, e.g. "mechanism", "computation", "function", "causation", "implementation", "architecture", "emergence", "supervenience", and "design".

It is not possible to start with any simple and clear definitions of any of these terms: even dictionary definitions tend to be biased towards a particular philosophical theory. The clarification we require can only come gradually, via development of powerful explanatory theories (Sloman 1996b). This contrasts with the common view that scientists *start* by defining their concepts and then use them to formulate hypotheses which are tested empirically. This view is refuted by the history of science, in which important new concepts often emerged only in the context of new theories. Only in relatively trivial cases do scientists start by understanding their concepts and the questions they formulate. Normally, after we have developed good theories we can use them as a basis for defining good concepts, and then, finally, understand what it was we were originally trying to explain.[8]

[8] I shall not attempt to defend this claim here as it would require an excursion into the history and philosophy of science. However, this paper is an illustration of the claim: for it attempts to show how new clearer concepts can emerge out of a new theory of how minds work.

In particular, I suggest that we should abandon the very compelling belief that we all know exactly what consciousness is or what we mean by “consciousness”, and related words like “awareness” and “experience”. I’ve given some reasons above for thinking that our concepts are muddled, because of their use in formulating apparently unanswerable questions. Moreover, if we start with a strong commitment to a mistaken analysis of our concepts we may be duped by fallacious arguments claiming to show that the phenomena we are interested in cannot be explained by particular types of theories, or perhaps by any scientific theory.

This includes trying to avoid assuming that consciousness, or the class of mental phenomena of which we are conscious, is some sort of self-contained realm of events and processes. We shall see that it is more accurate to regard what we are aware of as the small tip of a large iceberg of information processing mechanisms and processes to which we normally have no access and whose existence we unwittingly take for granted when we talk about consciousness.[9]

For instance, as the philosopher Immanuel Kant began to see over 250 years ago (Kant 1781), we cannot experience a visual scene, e.g. a view of a fast flowing river under a bridge, without relying on an infrastructure of unconscious processes analysing, comparing, grouping and interpreting. It is now clear that this involves vastly complex processes relating visual information of many types and on many scales, from small-scale intensity discontinuities through to global patterns of optical flow, or structural relations in scenes (Sloman 1989).

Likewise, you cannot experience a poem without relying on mechanisms involving the rules and structures of your language, lexical, syntactic, and semantic information and knowledge of the items referred to in the poem which resonate with each other and the sounds of the words.[10]

It is difficult to acknowledge our conceptual muddles and confusions when we think we have some sort of direct and unmediated access to what we are talking about, just as people once thought they had a very clear notion of *simultaneity* before Einstein revealed hidden complexities by asking what it means to say that two events some distance apart occur simultaneously. Later I shall present questions designed to undermine confidence that we all know what we mean by “consciousness”.

Much writing on consciousness, including most of what I have seen produced by scientists, ignores deep and difficult conceptual questions, or if they are noticed it is assumed that somehow we already know enough for the lack of analysis not to matter. For instance, some authors assume that first-hand knowledge about consciousness is enough to guarantee that we know what we are talking about. That is as unjustified as claiming that long before Einstein we all knew what simultaneity was because we had first hand, direct, knowledge about “it”. Einstein’s analysis showed that a concept that appeared to be simple and clear had hidden complexity, which could be unravelled in surprising ways. I shall try to explain why this is also true of many of the concepts used in discussions of consciousness.

The need for conceptual clarification becomes more evident if we consider diverse examples instead of thinking only about the obvious cases. We may think we know what we mean by “freedom”, until we consider examples of peer pressure, parental influence, genetic influences, advertising, cultural norms, hypnosis, various kinds of duress, effects of poverty, religious indoctrination, etc. We then realise that instead of freedom being a unique property which an individual either has or doesn’t have, instead there’s large collection of capabilities which can be present or absent in different combinations, where no particular combination is uniquely the *right* one (compare (Dennett 1984; Franklin 1995)).

[9] As Brian Logan pointed out to me: the iceberg metaphor can be misleading if it suggests that the same part of the iceberg is always above the surface. A rotating or oscillating nearly-submerged floating structure gives a better metaphor for the contents of consciousness in a person with constantly shifting attention, though all such mechanical metaphors for mental processes are ultimately inadequate.

[10] A point made by Turing in his famous 1950 article.

Likewise, thinking about the capabilities of other animals, of infants, of people with various kinds of brain injury and degenerative diseases, can help to focus attention on conflicting criteria for applying the concept of consciousness. When we understand that our concepts are far more muddled than we had realised, we can open our minds to new ways of thinking about them.

4 More evidence for conceptual confusion

Contradictions in our ordinary ways of thinking about consciousness can sometimes be revealed by asking probing questions that generate a tendency both to answer "yes" and to answer "no", or questions which are difficult to answer because boundaries are unclear. Examples of such questions (some of which were raised in the opening section) are:

- Are you conscious when terrified in a dream?
- If you first become aware of a noise when it stops, were you conscious of it before it stopped?
- Is a sleep-walker who dresses himself and walks downstairs conscious?
- Is a foetus conscious X weeks after conception? (Try various numbers in place of X.)
- Can we draw a line between animals with and animals without consciousness?
- When a degenerative brain disease gradually reduces a normal person to an apparent vegetable, at what point does consciousness disappear?
- Some forms of brain damage produce "blind-sight" – people claim not to be able to see, and yet they can answer some questions about where a light is. Are they conscious of the light without being aware of that fact? Is there some subsystem that is conscious of the light?
- Dissociations can be produced by brain damage, e.g. in people able to make precise anticipatory movements while performing a manual task, yet unable to state in words or show with a hand movement, what the movement would be without actually doing it. They are unable to indicate the angle at which a letter needs to be held to be posted into a slot even though they can rotate the letter to that angle if posting it. Are they or are they not conscious of the orientation of the slot?

4.1 Seeing yet not seeing

There are many examples where it is not clear whether something is or is not in a person's consciousness. Here is an example. Examine it carefully:

A
BIRD
IN THE
THE HAND

Many readers will have seen the phrase, sometimes shown displayed inside a triangle in books on vision. Some people can stare at it for several minutes and not see anything amiss, even when told there is something wrong (e.g. a subset of those reading this).

Yet if they are told to shut their eyes and are then asked certain questions (such as how many words there are, or where the "the" is), they sometimes realise that they did see something wrong, and can suddenly report it accurately. They saw it, but did not attend to all the details. Were they somehow conscious of it and not conscious of it at the same time? Can there be things which are in your consciousness but which you do not notice, even though you try hard to find the oddity which is staring you in the face? Were you really aware of it, in that case?

A related example involving no linguistic errors was reported in *New Scientist* on 24th May 1977 (inside back cover): count the number of occurrences of "F" in this sentence:

**FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC
STUDY COMBINED WITH THE EXPERIENCE OF YEARS**

It seems that most people at first manage to find only half of them. There really are six occurrences. Eventually everyone finds them by looking very carefully. Before that were they or were they not conscious of all the occurrences of “F”? After all, they read all the words and saw the 2-D printed patterns making up all the letters, including every “F”. Or did they?

Perhaps being conscious of a particular letter requires something more than that the information about the 2-D distribution of colour be registered. It seems to require that certain recognition abilities (which a video recorder does not have) be activated, and for some reason they are not (in most people) activated by all the appropriate “F” configurations in the above sentence, even though all the word configurations using those letters are recognised. So perhaps what you are conscious of depends on which of your capabilities are activated. It just so happens that activation of different collections of capabilities may be triggered by the same sensory data in different contexts. Different subsets may also be removed or impaired by brain damage, for instance in people who can recognise faces as faces yet can no longer recognise the faces of their spouses.

If being conscious involves deploying various capabilities, then we need a theory to explain which sets of capabilities are involved, how they are turned on or off, how they vary from one person to another, or one species to another, and how they can be impaired temporarily or permanently. In particular, we need to understand how different architectures can support different collections of capabilities: i.e. which kinds of consciousness are associated with different regions of design space.

Another case that causes confusion is how to describe what is happening at the ‘blind spot’. Look at the “X” below with your left eye closed or covered, and move the page back and forth. At a certain distance the “O” disappears. That’s because it is projected onto the part of your retina where all the optic nerve fibres dive through to get to the brain, so that no retinal cells can detect the light falling there.

X **O**

Shut one eye and look around you: where is the gap in what you see? Is there a gap or isn’t there? Is the blind spot actually filled it, or do we merely think it is because we don’t know where to look for the gap? Is the gap there, but unnoticed? What’s the difference between an unnoticed gap and a filled in gap?

Can you describe what you see where the “O” was before it disappeared as you moved the page nearer or further? Do you see anything there? Perhaps there is neither a gap nor something which fills the gap, but simply complete ignorance about what’s going on in that region of space, which we don’t normally notice because when we need information about a region we fixate it. It is only in specially contrived situations where at first we see something (e.g. the “O”) and then suddenly we don’t know where it is, that the gap is drawn to our attention. But even then we don’t know whether it is a gap or not. Perhaps we don’t even know what the question means, though superficially it seems unambiguous because all the words are so familiar.

Compare looking at the left hand edge of a page of text, while the left eye is covered. Move the page far enough away for the unattended right hand edge to be clearly visible. Move it closer while you fixate the left hand margin. Do some words on the right disappear? If you move it nearer than the distance at which a separate symbol would disappear (like the “O”) are you conscious of the right hand edge of the text? Are you conscious of a gap where some words are missing?

Notice how many of these questions concern the availability or possession of information: information about your own visual field and information about what is on the page or the screen or the wall you are looking at. The discussion below will often return to this point about having and using information.

4.2 Dealing with conceptual confusions

There are different ways of dealing with these confusions, some of which will be discussed in more detail later. One is to assume that the word "consciousness" just happens to be a bit vague, like other words whose boundaries are unclear (e.g. where's the boundary between a hill-top and a hill-side, or between a hill-side and a valley?).

Another common suggestion, mentioned previously, is that the word really labels something which varies quantitatively: there are no clear boundaries because there are differences of degree, forming a continuum of cases. On that view instead of asking which things have consciousness and which don't, we should ask which have more of it and which less, as we might do with wealth, or popularity, or speed. Similarly instead of asking whether a person recovering from concussion or an anaesthetic is or is not conscious we should ask to what degree he is conscious, and instead of asking whether some perceptual item is or is not in consciousness we should ask to what degree it is in consciousness.

A variant of this would propose that consciousness is multi-dimensional, with each dimension varying in degree, e.g. degree of self-awareness and degree of awareness of the environment.

Later I'll explain why instead of being tempted either by the notion of differences of degree in a continuum or multi-dimensional continuum, or the notion of a single major discontinuity (e.g. a dichotomy between things with and things without consciousness) we should instead consider the possibility of a large collection of different sorts discontinuities, some large and some small. This is a common feature of "cluster concepts".

I'll try to show how to think of "consciousness" as a "cluster concept" related to a cluster of capabilities that may be supported by an animal's or a robot's information processing architecture. Different combinations of these capabilities define different sorts of consciousness. Thus instead of animals being conscious to different degrees, we'll suggest that they have different kinds of consciousness, involving different collections of capabilities. E.g. some nut-eating birds, but not all, can remember a large number of different locations at which they have buried nuts.

Similarly, instead of saying that things can be in our consciousness to varying degrees, or that a person can be conscious to varying degrees, we'll have to develop a much richer vocabulary for describing the differences and discontinuities in particular episodes of human consciousness. Some of this work is already being done by vision scientists examining ways in which abnormal experiences occur after various kinds of brain damage, or in various artificial experimental situations. However, it is much harder to get a good set of descriptive concepts by induction from a large collection of data than by generating them systematically from a good theory, e.g. a theory of the architecture underlying the observed facts. Compare the terminology available for describing chemical compounds before and after development of the theory of the atomic structure of molecules.

5 A mind is an information processing control system

Mind and brain are two aspects of the same thing, namely, a very powerful and versatile control system, which makes things happen and controls how they happen. Most of this activity is not external behaviour, but in some sense "internal", though not physically internal like the chemical and electrical processes in neurons, for it cannot be made visible by opening up the brain, anymore than the operations of a spelling checker will be made visible by opening up a computer.

At the very least the information processing is internal insofar as nothing externally visible is required for or implied by its occurrence: when sitting still and neither looking at or listening to anything, nor making any sound or movement, you may still have a stream of changing mental states concerned with matters totally unrelated to the current physical environment, which you never reveal in words or deeds throughout the rest of your life, either because you forget because some interruption clears the processes or because you do not wish anyone to know what you were thinking, or because the occasion to reveal the thoughts never arises. In the case of some visual experiences, e.g. watching waves breaking on a rocky shore there may be no possibility of expressing or communicating either verbally or otherwise the full contents of the experience. The brain may lack output channels of sufficient bandwidth, and the body may lack appropriate physical mechanisms to express the content of such channels even if they had existed.

I'll often use quotes in talking about such processes as "internal" as a reminder that they are not like physically internal processes. They need not have any location within the body, and they need not be detectable or measurable by physical inspection whether the body is opened up or non-invasive physical measurements are used.

The internal processes are described as part of a *machine*, albeit a virtual, non-physical, information-processing machine, because they interact causally with one another and the physical components of the body. They are not purely random, totally unrelated events: they depend on who we are, what we know, what we are interested in, and the interactions are law-like, just as the interactions within a chess computer or a word processor are. But the laws are not laws of physics or physiology.

Neither do the laws imply that all minds are alike. Both the contents of their information states and the capabilities available for using or changing them are different in different people. A normal three year old child is incapable of wondering about the authorship of a sonnet attributed to Shakespeare, or about the nature of quantum indeterminacy. An adult is equally incapable of having the thoughts and desires of the young child he once was. Thus, each individual has a collection of capabilities and constraints, shaped by many different factors, including genetic history, individual development, cultural influences, and even types of self-modification discussed later. Within information processing machines there is tremendous scope for individual differences.

There is also scope for considerable change over time within an individual, both through self-monitoring, self-evaluation, and self-modification, if the architecture is rich enough to support such processes. But despite all this flexibility such systems are constrained to some extent: if they they change too suddenly or in very unusual ways, we may wonder whether something has gone wrong with the person, and often, tragically in the case of strokes, it has.

Many theorists offer definitions of consciousness which are short enough to be expressed in a few lines. What I am talking about is far too complex for that. It presupposes that there is something about us, some sort of evolving, developing, abstract machine, intimately connected with (and largely implemented in) the physical machine we call the brain, yet very different from it. Its operations both include the phenomena we are conscious of, and also, without our awareness, support and control what information we have conscious access to and which capabilities for processing it are turned on and off.

Exactly what sort of machine this is, and what sort of control the machine has over its own processes and long term development are still unclear. Some of the control mechanisms correspond to conscious decisions, or voluntary actions including internal actions, like starting to think about something different. Others are definitely not voluntary: an idea simply strikes us, our thoughts wander, and to a large extent, when awake, our mental processes are driven by incoming data, including what happens when you read this text. Similar things could happen in a suitably designed robot, even if the details of the processes are different because the sensors, motors, and bodily feedback are different.

In the robot, as in us, there will be many processes that are "owned" by it even though they are not voluntary. If we required reasons and decisions before anything happened, nothing would ever happen. We are able to choose between two culinary offerings even when we have no preference, and want both. Although we are sometimes surprised by our thoughts, and even our decisions, for instance, finding oneself rushing into grave danger to help a child in a burning house, we are normally willing to accept them as ours (except in pathological cases, where patients attribute the thoughts to others who are controlling them).

It is clear to many people that the sort of machine that makes us what we are is essentially some kind of information processing machine, involved in the short term and long term processes of production and modification of thoughts, desires, feelings, moods, emotions, attitudes, preferences, goals, memories, personality and other "internal" states and processes, as well as interpreting sensory data and controlling external actions.

Others, however, object to this hypothesis for various reasons, some of which are based on argument, some based on finding the idea objectionable. Sometimes that is because they have too narrow a conception of an information processing machine. E.g. they may think an information processing machine must be some kind of rigidly programmed computer, or a physical mechanism, or something which cannot do anything but operate on bit patterns. In what follows I'll try to explain a broader concept of information processing machine. But before that I'll continue presenting some of the reasons for thinking of consciousness, even in its simplest purest forms, as necessarily involving information manipulation.

5.1 Being conscious of and having information about

Although having information about something is not *sufficient* for consciousness of it (since your posture control system has information about optical flow patterns of which you are not conscious), what you are conscious of is conceptually linked to what information you have immediate access to.[11] Later I'll try to explain what sort of immediate access this is. The explanation has to be part of a general account of

- what sorts of information a system has access to (e.g. about itself, about its environment, about other agents),
- how it has access to this information (e.g. via some sort of inference, or via something more like sensory perception),
- in what form it has the information (e.g. in linguistic form or pictorial form or diagrammatic form or something else), and
- whether or how it uses the information.

The information may be information about yourself, or about the environment. What having that access implies is another matter: there is no implication that any particular use is made of it, least of all any external behaviour based on it. What having "immediate access" means also needs to be made clear.

There is also no implication that having access to information requires the existence of some internal human-like entity (which might be called a "self" or an "ego"). For that would simply lead to an infinite regress: if the self were human-like it too would be conscious and would contain another self, and so on forever. So your having information about whatever you are conscious of may involve something about a part of you, perhaps part of your architecture in a sense to be explained, but that part cannot be something like another human. So we'll need to find a way of thinking about sub-human components that manipulate information, and show how the existence of an appropriate architecture composed of such things constitutes a whole human having information, and in particular cases how that amounts to being conscious of something.

Some philosophers will object that only a *complete* human can have semantic competence, i.e. be able to acquire, use, or manipulate information. But this is just linguistic legislation, often based on ignorance of current technology. If we open our minds and look at the kind of information which an office information system may have about which orders have been processed, and the information a word processor may have about the number of pages in a document, we learn that these are examples of a type of semantic competence

[11] For this reason I find Block's distinction between "phenomenal consciousness" and "access consciousness" (Block 1995) incomprehensible. It seems to be typical of distinctions grounded only in differences in verbal forms, not in a theory of the underlying architecture and the different sorts of states and processes it can support.

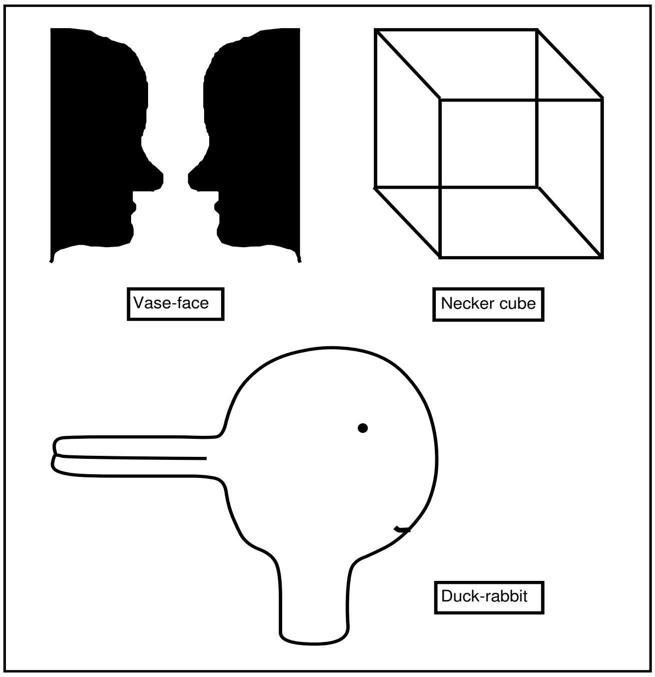

Figure 1: **Different kinds of ambiguity: geometric and nongeometric**

that requires less than a whole human being. By pondering such cases we can come to understand why an intelligent ghost, especially a conscious one, must contain an information processing machine.

5.2 Experiencing a red patch

Consider the typical philosopher's example of a paradigm case of conscious: I am now conscious of a red patch on a black background. That implies that the patch must have a spatial pattern with a particular shape, e.g. square, round, elongated, or irregular, with a sharp boundary or a fuzzy boundary, or a mixture of sharp and fuzzy bounding portions, or possibly in some peripheral areas of my visual field it may be unclear whether it is sharp or fuzzy or where the boundary is.

My consciousness of the patch and its background implies that various shape-related and space-related conceptual capabilities have been applied to the region of the patch and its surroundings, and more can be in principle. I am capable of attending to different locations in the pattern, comparing distances within the pattern, recognising and comparing shapes, colours, textures, and types of motion, for instance, even if I don't actually do so. In the case of normal adult humans many other capabilities are readily available, including imposing other known patterns on those experienced, for instance when those who are privileged to live away from light pollution gaze at the stars and see letters and other shapes, or recognising a mother's features in a young child, or seeing the patch as dog-shaped.

I have no idea how many other animals can do similar things with their spatial experiences. But if they cannot do *any* of this, then they do not have what we understand by experience of a red patch on a black background. Neither, at present (1998) do any current artificial visual robots as far as I know.

Different types of consciousness may vary not only in the capabilities they service, but also in the degree of control the person or organism has over which capabilities are applied and how they are applied. It may be that for some organisms the capabilities relevant to having visual experiences are completely controlled by the incoming data. In others, the capabilities invoked may be controlled by a mixture of incoming data and present needs or goals, or a short term memory recording recent processing results.

In humans it seems that our ability to attend at will or at whim to different aspects of our experience is relatively untrammelled: without needs, desires, or the external data changing at all we can switch between attending to different things. You can voluntarily switch your attention between an object's shape, its colour, its axis of symmetry, its top left corner, its top right corner, etc. Not all of these involve physical redirection of gaze. In the case of an ambiguous figure, such as the necker cube in Figure 1, the ability to see it one way and the ability to see it another way may sometimes be voluntarily switched, sometimes involuntarily. Typically, learning to be an artist involves learning to use these capabilities in a more systematic way and possibly also developing new ones.

Some people can develop skills which enable them to have simultaneously two views that are normally thought of as incompatible. E.g. the famous Rubens vase-face figure can be seen as two faces gazing at each other or a vase in the space between, and it is common for books on vision to say that only one or other of these can occur. However many people can easily see two faces with a vase wedged between them, as soon as they have been asked to try. They have a capability which, in most people, is never invoked unless the suggestion comes from outside.

Other ambiguous figures, such as the necker cube, may strongly resist such unusual co-application of normally mutually inhibitory capabilities, but with practice some people can simultaneously see at least portions of the normally incompatible interpretations.

All of this helps to support, but does not prove, that what is experienced is the result of application of some set of abilities, and potentially involved in every experience there will be lurking abilities ready to be applied.

Some of the potentiality can be invoked voluntarily, perhaps after special training. However, some aspects of what we see cannot be turned on and off at will. E.g. in the case of searching unsuccessfully for all occurrences of the letter "F" and for the mistake in the phrase with two occurrences of "THE", the voluntary decision to turn on all relevant reading capabilities failed at least for a while.

Moreover, you probably cannot voluntarily turn off your recognition of the phrase "The Cat" in this sentence so that it becomes undetectable in the same way as occurrences of the letter "F" were undetectable in the example above. If someone nearby mentions your name audibly you will probably hear it and be distracted, even if you are trying hard to concentrate on a conversation about an important unsolved problem and ignore everything else. You are hereby invited to banish completely the concept "elephant" from your thoughts for the next 30 seconds. Try it. Or try going back to seeing the textual patterns in front of you as merely meaningless patterns, like a child, or animal, who has not learnt to read. Normal people cannot turn off the powerful recognition and interpretation capabilities they have developed over many years and use every day, even though once it was an effort to turn them on.

It is clear that even if the mind is an information processing control system there are many different sorts of information and many different sorts of control at work. What they are, and how they interact will need to be explained in terms of the underlying information processing architecture, most of whose contents and activities are not consciously accessible.

Some readers will object that I am making too much use of introspection in all these examples, and that introspection is totally discredited as a method of discovering any objectively significant information. Such readers are invited to join me on a journey towards a view of introspection as deploying a biologically significant "inwardly focused" perceptual capability which is part of an information processing architecture that evolved to give organisms some information about themselves. Robots with similar capabilities will also be able to use introspection to find out things about their internal information processing.

Moreover, like any other perceptual ability, introspection abstracts from details, sometimes omits important information, and sometimes over-interprets, reporting what isn't there (otherwise known as self deception). Thus, as with all perceptual information we can start by taking it at face value, and then correcting and modifying on the basis of a good theory, just as we correct our normal perception of a table as solid, smooth, rigid and impenetrable, when atomic theory teaches us that it is mostly empty space.

From this biological design viewpoint, Using introspection is no different in principle from using ordinary perception in a laboratory experiment. Both are indispensable for obtaining certain kinds of information. Yet both are partly incomplete and partly misleading. We can learn what to trust and what not to trust when we have good theory of how everything works.

5.3 Dispositional states

A subtle point about information processing machines which may be very unfamiliar to most readers is that a huge amount of what goes on in such machines is concerned with things that might have happened but did not, and that many of the changes that occur are changes concerned with such "might haves". That is an important aspect of the kind of control I am talking about, which is different from more familiar mechanical types of control involving a rudder changing position, or a current being turned on or off, or a string being pulled.

Consider homeostatic systems, like a steam engine whose speed is controlled by an old fashioned governer where an increase in speed of rotation of an axle causes centrifugal force to cause a change in a valve which reduces the speed, and a reduction in speed has the opposite effect, so that the speed remains constant. Over a period of time, the configuration of springs, masses, levers, valves and the speed of rotation may actually remain unchanged, and yet there is a control system which would have reacted in certain ways if it had changed. Very many mechanical, electronic and physiological homeostatic systems work on the basis of such "might haves" and "would haves". They are designed, by human engineers, or by evolution, to make a whole lot of counterfactual conditionals true.

The same thing is true of computational systems, except that the variety of forms of control and types of counterfactuals is far greater, and more interesting!

It is worth reflecting on the differences between a multi-user operating system (like Unix or VMS) which allows multiple processes owned by different users to run on the same machine, but detects and prevents attempts by one user to access protected files belonging to another, and an operating system which cannot do this and is restricted to one user at a time (like Windows 3). This is an important difference in the nature of the control states and processes in machines running those operating systems, even if no user ever tries to read files belonging to another. The states which involve unactivated capabilities are dispositional states. The prevalence of such dispositions in human mental states was emphasised in (Ryle 1949).

Many systems involve these and other security mechanisms which, fortunately, are never activated. However, whether activated or not, the ability to detect and prevent such violations, and the disposition to invoke the ability under certain conditions, exists all the time in one operating system and not in the other.

Similarly, different parts of the same system may be linked to different abilities over a period of time, even though the abilities are never activated. If I am using an operating system which supports access privileges, I may have two files, one protected against being read by others and another unprotected. That dispositional difference between the two files exists even if nobody else ever attempts to read either of them. Similarly a file may change from being protected to being unprotected to being protected again, even though nobody other than its author ever tries to access it. Yet its causal powers change during those transitions, even if the differences are never manifested.

Some causal powers are second order powers. I have many files whose accessibility by others I can change, even though I don't. There are other files, owned by other people, whose access rights I cannot change. If I had changed the accessibility of file F, then that would have changed what other users could have done to

F. But I didn't and they didn't try. Nevertheless as long as the file exists that higher order ability to have its protection changed persists – unless, for some reason the disk it is on gets mounted "read only" in which case a third order ability will have been invoked which turns off the second order ability.

A typical computer operating system is a vast collection of multi-level dispositions and capabilities most of which are not realised most of the time. But when you buy the operating system you pay for them anyway, in the money that you have to part with, the amount of useful memory and disc space they occupy, and the reduction in performance which comes from all that flexibility, even when it is unused.[12] A mind is a vast collection of multi-level dispositions and capabilities most of which are not realised most of the time. But you pay for them anyway, in having large amounts of circuitry dedicated to them.

Some people also pay because of the dysfunctional interactions that occur in certain contexts, some of which have long term effects, including effects that go unnoticed until much later.

5.4 Consciousness and "might haves"

There are many ways in which two systems (including two virtual machines with partly different architectures) can differ in the kinds of dispositions and capabilities they support even though none of the dispositions or capabilities happens to be triggered, so that the actual occurrences in both systems are the same over an extended period.

A file of text on one machine may have exactly the same contents as a file on another, where the first machine has a huge collection of text processing capabilities, including searching, sorting, file transformation capabilities, file access capabilities, which the other lacks. In that case the mere existence of the first file enables the existence of a large collection of available but unactivated processes not enabled by the second file. The difference of context in which an information structure exists can make a huge difference to the control powers of that information structure, even if those powers are not used at all.

Similar comments can be made about human conscious states, including visual experiences.

Even when nothing much is actually happening in visual experience there is readiness for all sorts of happenings and that is what makes it *that* sort of experience. The apparently passive and static experiencing of a blank red portion of a surface includes *readiness* to experience and recognise a great variety of different sorts of shapes and colours that could appear and move in that surface. A blank space is a space where things *can* be or happen. Experiencing it *as a space* includes an implicit grasp of the space as a potential container for many different sorts of things, including things which don't occur but might have.

Just as different sorts of operating systems or software packages may be able to process different sorts of data, and may be able to process the same file of data in different ways, so different sorts of minds are ready for different sorts of occupants in spatial regions, and different sorts of processes involving those occupants, and have different abilities to process them, even if the abilities are not invoked.

If you read English or understand circuit diagrams you are (whether you think of this or not) ready for meaningful English words or sensible circuit diagrams in every spatial location you see, but not if you can read only Chinese and can understand only pictures of natural objects. One of the reasons large letters are used in books for young children is that they are not yet ready to see the words and letters in a small font suitable for adults. (Why not?) Likewise a fluent speaker of English has the ability to take in utterances at greater speed and in a more noisy background than a novice speaker, even though the novice is a fluent reader of English and can hear the same sentences if they are uttered slowly in a quiet place. So the fluent speaker and the novice will experience the same sounds differently when they hear English spoken quickly or in a noisy room. However, the capabilities in the novice will gradually change over time, with practice.

[12] For more discussion of interactions between possibilities at many levels see (Sloman 1996a).

Our consciousness of current sensory contents therefore involves readiness to apply a range of capabilities to those contents, only a small subset of which will actually be invoked at any moment. Since we are generally ignorant about the *full* range of capabilities we can deploy, it follows that we are unaware of most of what constitutes our current consciousness. This may seem paradoxical if, as recently suggested in (Baars 1997) you think of consciousness on the model of an internal theatre stage which you, or something inside you, can observe. This paper will gradually build up a different model, which has some features in common with Baars', but removes any sense of paradox.

5.5 Word processors and graphics packages *vs* cameras

One way of summarising the points made in the previous section is to say that to a first, very rough, approximation we can compare the mind of a person gazing at a blank region of space to something like an empty file in a word processor, rather than like a blank film in a camera, or an empty stage in a theatre.

If two word processors use characters in quite different alphabets, e.g. Cyrillic and Roman, they are ready for totally different contents to be inserted, with quite different rules for breaking lines, checking spelling, adjusting layout to line up margins, etc. By contrast, a camera made in Russia and one made in Italy will not differ in the kinds of text they can photograph. They do not have different sets of abilities to take in, analyse, manipulate, and use information.

If we replace the theatre analogy with an unscripted puppet show: then two puppet shows may have exactly the same things on stage during a certain period of time, but the teams of performers controlling them are able to detect different patterns and would produce different reactions to those patterns, thereby causing the rest of the show to go off in different directions. The teams of performers managing the puppets are a bit like the collections of capabilities in a word processor. But the word processor has far more variety and flexibility in its dormant dispositions, since puppet manipulators are severely constrained by physical linkages.

My wife, who is a keen orienteer, uses an excellent software package called "OCAD" to produce very detailed maps of terrain to be used for orienteering competitions. The package has an orienteering ontology built into it at a deep level: it knows (to some extent) about various kinds of depictions of objects such as buildings, boundaries, rivers, roads, and contour lines, as well as labels, scales, and orientations, and can handle them, though it cannot plan a route nor does it detect the absurdity in two contours crossing each other, whereas a more sophisticated package might.

By contrast there are other graphical packages which have information only about lines, polygons, curves and coloured regions, knowing nothing about physical geography, and many word processors handle only configurations of text items, knowing nothing of circles and polygons.

The different word processors and the different graphical packages have different architectures. Each handles a particular ontology (set of types of entities) and each is composed of collections of manipulative and recognition capabilities of different kinds, applicable to those entities. In each package the capabilities are assembled in such a way as to produce a total functionality which suits the package to particular sets of uses, to varying degrees. The systems have different designs and they fit different niches, different collections of task requirements. Or, in other words, they instantiate different areas of design-space and niche-space.

5.6 Internal *vs* external functionality

Some of the capabilities and requirements of word processors and graphical packages relate to *internal* manipulations of information structures, e.g. checking a document for spelling errors, or creation of a list of index entries. Others relate to *external* behaviour e.g. displaying things on a screen, sending instructions to a printer, responding to keyboard events or mouse events.

In general it is the abstract internal functionality which provides the core capabilities. Different external interfaces to those capabilities may be provided for different users, or different sorts of screens, or different sorts of printers or keyboards. This notion may be relatively unfamiliar to PC users where there is much more standardisation of interfaces and less separation of core functionality from the interface. However, it is a standard feature of Unix systems, where for example, different users logged in to the same machine typically access the basic electronic mail system via totally different mail-reading and sending packages, with very different kinds of functionality.

When using windows software on a PC or watching another person's face, it's the reverse: users are generally forced to employ basically the same rather limited interface to interact with very different information processing mechanisms. Human facial expressions are normally closely connected to particular internal information processing mechanisms, though in some cases (e.g. when acting on a stage) we learn to break the connection, and become more like a Unix system than a PC! The connections can also be broken by physiological damage, e.g. damage to facial muscles, to nerves, or to parts of the brain making the connections. There may also be social or cultural differences.

The core information processing capabilities can (in some, if not all cases) be left unchanged by changes in their connections to external manifestations, if the relevant portions of the brain are left undamaged.

Of course, where the central processing depends on close monitoring of proprioceptive feedback signals produced by the external changes, the central processing may be partly changed by that feedback. Thus if you are no longer able to laugh out loud the internal processes involved in finding something funny might be different. But even if your ability to weep and show sadness are completely disabled you may still be capable of being, and feeling, desperately unhappy about the death of a close friend, just as a word processor may go on with its reformatting, or index collating, or spelling checking activities even if something goes wrong with the screen on which its output would normally be displayed.

The suggestion I am offering, in this admittedly still vague and potentially misleading initial analogy, is that what makes possible the particular experiences and other mental states in an individual is an *architecture*: a collection of interacting mechanisms actually or potentially producing various effects on one another, and also, under some conditions, producing external behaviour. What sort of architecture will be discussed later. It may or may not be like some of the virtual machines that can be implemented on computers. It may turn out to require some entirely new sort of low level medium for storing and manipulating information. In being conscious we are making use of the mechanism. But we are not conscious of doing so.

5.7 Mental mechanisms as abstract (virtual) machines

What it means to say that a mind involves an architecture with a collection of interacting mechanisms should gradually become clear in the course of this paper. For now, it should not be assumed that "mechanism" necessarily refers to something physical or physiological, nor that all the effects of such mechanisms have to be externally visible: the mechanisms constituting a mind may be as abstract as the parsers, compilers, schedulers, formatters, theorem provers, mail handlers, internet browsers, nameservers etc. to be found in software systems.

None of these is a physical mechanism, though they are all *implemented* in (some would rather say "realised as") physical mechanisms. Computer scientists often refer to them as "virtual machines" (as in "the Java virtual machine", "the Prolog virtual machine").

Some architectures are physical: they are composed of physical mechanisms, and the processes they produce are physical, like a windmill causing a wheel to turn and grind wheat. Others, like the architecture of a business organisation or a computer operating system (like Unix, or Windows95), are not physical and neither are the processes they are mainly concerned with, although they can be implemented in a variety of physical machines (usually with several intermediate layers of virtual machinery).

The most interesting and sophisticated virtual machines involve far more internal processing than interactions with the environment, and as computer memories grow larger and their processors more powerful the ratio of internal operations to transactions across external interfaces will continue to grow, though it may take a long time before the internal/external ratio approaches that of human brains.[13] (Remember that the word "internal" here is being used in a special way.)

5.8 The variety of types of physical and abstract machines

The physical technology used in computing systems is constantly changing, so as to produce smaller, faster and cheaper engines with ever larger capacities. However, at a higher level, which defines the characteristics of a machine language, only a small variety of information processing engines is used, built on digital circuits designed to manipulate bit patterns in a CPU or in a randomly accessible memory (possibly supplemented by less volatile slower memories).

At a still higher level, however, there is enormous variation in types of virtual machines performing many different types of tasks. We still don't know what the full range of types of virtual machine architectures is that can be implemented on such systems, especially when many of them are put together to form information processing networks of various kinds. The space of possible distributed information processing designs is mostly uncharted territory.

Moreover, future information processors may employ a greater variety of low level mechanisms, perhaps using chemical or other new sorts of information storage and manipulation. This may make a large difference to the variety of higher level virtual machines that can be supported.

So it is important throughout this investigation to try to abstract away from the *specific* features of computers, operating systems, and programming languages, as we now know them, in order to grasp the central issues which transcend current technology. Readers who cannot do that will misconstrue what I am writing as some sort of plug for computers, or for digital technology.

That restricted vision of information processing as digital computation is common both among many AI researchers and among many of their opponents: their disagreements, based on their common narrow viewpoint, are irrelevant here. It should be obvious that if new, more powerful, types of machines become available they will be seized upon by AI researchers. Thus, just as it would be silly to define physics in terms of the technology and mathematics available to physicists at any particular time, it would also be silly to define AI as constrained by the information processing technology available at a particular time.

Equally it is silly to constrain philosophical discussion of the possibility of information processing models of mind to the particular forms of information processing models already investigated, even though existing forms provide a rich collection of examples which can be used to stretch our thinking. They are only examples.

5.9 Most of the functionality and causal interactions are "internal"

For machines whose operations are mainly internal, the definitions of the functional roles of most of the components need not refer to external behaviour. For example, the primary task of a spelling checker in a word processor is to compare words in the text buffer against an internally stored dictionary and a collection of morphological rules, and to identify the unacceptable words. Whether the result is shown on a screen, stored for later reporting, or fed into an automatic internal spelling corrector is irrelevant to the nature of the internal processes which cooperate to produce the result, though it is not irrelevant to the functional role of the spelling checking module within the larger system.

[13] The input bandwidth in human vision is admittedly very high. But many blind humans get along very well. Helen Keller was blind and deaf yet managed to write books.

Likewise, most of the internal virtual machine components in a human mind are not concerned with external input or external output, but with interactions of other internal components, from which they receive or to which they transfer information structures. This is a point that was never understood by either psychological or philosophical behaviourists. (Ryle grappled with it in the chapter on imagination in his 1949 book, which is often misread as a behaviourist manifesto.)

If all this is correct, then many, or perhaps most, of the processes produced by mental mechanisms are not perceivable externally nor with the aid of physical measuring devices. To that extent they are similar to the complex "internal" manipulations of a software system, i.e. virtual machines.

However, it should be stressed that the word "internal" here is potentially misleading. The processes in a virtual machine are not internal to a computer in the same way as the electronic processes are. The former cannot be observed or measured by opening up the computer and examining or measuring the components. Components of a virtual machine, and their states and processes, are "internal" only insofar as they form *a subset* of the total collection of processes within the complete virtual machine architecture. Thus treating "internal" as referring to a relationship of spatial containment would involve a category error, like treating the horse-power of a car engine as being something under the bonnet of the car.

Moreover, it should not be assumed that there has to be any one to one correspondence between components of the virtual machine and components of a physical machine. Different aspects of virtual machine functionality may be distributed in a complex way over different parts of the physical machine, and some of the "laws of composition" relevant to physical components may be broken by virtual machine components. For instance, no physical component can be a sub-part of another, yet it is commonplace in computational virtual machines for one list structure to be a component of another list structure which is also a component of the first one. Similarly two procedures can each form parts of the sub-mechanisms of the other.[14]

Of course, we cannot yet say how far the virtual machines implemented in brains are like or unlike the sorts of virtual machines that can be implemented in computers, or networks of computers: it could turn out that brains use mechanisms we have not yet dreamed about for some of their processing.

A special case of what I am talking about when I say that mental processes require interacting mechanisms in a non-physical architecture is the old and familiar Kantian point that there's no experience that does not involve concepts, which is one way of summarising my earlier discussion of the way in which the nature of your visual experience is in part constituted by which spatial information processing capabilities you are capable of applying to the contents of your current visual field. This point was expressed by Wittgenstein as "The substratum of this experience is the mastery of a technique" (Wittgenstein 1953, p208).

You cannot experience a 2-D set of lines as a cube, or a dot as a dot in a spatially extended surface, a face as a face, a red patch as red and extended in space, unless you have a mostly unconscious collection of abilities (techniques for producing effects in an abstract machine) some of which are involved in having the experience (e.g. applying the concepts which define the experience) and others which are ready to be deployed if triggered by a change in the sensory contents or a shift in attention or your goals (e.g. looking for symmetry in the cube picture). Some of the capabilities, if triggered, can generate waves of influence throughout the whole system, for instance the potential to be alarmed when peering nervously into a dark doorway.

My constant allusions to these internal capabilities which are ready to be deployed if needed are closely related to Ryle's notion that most of what constitutes the mental is dispositional. In other words, what is currently going on in you is to a very large extent a matter what "what would happen, or could happen, if ...". Logicians and philosophers who don't like counterfactual conditionals will not be able to make sense of this

[14] This undermines some common philosophical assumptions about requirements for supervenience of minds on brains. The issue is discussed in a draft paper accessible at **http://www.cs.bham.ac.uk/~axs/misc/supervenience**

paper. But equally they will not be able to make sense of much of engineering design which, increasingly, involves assembling mechanisms which support a wide range of counterfactual conditionals.

For instance, if you are lucky you can buy a relatively cheap computer which runs for about six years without breaking down, like the old Sun workstation on my desk, or, since you may not wish to bank on such luck if your business or your life depends on it, you can buy a very expensive computer with a lot of built in reliability features, which also runs for about six years without breaking down. What you have paid for in the second case was the truth of a large collection of counterfactual conditionals: what would have happened if a disk drive had failed, or memory components had failed, or the power supply had failed, or an algorithm had had a bug, etc.

Of course, we know what sorts of mechanisms support those dispositional properties in the computer, whereas we don't yet know what sorts of mechanisms underly all the unrealised capabilities that constitute our conscious experience. I'll return to that later, when discussing the meta-management layer in an architecture for human-like minds.

To understand all this better we need a theory of the kinds of architectures and component mechanisms that make various kinds of experiences possible, including all the supporting counterfactual conditionals.

People who reject this because of a firm belief that having an experience is an unanalysable inexplicable "given", will not be convinced by anything in this paper: their view probably cannot be changed by argument, though the theory sketched below can be used to explain why, if it accurately depicts the human mental architecture, that architecture will generate beliefs in the inexplicability of experience. The theory leads to the prediction that some intelligent robots will one day also share that view of consciousness.

Readers with more open minds are invited to join in the long term exploration of the consequences of the key idea that having a mind, or even having any single experience, involves having many coexisting states and processes, embedded in a variety of abstract mechanisms, many not yet active but ready to be triggered so as to change the experience or add new experiences, via a large number of *potential* interactions.

We are not aware of all the active and potentially active capabilities, the states and processes they produce or the interactions between them, though we may be aware of a subset. Because the interactions between these states are primarily concerned with production, maintenance or modification of states and processes I call them "control states". A mind is a control system within which things happen (Sloman 1993). The concepts of "control", "causation" and "what would happen if" are very closely related, and also very hard to analyse. I shall mostly use them without analysis in this paper, since I believe that we have to use them anyway: they inevitably pervade all our thinking both in science and in everyday life, and certainly many types of hardware and software engineering.

5.10 Information and "about"

As we have seen, the processes and capabilities constituting the mental virtual machine are not directly concerned with production of external behaviour. Rather they all involve acquisition, storage, manipulation, transformation, interpretation, retrieval or use of information *about* something (including information about one's current visual field or other mental states). I therefore call a mind "an information processing control system".

Here the word "information" inherently presupposes the notion of semantics, i.e. reference to something. Some people use the phrase "information processing" in a different way based on Shannon's information theory. The two should not be confused (Sloman 1994).

Semantic content pervades all our experiences. Even the "raw" visual experience of a plain red patch involves information about the colour of the patch, the size of the patch, the absence of any other object in the patch (if it is an "empty" patch), and the shape of the patch: even if it is experienced as having an indeterminate shape, that is still a spatial categorisation involving the application of spatial know-how. When an insect or a rabbit is faced with a red surface it may have totally different information states, for which we lack any suitable vocabulary at present.

We can begin to get some first draft (though possibly incomplete and inaccurate) ideas about the sort of architecture required for a human-like mind by reflecting on various aspects of experience, e.g. recognising a shape, noticing a spatial or causal relationship, seeing a necker cube flip, disliking a colour combination, seeing happiness or fear in a face, feeling puzzled about a movement, understanding a printed phrase, and so on. These all involve processes produced or modified (usually unconsciously) by various more or less enduring but not necessarily accessible features of our minds, i.e. our concepts, attitudes, preferences, beliefs, linguistic skills, intentions, personalities, etc. These in turn will have to be explained with reference to the types of information processing control architectures which can support them.

Ordinary language provides a very rich vocabulary of words and phrases for describing such states and processes, though the theory sketched here implies that as we discover more about the underlying architecture we'll find many ways of refining and extending that vocabulary.

The processes that can be supported by a mental architecture include episodes we would normally describe as: having new sensory experiences, learning things, taking decisions, becoming more (or less) unhappy or angry or envious or relieved, making plans, considering options, comparing things, making inferences, rehearsing arguments, coming to notice objects or processes or relationships, classifying or categorising things, feeling puzzled, forgetting things, reminiscing, switching attention from one thing to another, forming attachments, acquiring new tastes, having a new impulse to act or think in a certain way. Any theory of consciousness that does not provide a unified framework for explaining *all* of these things is clearly inferior to one that does.

However, we'll see later that there are also simpler, more primitive, architectures, such as we may expect to find in other animals, which support different ranges of mental episodes, and perhaps different kinds of consciousness. A general theory must not account only for human consciousness, but also the kind of sensory awareness that enables a fly to avoid your fly swatter. It should also account for differences between human beings at different ages, in different cultures, and affected by drugs and various kinds of brain damage or disease. What needs to be explained by adequate theory, therefore, goes far beyond what we are likely to think of when normal healthy, but untravelled, adults reflect on consciousness in their armchairs.

Of course, as in all scientific advances, a new explanatory theory can provide new concepts that extend and refine our grasp of what needs to be explained.

6 Ordinary and philosophical concepts of experience

The starting point for all this is not a philosophical or scientific theory, but common knowledge about what it is like to be a normal human being. That knowledge can be refined and extended both by factual information gleaned from laboratory and clinical reports, including observation of the effects of brain disease or damage, and also by the implications of an explanatory theory.

But our starting point, the origin of the philosophical notion of "qualia" is the familiar fact that in addition to paying attention to the rectangular table top out there or the circular penny lying on it we can also pay attention to the non-rectangular and non-circular shapes in our visual fields that are part of the same perceptual state, when we view objects obliquely. This is the sort of thing artists have to do, and a good artist learns to do it better than most people. Attending to our sensory states and noticing their detailed properties and relationships rather than the objects in the environment and their properties and relationships is not in general something that happens spontaneously for everyone.

Sometimes focusing on the structure of your own experience may provide a good way to instruct another person where to look for something: "Look just above and to the left of the point at which that hillside intersects the wall of the house". Of course the hillside does not intersect the wall. But there may be an intersection between two edge representations in one of your intermediate visual databases.

It can also be useful in explaining to someone what sort of experience to expect in a new situation, or in diagnosing visual defects, as in "Vertical lines look more blurred to me than horizontal ones".

Some of these uses of our ability to attend to our qualia may be part of the answer to how these capabilities evolved.

6.1 Qualia are not causally disconnected

Although it is hardly controversial that we have experiences and that we can attend to them, there is a more controversial *philosophical* claim that they are in some way "causally disconnected" e.g. because they cannot be explained causally, or cannot have any effects or functional role. This is not an agreed fact about common experience. It is not part of my experience, for my qualia are clearly causally connected both with other mental events and processes and my actions.

The idea of causal disconnection, far from being an obvious requirement for qualia, is simply additional philosophical baggage added by a subset of philosophers and some scientists. There is certainly no implication in the untutored notion of what it is like to have an experience that having the experience doesn't interact with anything else.

On the contrary, experiences can produce enjoyment, displeasure or boredom, which may or may not affect your utterances and other actions, and they can remind you of other experiences, or give you ideas for future actions, make you want to photograph or paint the scene, etc. Moreover even the mere fact of having the experience involves the actual or potential connection with other processes such as applying concepts to categorise the experience, recognising features of the experience, and a more subtle array of potential interactions to be described later. These are all causal, functional, relationships in an information processing system. So I conclude that whatever philosophers may claim the *ordinary* concept of experience has nothing to do with causal disconnection: that's just a theorists' invention.

When the causal disconnection requirement is added to the familiar concept of sensory experience, then it is not at all clear that we are left with a concept that is coherent (Dennett 1991). Even if the result is a logically consistent concept there is no reason why the rest of us should accept that requirement as part of the definition of any aspect of consciousness (e.g. qualia), for there may be nothing that satisfies that definition.

Adding the extra "causal disconnection" requirement would be like specifying as a requirement for simultaneity that the concept should be applicable to spatially separate events independently of any reference frame, a possibility ruled out by the special theory of relativity. If someone claims that that is how *his* concept of simultaneity works, the rest of us can just smile and attend to more important matters.

Likewise if some people wish to cripple the concept of qualia by attaching extra definitional baggage we can simply ignore them and get on with the search for a general framework that accounts for all the less controversial aspects of the concept.

The people who then reiterate that given *their* construal of qualia, qualia cannot be explained by any cognitive or physiological or physical mechanisms will simply be drawing attention to a tautology, though many seem to mistake this for a profound metaphysical truth. Likewise if I define a new type of plant growth which, by definition, has no causal connections then I shall not have achieved much by using that definition to prove that there is a type of plant growth that biologists cannot explain (even if highly trained philosophers can use their conceptual skill to imagine its possibility). I believe this line of argument refutes the main thesis of (Chalmers 1996).

Let's instead (like Ryle) go back to the philosophically uncluttered, everyday, concepts of kinds of experiences, thoughts, decisions, motives, emotions, imaginings etc., and see how far those strongly causally connected phenomena, rich in dispositional properties related to "internal" information processing, can be incorporated into a larger picture including powerful explanatory mechanisms that are not part of common sense.

In other words, let us search for types of architectures that can account not only for sensory qualia (unencumbered with unreasonable constraints) but also a host of other familiar and unfamiliar types of normal adult human mental phenomena.

If we can show how this is just one sort of architecture in a larger space of types of explanatory architectures, some of which fit other animals, some the minds of infants, and some the minds of humans with abnormal or damaged brains, then we shall have achieved something that is potentially not only deep but also very useful. We may even derive new ideas relevant to the design of more or less human-like artificial agents.

6.2 Some requirements for an architecture

Mental phenomena do not simply occur at random: conscious states and processes in normal humans are not a disorganised unintelligible mish-mash. There are patterns and principles, with relationships of varying depth and precision between the mental occurrences within an individual.

For example given a 2-D visual experience there may be a well defined set of possible 3-D interpretations each person can impose, even if the sets are different for different individuals. If you stare at the 2-D pattern of lines underlying the necker cube in Figure 1 you can see it flip between at least two different 3-D structures, but you cannot see it as a duck, or a rabbit. Likewise staring at the ambiguous duck/rabbit picture you may find that it "flips" between a duck facing one way and a rabbit facing the other way (without any change in experienced 2-D structure), but never becomes a cube. The linkages between experiences and facets of experiences are causally constrained.

Characterising that sort coherence in detail, for instance by specifying the particular collections of mechanisms and stored information that make possible those mental occurrences within each individual and how they develop, would help to explain the enormous amount of individual variation among humans.

It would also explain in more detail the sense in which a mind is a *control* system as opposed to merely being a large collection of interacting components, like the earth's weather systems. (Like minds, control systems do not have simple identity criteria.) A deep explanatory study of mind includes unpicking all these patterns and relationships and looking for underlying mechanisms that make various kinds of control systems, more or less like human minds, possible.

This will help us understand what difference it would make if the architectures and mechanisms were different, e.g. if the same sorts of mechanisms were put together in different ways, or if different mechanisms were included, or some were left out (Compare (Dennett 1996) for a partly similar view). In that way we can begin to understand other sorts of minds, and eventually see how all of these fit into a larger space of possible behaving systems. We may even have a framework showing various ways in which the architecture can go wrong, either because it does not develop properly or because something gets damaged after it develops. This could help us generate a taxonomy of possible "disorders of consciousness" within which we would hope to explain many known pathologies, and perhaps find predictions regarding types not yet found.

At that stage we may find that the notion that some things have consciousness and some do not has to be replaced by a deeper more extensive characterisation of types of designs and the capabilities they explain, generating an ontology in which there are many types of minds, with many types of "consciousness", each precisely definable in terms of the types of capabilities involved. The exploration of types of minds is necessarily connected with explorations of design-space and niche-space.

7 Layers of implementation or supervenience

Not everyone is used to thinking about mechanisms and architectures that are non-physical, though more and more such systems are being developed in computing systems, and there are there are old and familiar examples in social and economic systems.

In a particular behaving system, such as a human being or a robot, or an economy, there may be many layers of mechanisms at different levels of abstraction in different virtual machines: quantum mechanical processes, electronic and chemical processes, neuronal processes, computational processes (perhaps!), processes that manipulate semantic information, and the sorts of mental processes whose existence is presupposed in our thoughts and conversations about ourselves, our friends and our neighbours.

We know from other fields of study that complex systems with different levels of implementation are possible, where emergent states and processes occur at various levels of abstraction.[15] We know this in some cases because we have designed and implemented such systems and we know how they work, although I do not believe philosophers have yet paid enough attention to the task of analysing the relations between ontological levels in such systems, and the various kinds of causality they support, though software engineers already have considerable knowledge about this. Examples of fairly well understood man-made systems are states and processes in a word processor, or operating system, or office management system or game playing computer, or the internet.

Naturally occurring multi-ontology systems such as social systems and individual human minds are much less well understood and we are still struggling to come up with good sets of concepts that might be used to formulate theories to describe and explain them. Current theories may have to be replaced not because they are false, but because they use inadequate conceptual tools, and therefore cannot be true or false.

In all these cases we can say that the more abstract system is "implemented" in the "lower level" system. Philosophers are more accustomed to saying that the more abstract system "supervenes" on the more concrete. Simple cases of supervenience are relatively easy to understand. E.g. a mechanical clock's time-telling ability supervenes on its physical structure of cogs and wheels and springs and levers. However even here it is not so simple: for what makes its time-telling *correct* for its current location is not merely its internal structure and processes but also its external relations. When it moves from London to New York it has to be adjusted internally also. There are similar exceptions to the supervenience of mental states on internal architectures: sometimes causal connections with the environment are also relevant to the state, e.g. when the mental state includes a reference to an external object, such as the Eiffel tower.

Though clocks are relatively easy to understand, other cases of supervenience (implementation) are far more subtle and complex and we still lack a good general vocabulary for describing them.[16]

In particular it is important not to confuse the notion that one working system is "implemented" in another with the notion of an algorithm being "instantiated" in some structure, a notion used by Searle in his attacks on AI (Searle 1980). Instantiation, in its most general sense, does not require any temporal embedding or causal relationships, whereas we are here talking about functioning mechanisms *which make things happen* being implemented, e.g. in physical machines which support the abstract capabilities. (This distinction is discussed more fully in (Sloman 1992; 1996b).)

If an abstract machine X is implemented in a physical machine Y it does not follow that there is any

[15] "Emergent" is defined in (Sloman 1994).

[16] A draft paper discussing this issue is available (Sloman 1998in preparation).

simple relationship between components of the machines or that comparable laws relate them. For instance in a computing system X might include an infinite list or array even though Y is finite, and X may contain two parts A and B each of which is a component of the other, even though it is impossible for two physical components to be parts of each other. There is not even any requirement for regular correlations between events in X and events in Y, since the details of the mapping from X to Y may be constantly changing, as happens in computers with virtual memory systems or programs that use garbage collectors that relocate data structures within the physical memory. It is not clear whether brains use similar techniques.

Not all abstract machines are computational in the sense of using symbolically defined internally stored algorithms to operate on discrete structures in accordance with well defined rules, though many are. (Some of the interesting ones don't run a single algorithm, but have many concurrent interacting processes, as discussed in (Sloman 1992).) It is already clear that existing computers can support some capabilities previously found only in humans, e.g. the ability to play chess, make plans, interpret diagrams, or parse sentences, but it is not yet clear to what extent they are capable of supporting a full range of human-like mental capabilities, including enjoying philosophy, feeling sad, or experiencing a red patch as we do. However, as predicted in (Sloman 1978), it is clear that designing and using computers and software systems as a basis for designing other things has extended our thinking tools, including our ability to think about systems that are not always regarded as computational, e.g. neural nets.

Our concepts for formulating theories about various kinds of information processing control systems have developed rapidly during the last half century, partly, though not entirely, through the development of computer science and software engineering, but also through developments in theoretical biology, physics, and mathematics. We must be ready to explore and extend the explanatory potential of these new concepts, instead of assuming that we can formulate all the important questions and all the correct answers in the old concepts known to philosophers hundreds of years ago.

In particular, in the design of such things as office automation systems, plant control systems, computer operating systems and the internet we have learnt how to make machines that acquire, transform, store and use information in all sorts of ways, including information about their own information processing activities. For example the internet depends on machines having information about where to transmit email messages, and where to collect information for web browsers. Some operating systems monitor the amount of time they spend on various tasks as part of the process of load-balancing. Electronic mail systems need to keep track of how long it is since they last attempted to send a message to a site that doesn't respond, and how many attempts they have made. We are learning more and more about the design and implementation of (abstract) machines that can grasp and manipulate semantic contents referring to both external phenomena and their own internal states and processes. Evolution, of course, discovered the power of semantic bootstrapping engines, and how to design and implement them, long before we did: we are still groping to catch up.

Some people feel uncomfortable with the idea that states of an abstract machine can have causal powers or enter into functional relationships. They forget that very many of the kinds of causation that we are interested in are exactly like that, for instance when poverty tends to increase crime, full employment tends to increase inflation, and so on. Similarly abstract states of a software system can form part of a control system, e.g. for a factory or aeroplane, and the interactions are not only real, but often very important for the functioning of the factory or aeroplane.

The rest of this paper sketches the main outlines of a mixture of sketchy explanatory theory and research methodology on the basis of which we can try to explore the consequences of these and other ideas through a combination of theoretical and practical design work, empirical investigation, and conceptual analysis.

This is primarily an introduction to a new type of research programme,[17] not an argument nor a report of

[17] It is not totally new. See, for example, (Boden 1977; Dennett 1996; Hofstadter 1979; McCarthy 1990; Minsky 1987; Ryle 1949;

results.

8 The importance of architecture

This research programme investigates the implications of assuming that many familiar concepts, such as "consciousness", "pain", "experience", "emotion", "personality" can be clarified via study of possible designs for new self-interpreting sorts of information processing systems rich enough to support human mental states and processes. One of the presuppositions is that our normal concepts implicitly refer to things of which we are not aware. This is nothing new: many previous conceptual developments, especially in mathematics, included making explicit features of familiar concepts that were previously implicit, often thereby revealing that we were wrong about our own concepts.

Understanding the architecture of a system that is capable of explaining some of the more obvious and familiar features of mentality will help us grasp many other not so obvious features of our own concepts. In particular we can use the architecture as a basis for generating and classifying concepts describing possible kinds of mental states, just as theories about the architecture of matter led to a new set of concepts for classifying kinds of stuff e.g. filling in and explaining the periodic table, and a new set of concepts of kinds of processes that can involve such stuff, such as the production and destruction of various sorts of complex molecules.

As is often the case in science the proof of the pudding will be in the eating not in justifying the recipe: only after developing the consequences in great detail can we begin to understand the nature of a theory and grasp its explanatory power. At that point we shall be able to find where the consequences are wrong, so that the theory needs to be modified or rejected.

We can already say a few general things about the implications of this approach. For instance, in any system, no matter how sophisticated, self-monitoring will always be limited by the available access mechanisms and the information structures used to record the results. The only alternative to limited self-monitoring is an infinite explosion of monitoring of monitoring of monitoring ... A corollary of limited self monitoring is that whatever an agent believes about itself on the basis only of introspection is likely to be incomplete or possibly even wrong. It will not *all* be wrong: perceptual systems designed by evolution or engineers will not survive if they are completely unreliable, and the same goes for internal perceptual systems. But incomplete, partly accurate internal and external sensors may suffice for a particular niche: an elephant manages most of the time without seeing the microbes on the leaves it eats.

Another implication of the theory is that there can be no "direct" proof or refutation of the claims made here, about the nature of our minds, since the theory implies that we don't have direct access to the nature of our own minds. The self-monitoring (and therefore self-consciousness) that is naturally available will be designed to serve local biological functions, not to answer global scientific or philosophical questions about the nature of mind or the relation between mind and matter.

Thus progress can only be roundabout, including using ideas about the architecture of mind to generate a taxonomy of types of states and processes that the architecture can support and then checking those ideas empirically and through design studies. This will be problematic only for naive empiricists who believe that all knowledge must be systematically constructed from experience.

Much of the work has not yet been done. We have yet to develop an architecture-based "periodic table" for types of mind, or even for types of states that can occur in one interesting kind of mind. Nor do we know in any detail which lower level physical architectures can implement the "mental" architectures supporting human states, so there is still much to be done, though a very useful collection of ideas can be found, for

Simon 1967; Sloman 1978; Sloman and Croucher 1981; Sloman 1984; 1993; 1999).

example, in (Dennett 1996).

In particular, within the framework of a theory that allows many possible architectures fitting more or less well into many different sorts of niches (or sets of requirements), we may expect that there's not just *one* set of concepts for describing mental phenomena, but a host of different sets of mental concepts, appropriate to different sorts of architectures, such as we may find in different animals, different types of humans (including newborn infants, normal adults, people with brain damage) and possibly different future robots.

Exploring that space of possible architectures and the sets of states and processes they can support includes identifying architectures capable of explaining both familiar widely observed and more esoteric human capabilities, as well as various kinds of animal competences. Such architectures can provide us with a new set of concepts for thinking about mental states and processes in normal and abnormal humans, in other animals and in machines. We may then be able to refine muddled pre-scientific notions of "consciousness", "qualia", etc. into new precise concepts which are both theory based and fitting to the phenomena, as happened with our developments of notions of kinds of stuff.

If all this is correct, we can expect to find that many of our ordinary concepts related to notions of consciousness, awareness, perception, self-awareness, attention, experience, etc. implicitly refer in not very well defined ways to complex collections of states and processes. The feeling that we already know what we are referring to may be as mistaken as the feeling that we are clear about what we mean by notorious words like "simultaneous", "continuity", "set" or "causes" even though we use these concepts all the time and have direct experience of some of their instances.

9 Discontinuities and dichotomies

When we explore design-space and niche-space we should not assume that these spaces are continuous: on the contrary, design-space is full of discontinuities, and understanding those discontinuities can give us new insights into similarities and differences between different systems, and also the possibilities for development and change within an individual, and the possibilities for evolutionary change in groups of individuals.

Such discontinuities are ignored by those who claim that possession of consciousness is just a matter of degree, and there is a continuum of cases. The fact that we cannot find one single important division among animals does not imply that there are no important discontinuities: there could be many.

When different subsets of the collections of states and processes are found in abnormal people (e.g. after brain damage) or in other animals, or machines, or infants, the question whether they are conscious, aware, perceiving, self-aware, attending, experiencing, etc. etc. will not be well defined, and then it is pointless arguing about the cases. Instead we need to understand the similarities and differences and where necessary replace our old ill defined concepts with new theory-based ones, tested by their usefulness in constructing powerful explanations.

Then we'll have a better grasp of design-space and niche-space and be able to ask how various collections of capabilities described using those concepts, might have evolved, or how they actually evolved. By contrast asking or arguing about the evolution or the biological function of some ill defined "it" identified by an introspective process of pointing is a waste of time.

The assumption that there is a well defined "it" leads directly to the belief that there is a dichotomy in nature between things which have "it" and those that don't. Many who think they know about consciousness from direct acquaintance claim that there is a binary division (a dichotomy) between things that do and do not have consciousness, or a binary division between states in which we are conscious and states when we are not, or a binary division between those things of which we are conscious at a time and those of which we are not.

That illusion tempts us to ask questions such as 'Which animals have consciousness and which don't?' How did 'it' evolve? Does 'it' have a biological function? Is 'it' reducible to physics? Could a robot have ‘it’? Could software running in a computer have ‘it’? Could there be a machine which is a “zombie”, with all the appearance and behaviour normally associated with consciousness but totally lacking ‘it’? All these (and many other) questions become unanswerable if the presumed concept of consciousness turns out to be a muddled collection of very different concepts. Or, to be more precise, they need to be replaced by a collection of different questions corresponding to those concepts. We’ll then replace a single mythical dichotomy with a host of important discontinuities that required detailed investigation.

The enormous diversity among living things (e.g. sunflowers, carnivorous plants, amoebas, rats, bonobos) reveals no obvious place to draw a single boundary. Similarly when a foetus develops, it is at first simply a cell and then it starts dividing. Eventually it pops out and yells, and after a few more years gives lectures on philosophy. Is there a time at which it switches from something without consciousness to something that not only has ‘it’, but talks about ‘it’?

Later the individual may have a degenerative brain disease, suffering slow degradation until what is left is just a “vegetable” (or rather a piece of meat). At what point does consciousness cease?

9.1 Don’t be tempted to fall back on continuity

Many people who are aware of these difficulties in specifying boundaries agree that there is no dichotomy, but then go on to propose a continuum: claiming there is no sharp division between things with and without consciousness, or between what you are and are not conscious of, because it is all a matter of degree, so that all we can say is that animals differ quantitatively in their capabilities.

This is partly correct, because there may be gradual changes between different cases, but also seriously misleading, for two reasons:
(a) The idea of a smooth continuum is inappropriate to design-space since there are many discontinuities where no intermediate cases exist between possible designs (between possible architectures and mechanisms).
(b) The space of possible designs is not linearly ordered as implied by the phrase “differences of degree”. It has a far richer structure, as indicated below in Figure 9, discussed later.

Many people don’t realise that a continuum is not the only alternative to a single major discontinuity: there could be large number of different sorts of discontinuities. That is a better view of the variety of types of architectures. This is the basis of the exciting research programme of exploring all the myriad discontinuities, instead of seeking the mythical unique boundary line, or wallowing in the intellectually un-challenging idea that there are no divisions because it is a continuum.

One way to explain how there could be many discontinuities is to allow that consciousness may be a *cluster* concept. To say “consciousness” is a cluster concept is to say that it refers neither to something unitary that is always wholly present or wholly absent, nor to something smoothly varying in degree: rather it involves a large collection of re-combinable capabilities which can be present or absent in different combinations. These capabilities (some of which are listed below) do not differ only in quantity. They are different in their function, structure, origins, and ways of going wrong. Those are differences in kind, not in degree. (There may also be some differences of degree, e.g. speed of processing, memory capacity, or maximum depth of nesting of plans and sub-plans, and also some continuously variable probability distributions.)

Different clusters can occur in different organisms, in different sorts of machines, different people (we are not all exactly the same - we have different kinds of capabilities). Even in the same person at different times the collection of capabilities changes: between infancy, childhood and adulthood and in senile dementia. Brain injury and drugs can also make a difference to which capabilities are present.

When I say consciousness is a cluster concept involving a collection of properties – a, b, c, d, e, etc. – I am not saying that consciousness is some logical combination of them such as a disjunction, conjunction of disjunctions or disjunction of conjunctions or whatever.

A cluster concept can have a kind of indeterminacy as to what is and is not required in its instances, partly because our grasp of what is possible is too limited for us to have clear notions about how to divide things up, and partly because our previous history has not forced us to agree on criteria to deal with all possible cases. So cluster concepts can be indeterminate, and consciousness is no exception. A good example of this is the clash of intuitions as to whether consciousness can be present while someone is fast asleep and dreaming.

What sorts of capabilities are implicitly involved in the consciousness cluster? A partial list is given below. Readers should be able to produce many more.

10 Some capabilities involved in having experiences

Humans have many kinds of distinct perceptual capabilities which are not always clearly separated.

Being able to recognise something as an 'a', a 'b' or a particular word, or 2-D pattern, is different from interpreting it as having a meaning, e.g. referring to or depicting something else. For instance recognising the word "cat" as one that is associated with a particular spoken sound is not the same as grasping its meaning in a sentence. Similarly recognising a 2-D picture as being, for instance, the one you saw yesterday is quite different from seeing it as depicting a 3-D object.

In general the ability to *recognise* or *classify* an object (e.g. a 2-D pattern) is different from the ability to *interpret* it or give it a semantics (e.g. as representing a 3-D structure).

Moreover, within the class of interpretative abilities there are many different sub-cases. For instance being able to interpret abstract symbols that don't have any meaningful components (e.g. using ticks and crosses to label things as right or wrong, or knowing the meanings of simple words) is different from interpreting a complex structure whose meaning comes from the meanings of the components and the structure of the whole. Examples of the latter (compositional meanings) are understanding a phrase or sentence and understanding a picture whose parts are meaningful, e.g. a picture of a cube or other 3-D shape.

Within the category of *compositional* interpretations there are differences between understanding linguistic forms and understanding pictorial forms. And among pictorial forms there are differences between the cases where the pictures are isomorphic with what they represent and those where they are not (e.g. a 2-D picture cannot be isomorphic with a 3-D object). Within the non-isomorphic cases there are distinctions between the pictures which replicate the appearance of the object in varying degrees of accuracy (e.g. photographs and "realistic" paintings) and those which do not replicate the appearance at all though they may cleverly suggest aspects of the appearance, e.g. cartoon drawings. In some cases, e.g. maps, understanding the depiction cannot depend on comparison with a view of what is depicted because we never have aerial views of most places shown on maps.

Some forms of interpretation map spatial structures onto complex abstract states and processes, for instance seeing one object as "supporting" another, which involves a causal relationship – preventing downward motion. A class of perceptual experiences which we probably share with some other animals involves seeing facial expression and posture or motion as an indication of a particular internal mental state, e.g. seeing a face as happy or sad or threatening, or a posture as submissive or aggressive. These experiences may link the current percept to some very important decision making capabilities. Whether we are aware of the fact or not the intrinsic nature of an experience involves causal functional relationships, according to the theory being sketched here. (As already indicated the idea is not original: I am merely extending ideas of Kant, Ryle and Wittgenstein.)

One reason we are unaware of these relationships implicit in our experience is that there are so many of them and they are so diverse. The nature of our experience is too complex to be experienced: if it were it would explode in an infinite regress, as explained above.

The recognition and interpretation capabilities can also be combined with the ability to experience aesthetic and other affective qualities, including grace, symmetry, elegance, and sexual attractiveness. But not everyone has the same collection of abilities, and it is not clear whether other animals have all of these. Neither is it clear whether the capability of aesthetic evaluation is inherent in all sensory experience or whether it is an optional "add-on". We need to understand the mechanisms underlying these abilities to understand what they are and how they vary.

Yet another important class of capabilities (described in more detail in (Sloman 1989)) involves being able not only to perceive structures, but also to perceive possibilities and constraints on possibilities, for instance, seeing that a window catch admits certain possible states and some of those states constrain motion of the window while others don't. I think that this is what J.J.Gibson called the detection of "affordances" (Gibson 1979). Seeing that a red surface admits the possibility of other colours and shapes in the surface is a more subtle affordance. Grasping such possibilities and constraints on possibilities seems to be an important aspect of seeing how to act in a situation, e.g. planning a path across a cluttered room, seeing how to grasp an object with an awkward shape, or seeing that a stick on the floor could be picked up and used to move an object that is currently out of reach, an ability Kohler demonstrated in some chimpanzees (though it is always dangerous to derive hard conclusions from observations of behaviour).

If instead of trying to identify the nature of experience by attending to it, we ask ourselves why different experiences matter to us, and what sorts of roles they have in our lives, we can move towards a deeper understanding of what we are, going beyond naive, untutored, introspection, which we can be sure is inadequate for reasons given previously.

All this makes it easier to accept the possibility of diversity in types of minds or perceptual experiences: something not revealed by introspection, which provides at best information about only one mind. Diversity exists not only between species, but also among humans: people do not all see the same possibilities and constraints. Even an individual's abilities to experience change with experience, or the results of brain damage. Some mental states are culturally based (e.g. feeling patriotic or sinful would be impossible in some cultures), and this can add to the diversity of types of minds and mental processes.

10.1 Remembrance of things past, and future

The contents of experience are not restricted to interpreting current sensory input. You can be aware of what happened recently or a long time ago, or what might happen.

There are different sorts of memory capabilities. For instance besides the enormous long term stores of information about our language, our environment and a host of skills learnt in a lifetime, there are limited capacity short-term memories, like the ability to store and repeat the numbers 7, 3, 5, 2, 8, 2. The ability to remember some things for longer may enable an animal to recall that it has a nest a long way from its current location, and remember how to get back there. Some birds can remember 50 or more separate locations where they have buried nuts.

Humans can also remember things concerned with the future! E.g. we often need to contemplate several possible future actions and events and move back and forth between them before deciding. Here short term memory is not concerned with remembering facts, but remembering what might be.

This deliberative capability is shared with some animals, but probably not all. For instance it seems unlikely that insects have it. (We must beware of dogmatism however: it remains an empirical question, and it could turn out that an ant colony has powers that individual ants lack.)

There are also many kinds of learning that produce short term and long term changes, in specific knowledge, concepts, strategies, perceptual skills, problem solving skills, physical skills, social know-how etc. Different combinations of these learning capabilities would define different sorts of minds. Further differentiation would come from individual histories triggering activation differently, leading to different sorts of long term learning.

The architecture of an expert tennis player and an expert sight-reader of piano scores would be significantly different, and as a result how they experience the same situation could be very different. Most animals cannot learn to become either: not all intelligent architectures have the same developmental potential.

Motivational capabilities and processes (feeling hungry, thirsty, sleepy, tired, hot, cold, sexy or wanting to solve a mathematical or philosophical problem) can also vary enormously between individuals, or between species. Some humans, though regrettably not all, have what we call moral feelings or can feel indignation about what is being done to something else. I don't know how many other animals can do that.

All of these ways in which minds can vary could be the basis for differentiating the types of mental states and processes that can occur in different sorts of minds. It should already be clear that the space of possibilities has a very rich and intricate structure, not easily captured by a few dimensions of variation.

10.2 Kinds of self consciousness

Many human capabilities involve self monitoring and self control. You can not only look at things around you and do things, but you can also attend to what you are looking at, think about what you are paying attention to, and notice aspects of your experience, such as how what you see varies as you move, without which much art would be impossible.

This requires an architecture that includes a 'meta-management' component (Beaudoin 1994; Wright *et al.* 1996) that has self-monitoring, self evaluation and self modification abilities. It's an empirical question whether other animals have this: I conjecture that very few species have, fewer than have deliberative capabilities required for planning external actions. It may be connected with the "global workspace" postulated in (Baars 1988).

Possession of this extra architectural layer has many implications. For instance, sometimes we control our thoughts, because we monitor them, evaluate them, and decide that they should change, e.g. deciding to cease dwelling on yesterday's humiliating experience and instead pay attention to an important task, or deciding that our decisions are too hasty. But sometimes this attempt at self control fails. So we also have the ability to lose control of our thought processes, at least partly. How many other animals can monitor and control thought processes, and partially lose control? Those that cannot are incapable of certain sorts of emotional states, which I call "perturbant". (Many examples are given in (Goleman 1996).)

These notions of controlling attention and losing control of attention depend on the existence of the meta-management component in the architecture mentioned above, and described further below.

The evaluation of internally monitored processes can take several forms including experiencing the state or processes as pleasant or painful, as successful or unsuccessful, as having positive or negative aesthetic qualities. We can also judge some of our own processes as ethically good or bad, e.g. noticing whether we are letting unselfish or selfish motives influence our decisions.

10.3 Towards an explanatory framework

Different sub-collections of these abilities to sense, interpret, remember, contemplate, grasp possibilities, evaluate, and control may be present in different people, in the same person at different times, in different organisms, and in different sorts of more or less intelligent machines.

Asking which of the capabilities constitute consciousness is pointless: like many other cluster concepts, the concept of "consciousness" is too indeterminate for there to be any answer. And in any case there are far more important questions concerning the study of all these different capabilities, the mechanisms underlying them, how they evolved, which ones can be implemented in computers, which require new sorts of machines, and so on.

I have listed only a tiny fragment of the vast collection of familiar capabilities that make up what it is to be a normal human being having normal human experiences. A unified theory of consciousness would not only include a more detailed and systematic survey of all this richness and diversity, but would also provide some sort of generative explanatory framework, able to account for diversity within an individual, between individuals, between species, and in possible robots. A specification for an architecture able to support these capabilities could explain the mental states and processes in an individual. A survey of the space of possible architectures including an account of the dynamics of individual architectures and (as Dennett has frequently argued e.g. (Dennett 1996)) the evolution of collections of architectures, would do it for a much larger class of types of minds. This includes understanding trajectories in design-space and niche-space, discussed below.

11 Evolution of architectural layers

It is conceivable that human minds, and the brains on which they are implemented are essentially too complex to be understood by human minds. If so the search for an intelligible unifying theory is a search for pie in the sky. However, such pessimism is premature. We have by no means exhausted all the possible forms of explanation, and the rest of this paper attempts to sketch a framework that has hardly been developed so far. Although there is no question of proving that it will work there are some indications that it will prove fruitful.

One of the key ideas is *approximate modularity*. If we try to understand a complex system as one undifferentiated system we may not be able to make much sense of it. But if we can discern different subsystems which can be understood separately to some extent, then we may be able to see how they can be assembled in a single complex system, even if their interconnections are so rich that they are not really separate systems. In particular, while it may be impossible to understand the evolution of the total collection of capabilities in the whole system, we may be able to understand the separate evolution of the subsystems, followed by the evolution of new linkages between them which reduced their independence and changed their functionality.

The conjecture that I wish to explore (on the basis of considerable interdisciplinary collaboration) is that we can understand the characteristically human clusters of capabilities outlined above by seeing how they arise out of an architecture composed of layers with different evolutionary histories, for instance a reactive layer, a deliberative layer and a meta-management (or self-monitoring) layer. In humans the different layers seem to coexist and operate in parallel though sometimes one layer may be dominated by other layers. A reactive layer can involve a high degree of internal parallelism because of the use of dedicated circuitry, whilst the other two, though implemented in parallel mechanisms, may be inherently serial in the tasks they perform, as discussed below.

Some of these layers may occur in different combinations in other animals, without the full panoply of human mental functioning, and it may help if we start by thinking about how they might function in simpler systems.

11.1 Reactive architectures

For example one kind of agent (Figure 2) involves a purely *reactive* control mechanism. Information is acquired through external sensors and internal monitors and propagates through and around the system, and out to effectors of various kinds. Everything happens in parallel because there are dedicated coexisting circuits. They may be entirely analog circuits (as in (Braitenberg 1984)), or entirely digital circuits or some sort of mixture. If well tailored to their niche such systems can be fast, flexible and successful, and they have recently been attracting a lot of attention among some AI researchers e.g. (Brooks 1991).

It appears that many organisms (e.g. plants, insects) are entirely like that, whereas in humans the reactive mechanisms are part of a larger architecture, which, as explained in some detail in (Goleman 1996), uses

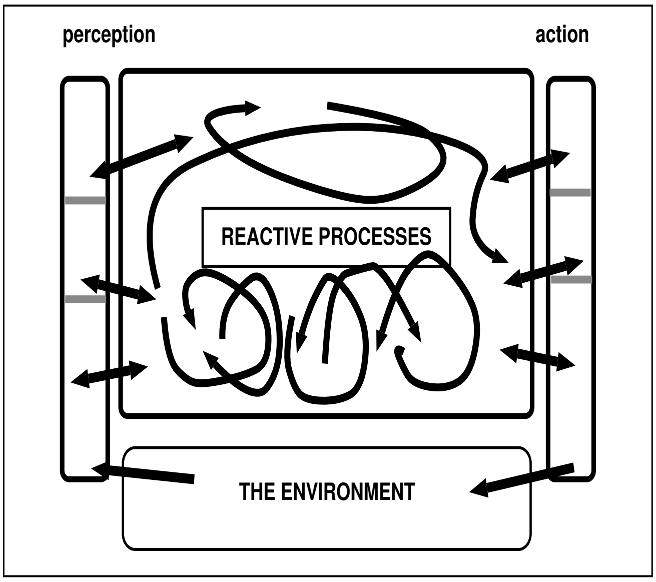

Figure 2: **An architecture for a reactive agent**

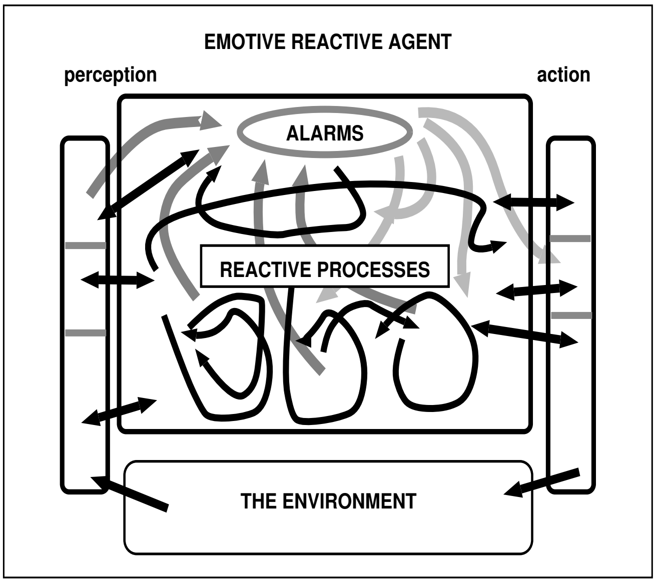

Figure 3: **A reactive agent with global "alarms"**

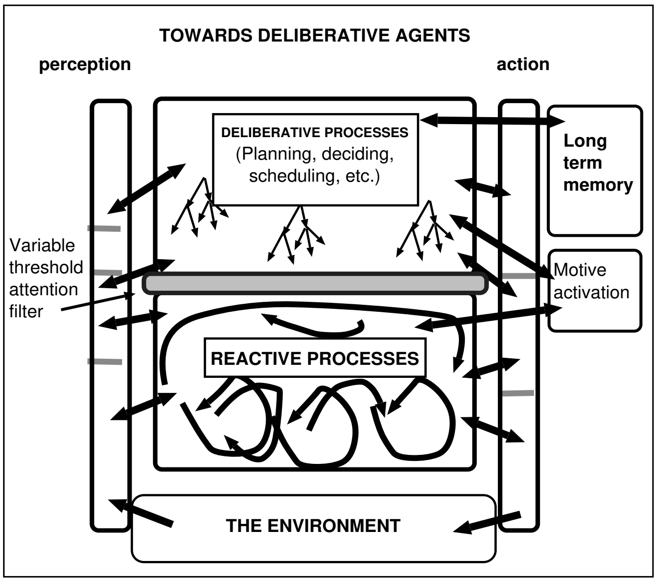

Figure 4: **A hybrid reactive and deliberative agent**

older parts of the brain also found in other animals, though subsequently modified to integrate with newer brain mechanisms. A more detailed survey than there is space for here would investigate the advantages and disadvantages of purely continuous reactive mechanisms against digital mechanisms in a variety of niches. As we'll see, some of the mechanisms required for human minds are inherently digital.

11.2 Global interrupts and alarms

Many theories of emotions postulate a system that operates in parallel with normal functions and can react to abnormal occurrences by generating some kind of interrupt which overrides everything else. Consider an insect-like organism with a purely reactive architecture, which processes sensory input and engages in a variety of routine tasks (hunting, feeding, nest building, mating, etc.). It may be useful to detect certain patterns which imply an *urgent* need to react to danger or opportunity, causing immediate freezing, or fleeing, or attacking, or protecting young, etc. Aspects of the limbic system in vertebrate brains seem to have this sort of function. We can depict the combination of reactive mechanisms with alarms in Figure 3.

11.3 Adding deliberative capabilities

Another architectural layer, a *deliberative* mechanism, is able to create new options for action in advance, evaluate them and select between them. This includes constructing a plan for a new complex action composed of smaller action steps. Creating a new plan involves having knowledge about the consequences of adding new steps in various contexts. This requires an associative memory including knowledge that can be used for selecting steps with the right sorts of consequences. The plans need not all be linear: they can include contingency branches to deal with the cases where there is insufficient prior knowledge to determine which step is appropriate in every case. Deliberative processes involve thinking about "what would happen if", but in more sophisticated systems can also be used for thinking about what might have happened in a situation that occurred previously. Deliberative and reactive architectural layers can be combined in hybrid architectures, as indicated in Figure 4.

Whereas a purely reactive architecture can make do with hardware circuits permanently dedicated to parallel activities, a deliberative mechanism needs a reusable common store of temporary memory for constructing representations of these "advance" or "hypothetical" possibilities. Moreover because the construction of exploratory plans is inherently a stepwise process (i.e. plan space, like design-space is full of discontinuities) a deliberative mechanism is largely digital and discrete, though there may be continuous modulatory mechanisms (perhaps including mood changes in humans).

For creating new plans, a long term associative memory is needed, linking types of occurrences in various contexts to their likely effects. This memory can be used used to guide the construction and evaluation of a novel plan and support the reasoning about "what would have happened if..." The existence of such a memory would in turn require additional mechanisms for extending the stored associations. These could take many different forms, which will not be discussed further here.

In a social animal it is possible to absorb plans created by others and follow them without having to go through the laborious process of construction and evaluation. This can use either a process of plan induction based on observation of the successful actions of others, or plan communication where a rich enough language is available for direct transfer of ready made plans. Information about the execution of plans by others can also allow individuals to benefit from the observed errors of other planners whose mistakes lead to disappointment or disaster.

For a variety of reasons (e.g. sketched in (Sloman 1997)), such a deliberative system, even if it is implemented in highly parallel mechanisms, would be intrinsically mostly *serial* in its functionality. The serial nature is empirically evident in the fact that although we can walk and talk and admire the view in parallel we cannot recite several poems in our heads, perform a calculation, and sing several tunes to ourselves, all in parallel, even if we can do each of them fluently on its own. (I am talking about purely internal performances.)

This implies that there is a powerful resource that has to be shared between different goals and needs. One aspect of attention is selection of tasks and subject matter for the deliberative mechanisms. This requires a mixture of top-down and bottom up control, because of the need for important tasks to drive the processes some of the time, whilst allowing new signs of important dangers or opportunities to redirect attention. This has implications for the study of emotional states where there is partial loss of control of attention, e.g. grief or excited anticipation (described in more detail in (Wright *et al.* 1996; Sloman 1997))

Some of the mechanisms for controlling the resource allocation, and directing it towards self-improvement could include filters with dynamically varying thresholds (e.g. filters between reactive and deliberative layers indicated in the figures), and mechanisms involved in what we loosely describe as reward and punishment.

A feature characteristic of the human hybrid architecture seems to include the ability of the deliberative layer to transfer plans that it has created (or learnt from others) into "reactive programs" in the reactive layer. A familiar example is learning to drive a car, which after much practice eventually becomes a semi-automatic process. It seems that similar processes of "compiling down" to reactive mechanisms occurs in many forms of athletic, artistic and intellectual learning (piano playing, reading, doing mathematics, programming etc). This requires an extendable store of reactive behaviours, which may not exist in all reactive systems.

One of the claims of this paper is that much of human consciousness is constituted by the ready availability of a host of such skills linked into every experience: Wittgenstein's substratum.

11.4 A hybrid system with global alarms

Where reactive mechanisms are combined with deliberative mechanisms, the sort of global alarm mechanism described previously can be extended to cause sudden changes also in internal behaviour, such as aborting a planning or plan execution process, switching attention to a new task, generating a new high priority goal (e.g. to escape from a predator, or to find the source of the noise just heard). Likewise processing patterns in the deliberative layer may be detected and fed into the alarm system, for instance if a planning process reveals

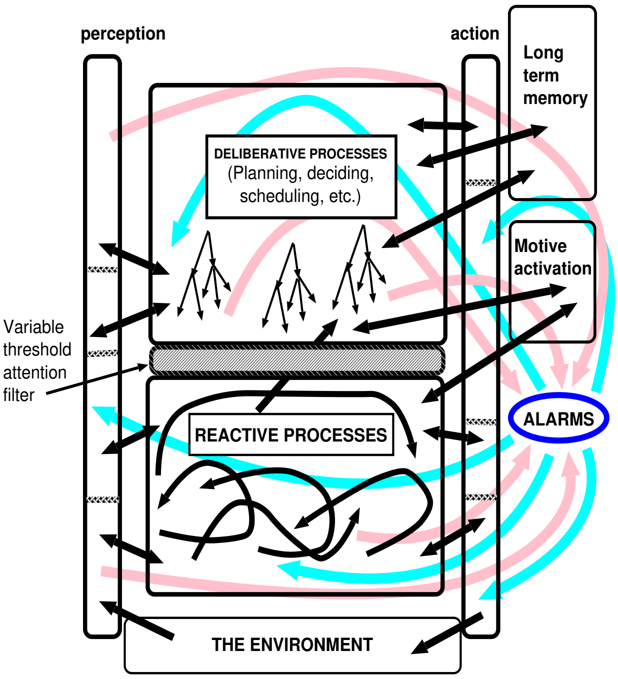

Figure 5: **A hybrid reactive and deliberative agent, with alarms**

that an important opportunity is likely to be lost unless very rapid action is taken. This is indicated in Figure 5.

The fact that deliberative mechanisms are resource limited, together with the fact that reactive mechanisms need to be able to cause interrupts of various kinds that are capable of redirecting deliberative processes to deal with urgent or important risks and opportunities, requires a strategy for managing the tensions between the two processes. (Some of the mechanisms generating such conflicts and the consequences of poor management are discussed at length in (Goleman 1996).)

With colleagues at Birmingham (Beaudoin 1994; Wright *et al.* 1996; Sloman 1997; 1999) I have been exploring the of third level of architecture mentioned above, which, in evolutionary terms, would be even more recent, and possibly a lot more rare. This involves a "meta-management" (reflective) subsystem, depicted crudely in Figure 6, which can monitor the strategies and behaviour of the deliberative system and some aspects of the reactive system (and possibly also the meta-management system itself) and take corrective action when the individual decisions do not seem to be producing an overall state that is valued highly or when they are evaluated as not conforming to some other standards, e.g. ethical, aesthetic, or efficiency criteria.

An example from human life might be noticing that one is switching attention between tasks too frequently, with consequent loss of efficiency. Corrective action might involve deciding to ignore interrupts and new motives for a time. Such control at the meta-management level is not perfect: we sometimes decide, and want, to think about X, but are continually drawn back to thinking about Y, a characteristic of certain sorts of emotional state, which probably occurs only in humans. For instance, it seems unlikely that a rat sometimes has control of its thought processes and sometimes loses control. I don't know whether a chimpanzee can decide that it would be better off thinking about something different. Only the most cognitively sophisticated animals have these abstract management mechanisms. (That's a tautology!)

These three layers in the proposed architecture, the automatic reactive processing layer, the deliberative layer which can construct and contemplate complex options in advance of acting, and the meta-management layer which can monitor, evaluate, and redirect high level internal actions, may each add new kinds of

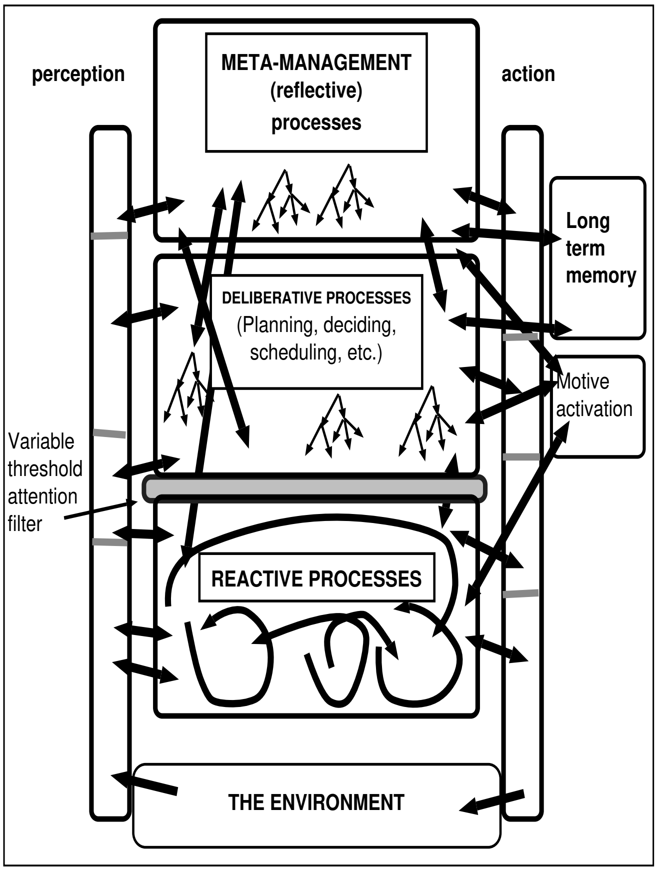
perception
META-MANAGEMENT (reflective) processes
action
Long term memory
DELIBERATIVE PROCESSES
(Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
REACTIVE PROCESSES
THE ENVIRONMENT

Figure 6: **Adding a meta-management layer**

capabilities compared with earlier more widespread systems. Trying to treat them as part of a *continuum* of types of mind is therefore seriously misleading. Moreover, for each type of layer there is a large collection of distinct capabilities, which can be present or absent in different combinations, adding to the diversity of designs, and discontinuities in design-space.

The existence of self monitoring mechanism along with a rich perceptual architecture containing several intermediate levels of processing of sensory data can account for the phenomena that make many people wish to talk about sensory "qualia" and "subjective experience", for these are aspects of perceptual states of the viewer as attended to by the viewer, as explained above. In the jargon of some philosophers they have "first-person" characteristics.

The ability to pay attention to our own experience is different from the ability to have the experience (i.e. the ability to see what is out there). Being able to attend to the experience can be useful for some purposes (drawing things, telling others how things look to us, helping someone identify a distant object in a complex scene by describing its relationship to others in the visual field, etc.) I do not know how many other animals which have sensory experiences also have the ability to attend to them, to report them, to compare them.

12 Reactive, deliberative and reflective mechanisms

This section adds a little more detail on the three architectural layers outlined above.

There are many variant forms of purely reactive agent, fitting the general ideas sketched in Figure 2. In particular, sophisticated versions require various kinds of mechanisms to resolve conflicts when different behaviours are triggered simultaneously. Sometimes it makes sense to compose the different behaviours, e.g. using something like vector addition. In other cases the conflict may be resolved by a "winner takes all" neural network, e.g. one in which the first subnet to exceed a certain activation threshold immediately suppresses all competitors.

A reactive network may have a fixed architecture or it may be modifiable by processes which create new links or kill off old ones. More subtle forms of adaptation use reinforcement learning which changes relative strengths of links between nodes. In a purely reactive system the set of behaviours may be fixed genetically with only marginal changes produced by learning. In a hybrid system if the reactive architecture includes spare capacity, it may be possible for new behaviours created by a deliberative architectural layer to be added to the reactive repertoire, as seems to happen when humans acquire new forms of expertise.

It is also possible for functional differentiation to occur in a complex reactive architecture. For instance, in addition to relatively sophisticated and fine-grained recognition of details of a situation which can control movement, a much faster, coarse-grained recognition process could trigger globally dominant reactions such as fighting, fleeing, freezing, ducking, catching rapidly moving prey, increasing arousal, etc. This seems to be how the mammalian limbic system works (Goleman 1996).

Where a reactive system detects an internal need, e.g. lack of food or water, it may be able to create a new internal state which then provides part of the context for other reactions, initiating or strengthening some while terminating or inhibiting others. In a hybrid architecture the same reactive mechanisms may be capable of generating new motives to drive processes of planning and decision making (e.g. find food, find water, get warmer, etc.). Thus a state which functions solely as the basis for controlling pre-existing behaviour patterns in a reactive system may have a quite different role in the process of creating a new behaviour pattern, i.e. a new plan, in a deliberative mechanism.

Figure 2 indicates that the perceptual and motor systems of a reactive agent may also have layered architectures. For instance, perceptual processes may perform a sequence of abstractions capable of feeding information into more sophisticated behaviours, for instance recognising a potential predator and giving information about its location and motion to an escape behaviour ((Sloman 1989)).

Although purely reactive agents are inherently less flexible than agents able to synthesise new plans, if their evolutionary history has been sufficiently long and varied to provide all the behaviours and behaviour control systems they need, and their brains can store all the information, then they can appear as flexible and intelligent as agents with deliberative mechanisms, or even more intelligent since their responses will be quicker. If termites have a purely reactive architecture, then the amazing cathedrals they construct and maintain illustrate this point.

It follows that one cannot determine purely from external observation of achievements whether a system is purely reactive or includes deliberative mechanisms. Detecting the difference requires the "design stance" (Dennett 1978). A purely reactive system, all of whose responses were completely determined by previously stored chains of condition action rules or neural circuits, designed either by a lengthy process of evolution or a super-intelligent designer, would be a sort of "zombie" indistinguishable from an agent that works out its own novel solutions to problems, compares alternatives, evaluates them and takes decisions on the basis of its long and short term objectives, preferences, etc.

It is because this difference is not definable or detectable if one adopts only the intentional stance (Dennett 1978) that that stance is inadequate as a basis for understanding mentality, and similarly the Turing test is inadequate as a basis for assessing the presence of mentality.[18]

[18] In the article (Turing 1950) describing the test, Turing himself did not make the mistake of describing the test as anything more

Of course, simply adapting the design stance does not in itself provide any simple method for investigating how a complex information processing system works. Even though opening up a modern computer might enable a digital electronic engineer to discover the main circuitry and identify some of the more important functional divisions (e.g. CPU, memory, interface devices), there is no straightforward way to find out the software architecture, algorithms, and datastructures used by the high level virtual machines running on the computer, not least because some of the techniques for probing the processes would change them.

In general only the people who designed the software can answer questions about how the system works. Moreover, if it is an adaptive, self-modifying system, even they may not know what is going on after the system has been running for a long time and has redesigned itself. But the *difficulty* in finding out what is going does not entail that it is impossible to do, or meaningless to ask, or that there is no difference between a deliberative or hybrid design and a purely reactive design.

In fact there are advantages in a hybrid system that could be useful both in an evolutionary context (faster evolution) and from the point of view of a designer concerned with physical requirements for memory stores and the need to be able to cope with the possibility of classes of problems unanticipated by the designer. These issues appear not to be understood by those who are convinced by the suggestion of Brooks (Brooks 1991) that nothing but reactive behaviours will ever be needed for intelligence.

The tree structures shown in Figure 4 are intended to indicate that within a deliberative architecture new hierarchically nested structures such as plans composed of sub-plans, or sentences composed of clauses and phrases, may be created. For a variety of reasons the process of creation is likely to be resource limited. For example the structures may be built in a re-usable workspace of limited capacity, so that it is not possible to create many of them in parallel. The process of construction requires frequent access to a long term associative store of information (to answer questions about what would happen if) and there may be only one such store capable of dealing with only one question at a time. Moreover if a system is to learn from its successes and failures (using meta-management) then it should not do too many things at once, since the resulting combinatorics could defeat the process of finding which combination of activities produced which results. There may also be advantages for integrated control if not too many high level decision making processes are allowed to run in parallel with equal powers.

For all these reasons resource limits in the deliberative layer may prevent indefinite growth of parallel activities. In that case, if the reactive layer and perceptual mechanisms are driven by a fast changing environment, feeding new information and new motives into the deliberative system in rapid succession, the resultant rapid switching of attention may be dysfunctional. That is why Figure 4 suggests that a variable threshold interrupt filter may be required. The Birmingham group's papers on emotions, inspired in part by the work of H.A.Simon (Simon 1967)) have explored some of the implications of this for the understanding of emotions such as long term grief.

We previously considered both a reactive system which evolution has designed to cope adequately with all naturally occurring situations and one which requires the additional flexibility of a deliberative layer. Likewise we can consider both a deliberative layer whose mechanisms for evaluating new goals, creating plans, comparing options, deciding what to do when, etc. have evolved so that they cope adequately in all situations and compare that with one that may need to improve itself by monitoring its performance learning which sorts of things work well and which don't, and determining conditions under which the interrupt filter threshold should be raised or lowered. This is the sort of thing the meta-management layer, depicted in Figure 6 is supposed to do. It may use mechanisms very similar to those involved in planning and deliberating about external actions, except that it needs additional mechanisms for monitoring and altering *internal* states, processes and strategies (including possibly its own). Such a system would then be involved in situations

than a technological challenge.

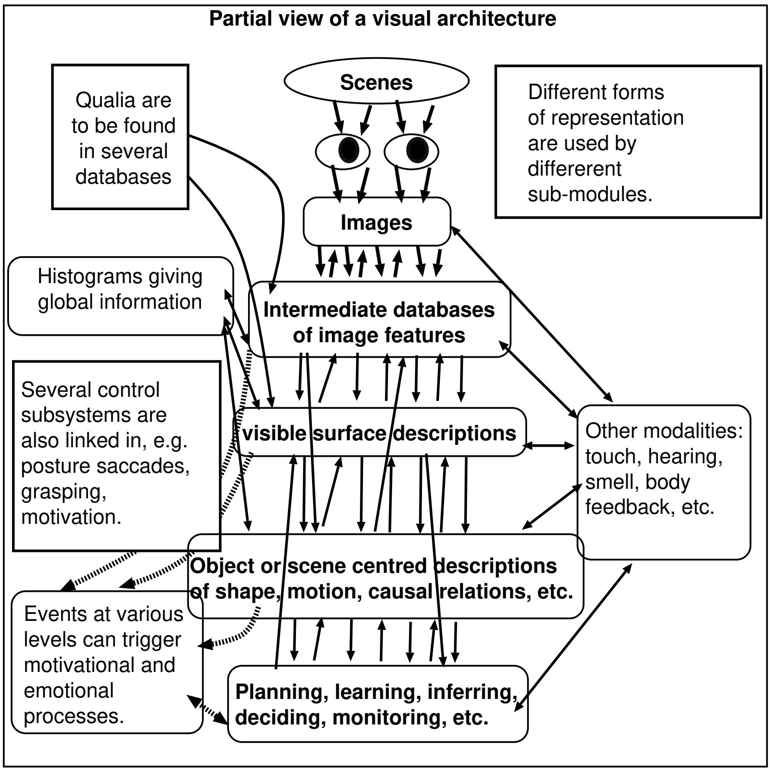
Partial view of a visual architecture
Qualia are to be found in several databases
Scenes
Different forms of representation are used by differerent sub-modules.
Images
Histograms giving global information
Intermediate databases of image features
Several control subsystems are also linked in, e.g. posture saccades, grasping, motivation.
visible surface descriptions
Other modalities: touch, hearing, smell, body feedback, etc.
Object or scene centred descriptions of shape, motion, causal relations, etc.
Events at various levels can trigger motivational and emotional processes.
Planning, learning, inferring, deciding, monitoring, etc.

Figure 7: **Towards an architecture for a visual system**

where an agent with deliberative capabilities decides that it should be thinking about something else and redirects attention, or wonders whether its motives are good ones, or notices that it frequently switches attention between unfinished tasks and therefore tries to achieve longer periods of undivided attention.

Of course, we all know that such control is not perfect, as we learn when our attention wanders from some boring but important task, or when our thoughts are drawn back to a painful episode that we'd rather forget.

In the diagrams I have tried to indicate that as the collection of layers in the architecture increases in sophistication, so will the collection of layers of processing and types of abstraction both in perceptual mechanisms and in action subsystems. Some of this can be the result of evolution, and some the result of individual learning, e.g. learning to read, to recognise artistic styles, or detect the strategy used by an opponent in some game. There are many complications hinted at in the diagrams which there is no space to discuss here.

However, I shall enlarge on the architecture of a visual subsystem to help explain how sensory qualia are to be expected within the sort of agent being described here.

13 Perception can use an intricate architecture

Figure 7 (summarising ideas presented in more detail in (Sloman 1989; 1996a)) is intended to indicate that visual perception is not just a matter of registering or recognising.

It also involves the following:

- Classification at different levels of abstraction: a square, a rectangle, a quadrilateral, a polygon, a figure.
- Interpretation: mapping from one domain to another. E.g. the 2-D optic array is interpreted in terms of a 3-D environment. Acoustic patterns are interpreted as meaningful speech.

- Grasping structure: seeing not only eyes, nose, mouth, arms, legs, hands, feet, but how they are related together. The hands are on the ends of the arms, but a finger may be touching the nose.
- Grasping patterns of change and motion: the wasp is flying towards the window, the car is moving forwards while its wheels are turning, the scissors are opening and shutting.
- Grasping more abstract possibilities and constraints inherent in objects in the environment (i.e. what J J Gibson called "affordances": a chair can support you, a table can obstruct motion, a door allows transfer to another room a window catch allows the window to be held open, a handle allows an object to be grasped, your prey is looking towards you, making detection more likely.)

In order to support this variety of capabilities, a human-like (or ape-like?) perceptual system needs to be able to create and manipulate a number of different sorts of rapidly changing representations of different sorts of information, making use of different intermediate information structures created during interpretation processes. For instance in expert speech understanding this could include the detection of phonemes, syllables, words and phrases. Fluent reading requires a different but related collection of intermediate levels.

Human perceptual architectures allow us to attend to some aspects of these *internal* information stores. E.g. learning to draw pictures, or sighting a gun, both of which require a person to attend to some aspect of how things look rather than how things are. I conjecture that this ability to attend to properties and relationships of intermediate structures in sensory systems, is the main source of philosophical interest in "qualia". But our access is both incomplete and unreliable, which is partly why there is a tendency to regard so-called phenomenal experiences as inherently simple and lacking in causal powers or causal explanations. An organism or robot that has access to some aspects of these internal states, presumably because this provides various biological advantages, will not necessarily also have access to the underlying mechanisms (Wittgenstein's "techniques") providing the substratum for the experiences.

Philosophically inclined robots with appropriate self-monitoring mechanisms within their meta-management layer can therefore be expected to wonder about the relationships between their qualia and the underlying physical mechanisms, and to consider the possibility that animals like humans (and perhaps some other species?) might have all the functional architecture of an intelligent robot, yet lack these qualia that are detectable only from the "first person" (or "first robot") viewpoint.

14 Varieties of perceptual consciousness

By relating the architecture of perceptual systems to other aspects of the architecture we can begin to dissect varieties of types of consciousness to be found in different sorts of animals, or artificial agents.

For example an insect with a purely reactive system, can be described as 'sentient' and capable of experiencing the environment in a rudimentary fashion, e.g. such as a fly that escapes an approaching fly swatter, or a bee that can orient itself in space to obtain nectar from a flower. It is also very likely that the perceptual mechanism of each type of insect has evolved to meet the requirements of that insect's niche. Thus two eyes with similar basic structure may obtain information from the same optic array (the same light cone) but process it differently, abstracting different properties and relationships and also link those properties and relationships to different forms of internal and external behaviour triggered by perceived situations.

Note that how a bee sees the world may be something we cannot describe because the types of abstraction and classification performed by its perceptual mechanisms need not map onto any relevant collection of concepts that we have developed for describing either our environment or our experience of the environment.

Any description in our normal language of what it is like to see like a bee will therefore *necessarily* be at best an approximation, and at worst seriously distorted. If, however, we learn enough about the design of a bee, then we may be able to produce new mathematical characterisations of the structure of a bee's experiences (e.g. their topology) and how they change as the bee moves etc.

An organism which includes a deliberative layer in its architecture will typically both use a different sort of perceptual system, able to provide additional information about the environment, and also use the information for quite different functions. In particular some chimps can, it appears, not only detect the presence of a stick lying on the ground but also see the possibility of grasping the stick, taking it to the bars of its cage, and using it to reach a banana lying outside. This requires not only perceiving actual structures and relationships in the environment but also *possibilities* and *constraints*, or what Gibson called "affordances". This is a necessary component of the deliberative ability to use such possibilities in contemplating in advance the possibilities of certain sorts of actions.

Such an animal both experiences far more complex and abstract aspects of the environment than a purely reactive agent, and also links the information thus obtained to a quite different collection of information processing capabilities. It may also have a reactive subsystem which responds at the same time in a far more primitive fashion: an example of this is the tendency to blink if a nearby object rapidly moves towards your eyes.

Neither a purely reactive agent nor an agent that includes deliberative capabilities can necessarily detect any of its own information states, or make use of those states in deciding what to do. If, however, there is also a meta-management layer which provides internal planning and decision making mechanisms with information about different aspects of the internal state, including for instance the contents of intermediate perceptual information stores, then this provides an additional range of types of awareness, for instance being able to attend to how things look to you as opposed to how they are. This may in some cases be related to an experience of being in control, for instance being able to switch attention, or to change viewpoint, or to move a perceivable part of one's body. This may also go with the experience of partly being out of control for instance when a reflex reaction occurs or a thought inexplicably comes to mind or one's thoughts merely wander.

I have tried in these comments to give an indication of the ways in which we may begin to distinguish different sorts of consciousness that can be supported by different designs, i.e. different sorts of mechanisms combined in an architecture to perform different roles. The full story is very much more complicated than what has been said here, but the reader should be able to extend the analysis by adopting the design standpoint and considering different ways in which information can be acquired, stored, transformed, and used. The implications for varieties of experience are very complex however, and hard to work out in the abstract.

15 Control states of varying scope and duration

I have tried to indicate, albeit very briefly, how the sorts of architectures sketched above could accommodate states involving perception, planning, practical knowledge, beliefs about the environment, motives generated by both external and internal processes, and various kinds of emotional states emerging from different features in the architecture. But I don't want to give the impression that this is anything like a complete and accurate survey. There are many additional types of control states that need to be explained by spelling out features of the architecture.

Some of them are global states, like moods (depression, elation, optimism, pessimism) which are not necessarily directed at anything in particular, but modulate many different kinds of processing including how perceptual contents are perceived, how decisions are taken, how actions are performed. This suggests a form of control whereby some global analog device or perhaps the degree of concentration of a particular chemical throughout the brain, might affect processing globally. It is easy to see how changes of these sorts might be based on learning about global features of the environment. It is also possible to see how such global control mechanisms could go wrong and produce pathological states.

Another type of mental state that is important for humans is an attitude which can lie dormant for some time

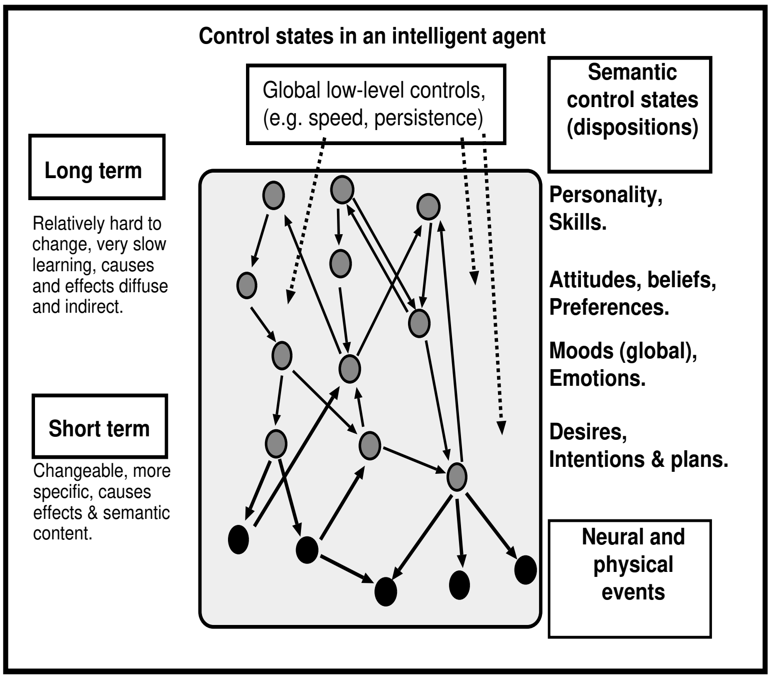
Control states in an intelligent agent
Global low-level controls, (e.g. speed, persistence)
Semantic control states (dispositions)
Long term
Relatively hard to change, very slow learning, causes and effects diffuse and indirect.
Short term
Changeable, more specific, causes effects & semantic content.
Personality, Skills.
Attitudes, beliefs, Preferences.
Moods (global), Emotions.
Desires, Intentions & plans.
Neural and physical events

Figure 8: **Control states of varying scope and duration**

and then be triggered into action. An example might be loyalty to one's country which most of the time does not enter one's thoughts, or affect one's behaviour, but which might interact with information about going to war or news of an international sporting competition, to produce new states such as jubilation about a victory or new actions such as joining an army or a group of supporters of the national team.

Attitudes, unlike moods as defined above, are states that are rich in semantic content, often involving a complex collection of beliefs, motives and preferences which can endure in a dormant state because nothing has occurred to which they are relevant. Nevertheless hearing news about someone or something to which one has an attitude, or seeing an event involving the object of an attitude can be an experience that is automatically coloured by this previously dormant collection of dispositions to react. If Fred is someone you care about then being asked "Have you heard the news about Fred?" may provoke an experience suffused with anxiety, whereas another person grasping the full meaning of the question is simply made curious.

Figure 8 is intended to indicate in a very crude way some of the relationships between different control states. The black dots represent individual events. The other dots represent more or less enduring states with particular information contents and particular potential influences on other states. The shaded background, with arrows pointing into it is supposed to indicate a global control environment which may be modulated by particular transient states and events produced by ongoing processes of various kinds. Some of these modulations may be focused only on particular sub-mechanisms, e.g. motive generation mechanisms or decision making mechanisms or attention controlling mechanisms.

The "higher" states depicted in the diagram are: harder to change, more long lasting, potentially subject to a wider range of influences, more general in their effects, more indirect in their effects. Some of them, e.g. certain aspects of personality, could be genetically determined, while others are largely a product of learning processes including very long term processes in which a whole culture is absorbed.

The lower level states are more transient, more directly under the control of current percepts and other processes, influencing short term feedback loops of various kinds. These are more easily changed by other

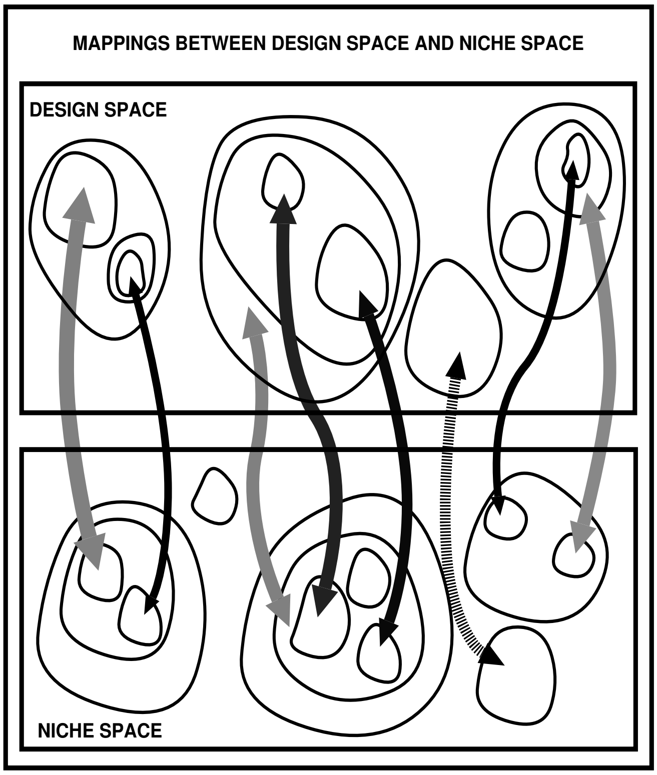

Figure 9: **Mappings between design-space and niche-space**

events and have more direct and specific effects on processing.

There is no implication that all or even most of this is consciously detectable or capable of being controlled consciously. However in humans and perhaps other animals that include a meta-management layer in the architecture it may be possible for some aspects of these states to be accessible and the information about one's current state (e.g. feeling depressed, feeling light hearted, etc.) to influence high level decision making. Sometimes this can be part of an undesirable positive feedback loop.

Items near the top of the diagram may be more difficult to detect, either because they are buried deep in associative memory mechanisms whose contents cannot be directly inspected or because they are implemented in a widely distributed control mechanism for which no synoptic internal percept is available, because they are parts of reactive subsystems not all of whose internal workings are accessible "from above", or for other reasons.

Although the comments in this section, like many others, have been extremely vague and sketchy they should suffice to rule out some of the more simplistic interpretations of the type of architecture being described here.

16 Designs and niches

Different combinations of capabilities correspond to different designs. As shown in Figure 9, there is a space of possible designs, describable at different levels of abstraction. There is also a space of sets of requirements, where a set of requirement is an engineering concept corresponding to the biologist's notion of a "niche". Both spaces have complex structures including many discontinuities (different discontinuities at different levels of abstraction), and there are various kinds of relationships between the spaces, as indicated in Figure 9.

Although we normally think of designs as created by people, there is a more general viewpoint according to which designs, like shapes, exist whether they have been thought of or not, and whether instances exist or not. Designs, like shapes, have properties, and those properties have consequences. In particular a design can *explain* the capabilities of a system that instantiates that design. A design presupposes an ontology of components, mechanisms, causal and functional relationships. Many designs use the ontology of an abstract machine, such as the virtual machine in a word processor, which contains chapters, pages, paragraphs, sentences, characters, fonts, diagrams, etc.

At a lower level the design may be implemented in a machine with a very different ontology, e.g. large collections of bit patterns in a typical computer.

Evolution can be seen as producing designs: though there is no designer or engineer, only natural selection. The notion of a design is an abstraction that has nothing to do with how the design was produced, or whether any agent intends it to serve any purposes. In that sense the design of a bee fits the requirement to pollinate plants and to collect pollen for the hive. Similarly the notion of a set of requirements, or "niche" does not presuppose anyone who imposes the requirements. We can also talk about possible niches, which might have existed, even if they don't actually exist in nature.

Actual niches for organisms may exist as pressures imposed by the environment, including other organisms. However the physical environment does not uniquely determine a niche, since different animals in the same physical environment can have different niches, and therefore require different designs, e.g. a bee and a hummingbird. These are natural niches. There are also artefactual niches arising out of requirements faced by designers trying to solve engineering problems. Some niches arise within an architecture, in the form of requirements for a sub-mechanism within the architecture.

Different sorts of designs correspond to different sorts of niches, though there is no one to one correspondence. In general, there are various tradeoffs: design D1 can fit niche N better than design D2 in respect of speed whereas D2 fits it better in terms of reliability. The different sorts of arrows in the figure are intended as a reminder that different sorts of design-niche relationships can exist.

AI can be seen as the general study of design-space, niche-space and their interrelations, for a very large, though ill-defined, class of designs and niches. This includes the study of trajectories in design-space and niche-space. There are different sorts of trajectories that have to be studied, for instance trajectories within a single agent involving development and learning, and trajectories involving evolution over several generations, as in artificial life studies and the use of genetic algorithms for solving problems.

Some trajectories are not possible within an individual (e.g. an cat's embryo cannot develop into a giraffe). However it may be possible for an evolutionary process over many generations to produce transitions that are impossible within an individual. There may be other kinds of trajectories in design-space and niche-space, e.g. those that require cultural development, or those that are not possible for a self-modifying adaptive system, but require external intervention, e.g. by an engineer.

When we have a better understanding of the dynamics of trajectories in design-space we may be able to produce theories about the evolution of various subsets of the capabilities involved in human consciousness.

17 Architecturally grounded concepts

It is time to return to conceptual issues. Each design for an architecture (together with the environment) determines a variety of states and processes that that architecture can support. This produces a family of concepts concepts, grounded in the architecture. In (Wright *et al.* 1996) it is suggested that the kind of architecture shown in Figure 6 can be elaborated so as to support many of the processes involved in prolonged grieving.

By studying different sorts of architectures, corresponding to different regions of design-space and of niche-space we can generate different families of concepts and provide a framework for comparative studies of animals and machines.

These theory-based concepts can elaborate and extend common sense concepts, as happened with our concepts of kinds of stuff and kinds of physical processes, when the atomic theory of matter explained the periodic table of the elements and the addition of the theory of valence systematically generated concepts of possible types of chemical compounds.

A similar architecture-based refinement of concepts of mentality could lead to conceptual advances in philosophy, psychology and biology to some extent parallelling the conceptual evolution in physics and chemistry that grew out of a new hypothesised architecture for matter. In both cases, we can expect the new concepts to elaborate and extend common sense concepts, instead of replacing them completely, for instance if the old concepts have already been used for many centuries to talk about instances of the architecture.

The new concepts will enable us to formulate new refined empirical questions. Instead of asking which animals are conscious and which are not, we might, for instance, define several different kinds of consciousness involving different collections of capabilities. We can then ask which animals have kind A, which have kind B, and whether infants of a certain age have kind C or D, and whether sufferers from Alzheimer's disease have kinds E or F, and so on. So the question "Which animals are conscious?" is replaced by "Which of the states based on such and such a class of architectures can this sort of animal have?" Likewise, the question how "it" evolved will be replaced with many questions about the evolution of different components or layers in the architecture and the different capabilities and forms of consciousness they support.

The answers may be different in different cases. For instance different niche pressures may have been involved in the development of awareness of features of the environment and development of awareness of states of one's own perceptual subsystem.

From this "design-based" standpoint, attempting to understand and model consciousness requires us not simply to explore one architecture or build one type of robot, but to think about varieties of architectures and trajectories in design-space and niche-space. Only in terms of similarities and differences between cases can we understand any particular architecture properly. E.g. we can then answer questions like: What difference would it make if this feature were absent from the architecture, or if this link between components were missing or damaged? That requires understanding a neighbourhood in design-space.

If we have an architecture-based theory, we can get a much deeper understanding of how a human mind normally works and also how it might go wrong: the more complex the architecture, the more ways it can go wrong. By applying all these ideas, we should be able to help therapy, counselling and education. For instance architecture-based theories of learning and development can replace educational theories that are mostly hunches and rules of thumb, and complex diseases involving multiple mental malfunctions, such as schizophrenia, can perhaps be better understood in terms of deviations from a normally functioning architecture.

Within the context of our exploration of design-space we can see how to replace some apparently endless debates with research that makes real progress, in philosophy and in science. In part this is because some of the debates are at cross purposes, with participants using words (e.g. "emotion", "consciousness") to refer to different sorts of things without being aware of the differences because they lack a conceptual framework in which to demarcate the distinctions.

The prizes to be won from such studies are high: including far better understanding of processes of evolution, processes of individual development, and also forms of treatment or therapy to help those who suffer from various kinds of abnormality or deficiency.

18 Relations to Dennett's work

It will be clear to anyone who has studied Dennett's work that there is considerable overlap between his ideas and mine. Curiously his book *Brainstorms* and my *The computer revolution in philosophy* were both published by Harvester press in 1978, though neither of us knew anything about the other's work before that, and our styles are totally different, despite the considerable similarities in our general approaches. However there are also important differences.

One difference is that Dennett has always emphasised the importance of the intentional stance, which I think is of little significance as a basis for explaining how our concepts of mentality actually work. I think our concepts are rooted in the design stance which we adopt implicitly. From an evolutionary point of view it is obviously advantageous for infants in a social species to have innate, genetically determined, abilities to take account of mental states and processes in others, instead of each one having to solve the mind body problem by some elaborate philosophical inference.

The idea is that we are born with innate dispositions to develop assumptions (both implicit and vague) about a cognitive architecture in other agents and ourselves, supporting a collection of distinctions between beliefs, desires, preferences, intentions, sensory experiences, hedonic states (pain and pleasure). These presumed mental states and processes and capabilities will be defined by their causal powers, of which we have only a very limited intuitive grasp, and to that extent the concepts are partly indeterminate.

Sometimes the genetic mechanisms are missing or do not work and a child develops without a grasp of these concepts (e.g. autistic children). In no case will they be totally determined by genetic mechanisms: like many other aspects of high level cognitive functioning our grasp of the design of a mind (our own and others) develops partly through interaction with the environment and cultural influences can play a deep role.

One thing we do not need to assume in ourselves or others is high level rationality: which is a presumption of the intentional stance. Given the right collection of causal powers in the mental architecture we can expect some processes to be rational and others not. I.e. the design stance makes the intentional stance redundant where the architecture is rich enough. In other cases (e.g. insects and some software systems) issues of rationality may be totally irrelevant: yet semantic belief-like states of a primitive type may be produced by perceptual mechanisms. I have discussed these issues and disagreements at greater length elsewhere (e.g. (Sloman 1994)).

However I believe that in Dennett's recent books the intentional stance has played an increasingly minor role compared with the design stance, so maybe that disagreement is disappearing. Moreover, if I say that an animal or child adopts the design stance in attributing mentality to others (or itself) and Dennett says it adopts the intentional stance, there will be a large measure of agreement regarding the predictions based on these claims: the core difference seems to be the extent to which the assumption of rationality is used and the extent to which an assumption of a network of causal mechanisms is used. Perhaps this can be resolved by empirical investigation.

Another difference is that Dennett apparently wishes to banish talk of qualia as incoherent, whereas I think there is a coherent concept to which some philosophers add incoherent and unnecessary extra baggage, which we can strip away. We are then left with a concept that refers to a host of familiar phenomena whose possibility needs to be explained, and the three layered architecture in "central" processes combined with multi-layered perceptual and motor systems, seems to me to provide a framework within which much can be explained (though a lot more detail is needed than I have space for here). I think Dennett just did not notice the important coherent concept buried in the incoherent one.

I suspect that he might be willing to go along with this sort of strategy without wishing to use the word "qualia", which I continue to use despite his protestations, because I think it has the right historical connections. (The word does not occur in the index to his 1996 book).

A third difference concerns the role of language. Dennett has always emphasised the role of an external communicative language as the basis for human consciousness, and even for thinking (see the discussion of animal thinking in (Dennett 1996) pages 159-160), whereas I have always thought that more primitive *internal* representational mechanisms are required for certain kinds of animal learning and plan construction and can then later be used for external language. For example production of a novel sentence (as children do frequently) requires internal compositional syntactic and semantic mechanisms which might turn out to have developed from more general action control mechanisms that evolved earlier. These are empirical issues not to be settled in our armchairs.

While I agree that we need to resist temptations to over-anthropomorphize, we also need to do further analysis of the requirements for the *internal* deliberative capabilities which make it possible for an animal (or robot) to create, evaluate and use new action sequences or plan structures, whether by merely "thinking" about them or by learning from the observed successful and unsuccessful actions of others. (It's likely that the ability to do the latter developed before the ability to simulate the process in internal planning, for the purely internal processes need a far more sophisticated ability to store generalisations about the consequences of actions in various circumstances, whereas learning from observation of the successes and failures of others uses the environment as the store of generalisations about itself.)

It may turn out that *before* an external human-like language can be developed, more primitive capabilities are required of the type that are needed in any case for the simpler sorts of deliberative mechanisms sketched above. In that case those mechanisms could provide a basis for primitive forms of thinking (e.g. about what would happen if), and even some primitive types of thinking about thinking e.g. the ability to detect that a proposed plan has unnecessary steps and can be shortened. This ability to detect and make use of information about plan structures seems to be related to a number of kinds of mathematical reasoning, e.g. reasoning about the equivalence of algorithms or mappings between structures. Perhaps some primitive and ancient ability to create and reflect on plans with loops provides the basis of our thinking about infinite structures, e.g. the natural number series and infinitely long lines. We should not take it for granted that somehow external linguistic ability was able to develop without presupposing any comparable internal mechanism and then fed back into the system as the major or main source of our abilities to think and reason. Perhaps further research will show that there was gradual development of both sorts with mutual feedback and a steadily growing dependence on cultural learning, based on ever more sophisticated internal mechanisms able to take advantage of cultural benefits.

19 Caveat: currently understood mechanisms may, or may not suffice

I'll mention one caveat. It may be thought that my references to information processing implies some sort of commitment to an implementation based on current computing concepts. There is no such *a priori* commitment, for it is an empirical question whether they will suffice. We may find that high level human-like architectures can be implemented on a number of different sorts of low level mechanisms – for instance, computer-based mechanisms as well as naturally occurring mechanisms.

Alternatively we may have to invent new mechanisms before we can make robots with human capabilities (ignoring, for now, the question whether we *ought* to do so). Our knowledge of what is and is not possible using computers is still very limited. We have been trying to explore what computers can do for only forty or fifty years. A comprehensive overview may take hundreds of years, especially as we discover new ways of linking large numbers of computers into an integrated system, and new ways of combining digital and analog mechanisms and perhaps new forms of chemical mechanisms capable of supporting complex information processing.

However, we may also find that, provided the global architecture is right, there are several *different* types of implementation mechanisms with sufficient richness and structural variability to support the architecture. Different implementations may turn out to differ no more than different humans do already. In short, we may discover that architecture dominates mechanism.

Alternatively we may find that because of laws of physics only systems that are largely like animal brains can satisfy the constraints on size, weight, energy consumption and processing capabilities required for human like features.

Sometimes people produce spurious refutations of the possibility of computer based implementations of mind by extrapolating from the properties of computers to the properties of virtual machines implemented in computers. For instance it is sometimes argued that all states in a computer are binary, and therefore everything implemented in a computer must be determinate in the sense that for any property, at every moment the system either does or does not have that property. This may be true of the simplest sorts of computing systems, but it is not necessarily true of all. This is an important fact because a kind of indefiniteness seems to be one of the features of human conscious states, as I've indicated above, e.g. in connection with peripheral vision and the blind spot.

There are many aspects of human experience that are inherently indeterminate. Besides peripheral vision there are cases of seeing something in fog or darkness with a shape that you cannot make out, states of indecision about what to do, propositions that one half believes and half doesn't, theories that one only dimly understands, past episodes that one vaguely recalls, a face that reminds you of some person though you can't be sure which person, and so on. E.g. when you look at the stars on a clear night there's a fixed number of them visible from your current viewpoint. But it does not follow that there's a definite number of experienced light points, including points in the periphery of your field of view.

This kind of indeterminacy is one of the things that makes imagining a scene quite unlike seeing a picture: a physical picture on a physical surface cannot have the kind of indeterminacy that we find in mental states. For instance a fuzzy picture is definitely fuzzy when you look directly at it, in a way that is different from the indeterminacy described above.

None of the widely used current (in 1997) forms of information storage appears to have these features: a data-structure or image array either does or does not have certain contents. However, there is no difficulty in principle in accommodating structures with indeterminate contents within appropriate virtual machines. An obvious example would be to generalise the existing mechanisms for "lazy evaluation". For example in languages with lazy evaluation it is possible for a list to contain infinitely many elements, because the Nth element is computed as soon as something attempts to access it. Other modules cannot tell that the components of a lazy list are not stored explicitly all the time. If the computation is partly probabilistic then the Nth element of the list may have a fluctuating character giving the rest of the system the impression of fuzziness.

Similarly a large 2-D array might have many cells whose contents are computed only when required. In these cases it may sometimes be useful to store information about the general type of contents to be found in different parts of the array, without having the actual contents already created. This may be useful for deciding which elements to examine in order to obtain the actual contents. In addition the process of creating the contents may be partly unpredictable, for instance, if it uses a probabilistic procedure or if it depends on the state of an external sensor. In such a case the information available to another subsystem about the contents of the array in advance of doing the actual evaluation would have a kind of indeterminacy. This could mean that decisions based on that information would have to be hedged in some way.

It is quite possible that the indeterminacy in human mental states is of a very different kind from this. The important point for now is that we should not extrapolate from properties of the underlying implementation medium (digital, discrete, determinate) to properties of higher level virtual machines implemented in that medium.

It is also important to counteract any impression that might have been given in earlier parts of the paper that I was talking only about architectures whose information states were determinate.

20 Conclusion

The ideas described here have many and varied implications. An important subclass of implications has to do with the fact that if you know about an architecture you have a basis for understanding not only how it normally works but also ways in which it can go wrong and what sorts of intervention can remedy the situation. In the case of human architectures this may eventually provide a far more solid theoretical basis for diagnosing various kinds of abnormality or malfunction, and recommending far better forms of therapy, counselling, education or other intervention.

A design-based theory can generate many new descriptive concepts, which can then be used to formulate a host of new empirical questions to be settled by neurophysiological, psychological and biological research.

Understanding virtual machines and how they relate to the mechanisms in which they are *implemented* and other mechanism in the environment is an important task for philosophy. The engineer's concept of one thing being implemented in another is a (comparatively) well understood special case of the philosophers' notion of one thing being *supervenient* on another. But there is still much work to be done clarifying these concepts.

In particular we need to improve our grasp of the constraints a particular high level virtual machine can impose on possible implementation machines. For instance, although this is still an open question, we may find that most of the high level aspects of a human-like architecture that enable people to form part of a human society can be implemented on quite different sorts of low level mechanisms (e.g. computer-based mechanisms). We can summarise this as the conjecture that for the purposes of fitting into a human social niche *architecture dominates mechanism*. To what extent this is true is a topic for further research.

It could be true for high level social capabilities even if many other aspects of human functioning, e.g. reproductive capability and states like feeling an itch or feeling thirsty cannot be implemented in some of the socially competent human-like architectures, because they lack the relevant mechanisms.

We need to study "design-space": the space of possible designs for systems, with different combinations of capabilities. This is linked to another kind of space, "niche-space", which includes what biologists study when they talk about the niche of an organism and includes what engineers study when they talk about the requirements against which a design is to be evaluated. Niche space and design-space are both very complex structures, containing many discontinuities, and the mappings between them are very complex, involving many trade-offs.

Recent research has opened up many unanswered questions regarding those spaces. What are the dynamics of these spaces: what sorts of trajectories are possible within them? Which sorts can occur in individual development? Which sorts of trajectories require evolution across generations involving many individuals? Which can occur in a laboratory? Which physical mechanism scan support which sorts of designs, and design trajectories?

We can study different ways in which transitions can occur from one collection of capabilities to another. Some involve change within an individual, some may not be capable of occurring without an evolutionary process involving generations of individuals. Perhaps an evolutionary process could produce human beings with a strong ability to detect magnetic fields. Their consciousness would be different from ours. Moreover the form of that consciousness would be different depending whether the magnetic sensors could influence only the reactive subsystem or whether they could also feed information about affordances to the deliberative system. It would also be different depending on whether intermediate information structures in the processing of magnetic data were accessible to meta-management processes or not.

Different combinations of capabilities correspond to different designs. Designing an overhead projector involves combining physical capabilities and some cognitive ones (it should be easy to find the controls). Designing a piece of software involves combining more abstract abilities, for instance to calculate, reformat a document, check spelling or grammar, solve puzzles. Some designs inherently involve a particular physical implementation (e.g. a violin) whereas others admit considerable variety in their implementation (e.g. a particular type of word processor may run on computers with very different physical structures).

The theoretical part of Cognitive Science coincides with the subset of AI that studies regions of design-space and niche-space containing human and animal designs. We still do not know much about the overall structure of design-space, nor which kinds of mechanisms are needed as substructures for particular sorts of designs. Nor do we know much about the classes of designs that are relevant to characteristically human capabilities, and we understand very little about the dynamics of such designs and the variety of possible trajectories in design-space. Thus Cognitive Science, like AI, is necessarily in its infancy.

In our current state of relative ignorance we should get on with the job, perhaps pausing from time to time to marvel at the complexity of the phenomena and mechanisms we are grappling with. By contrast pontificating about whether machines can or cannot be conscious or about whether consciousness does or does not need quantum gravity engines, or whether particular animals are or are not consciousness seems to me to be just silly, or at best premature. There is far too much research still to be done.

Instead of asking a few big, but empty and largely unanswerable, questions about consciousness we can ask, and perhaps answer, a large number of smaller, even more fascinating, questions about all the myriad components of the cluster. In particular, by considering how some sophisticated robots in the future including meta-management layers are likely to reflect on aspects of their knowledge of their own sensory states we can see how they are likely to be drawn into philosophical speculation and discussion about the problem of consciousness. Explaining in great detail why this is inevitable is part of the answer to the problem of consciousness.

A more profound long term implication is the possibility of a new science investigating laws governing possible trajectories in design-space and niche-space, as these form parts of high order feedback loops in the biosphere.

Acknowledgements and Notes

This work was supported by the UK Joint Council Initiative, the Renaissance Trust, and a grant from the University of Birmingham.

I thank colleagues in the Cognition and Affect project at the University of Birmingham, including Steve Allen, Luc Beaudoin, Chris Complin, Darryl Davis, Glyn Humphreys, Brian Logan, Christian Paterson, Riccardo Poli, Tim Read, and Ian Wright. Greg Mulhauser's comments on an early draft led to many improvements. I have also benefitted from discussions with Stan Klein, Roz Picard, and Henry Stapp. The work has benefited indirectly from a collaborative project on autonomous agents, funded by DERA Malvern, involving Jeremy Baxter.

This "position paper" overlaps with and extends the collection of papers in the Cognition and Affect ftp directory. **ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/**

The project overview is to be found at
http://www.cs.bham.ac.uk/research/projects/cogaff/

Bibliography

So much has been and is being written on consciousness, how brains work, evolution, the scope and limits of AI, artificial life, etc. that a full bibliography of relevant literature would take very many pages, and would quickly be out of date. A very small subset of directly relevant material is listed below. Readers may also find useful the ever growing on-line bibliography maintained by David Chalmers:

http://ling.ucsc.edu/˜chalmers/biblio.html

Additional papers and presentations from the Cognition and Affect group at the university of Birmingham can be found at

http://www.cs.bham.ac.uk/research/projects/cogaff/
http://www.cs.bham.ac.uk/research/projects/cogaff/talks

Further notes and discussions in various forms of incompleteness can be found in the author's "miscellaneous" web directory:

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/AREADME.html

References

Bernard J. Baars. *A cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, UK, 1988.

Bernard J. Baars. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, New York, Oxford, 1997.

L.P. Beaudoin. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham, Birmingham, UK, 1994.

N. Block. On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18:227–47, 1995.

M. A. Boden. *Artificial Intelligence and Natural Man*. Harvester Press, Hassocks, Sussex, 1977. Second edition 1986. MIT Press.

V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology*. The MIT Press, Cambridge, MA, 1984.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, pages 139–159, 1991.

David J Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, Oxford, 1996.

D. C. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, MA, 1978.

D. C. Dennett. *Elbow Room: the varieties of free will worth wanting*. Oxford: The Clarendon Press, 1984.

D. C. Dennett. *Consciousness Explained*. Penguin Press, London and New York, 1991.

Daniel C. Dennett. *Kinds of minds: towards an understanding of consciousness*. Weidenfeld and Nicholson, London, 1996.

Stan Franklin. *Artificial Minds*. Bradford Books, MIT Press, Cambridge, MA, 1995.

J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

D. Goleman. *Emotional Intelligence: Why It Can Matter More than IQ*. Bloomsbury Publishing, London, 1996.

Douglas R Hofstadter. *Godel, Escher, Bach: an Eternal Golden Braid.* The Harvester Press, Hassocks, 1979.

Immanuel Kant. *Critique of Pure Reason.* Macmillan, London, 1781. Translated (1929) by Norman Kemp Smith.

J. McCarthy. *Formalising Common Sense.* Ablex, Norwood, New Jersey, 1990.

M. L. Minsky. *The Society of Mind.* William Heinemann Ltd., London, 1987.

T. Nagel. What is it like to be a bat. In D.R. Hofstadter and D.C.Dennett, editors, *The mind's I: Fantasies and Reflections on Self and Soul*, pages 391–403. Penguin Books, 1981.

R.W. Picard. *Affective Computing.* MIT Press, Cambridge, MA; London, England, 1997.

G. Ryle. *The Concept of Mind.* Hutchinson, London, 1949.

J.R. Searle. Minds brains and programs. *The Behavioral and Brain Sciences*, 3(3), 1980. (With commentaries and reply by Searle).

H. A. Simon. Motivational and emotional controls of cognition. In H. A. Simon, editor, *reprinted in Models of Thought*, pages 29–38. Yale University Press, Newhaven, CT, 1967.

A. Sloman and M. Croucher. Why robots will have emotions. In *Proc 7th Int. Joint Conference on AI*, pages 197–202, Vancouver, 1981. IJCAI.

A. Sloman. Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, pages 209–226, London, 1971. William Kaufmann. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971.

A. Sloman. *The Computer Revolution in Philosophy.* Harvester Press (and Humanities Press), Hassocks, Sussex, 1978. http://www.cs.bham.ac.uk/research/cogaff/62-80.html#crp, Revised 2018.

A. Sloman. The structure of the space of possible minds. In S. Torrance, editor, *The Mind and the Machine: philosophical aspects of Artificial Intelligence*. Ellis Horwood, Chichester, 1984. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#49a.

A. Sloman. What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles, 1985. IJCAI.

A. Sloman. On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337, 1989.

A. Sloman. Notes on consciousness, 1990.

A. Sloman. The emperor's real mind. *Artificial Intelligence*, 56:355–396, 1992. Review of Roger Penrose's *The Emperor's new Mind: Concerning Computers Minds and the Laws of Physics*.

A. Sloman. The mind as a control system. In C. Hookway and D. Peterson, editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, Cambridge, UK, 1993.

A. Sloman. Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering*, 349(1689):43–58, 1994.

A. Sloman. What sort of control system is able to have a personality?, 1995. (Presented at Workshop on Designing personalities for synthetic actors, Vienna, June 1995).

A. Sloman. Actual possibilities. In L.C. Aiello and S.C. Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proc. 5th Int. Conf. (KR '96)*, pages 627–638, Boston, MA, 1996. Morgan Kaufmann Publishers.

A. Sloman. Beyond turing equivalence. In P.J.R. Millican and A. Clark, editors, *Machines and Thought: The Legacy of Alan Turing (vol I)*, pages 179–219. The Clarendon Press, Oxford, 1996. (Presented at Turing90 Colloquium, Sussex University, April 1990.

A. Sloman. Towards a general theory of representations. In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science*, pages 118–140. Intellect Books, Exeter, U.K., 1996.

A. Sloman. What sort of control system is able to have a personality. In R. Trappl and P. Petta, editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Springer (Lecture Notes in AI), Berlin, 1997. https://link.springer.com/chapter/10.1007/BFb0030576.

A. Sloman. Supervenience and implementation draft. 1998(in preparation).

A. Sloman. What sort of architecture is required for a human-like agent? In Michael Wooldridge and Anand Rao, editors, *Foundations of Rational Agency*, pages 35–52. Kluwer Academic, Dordrecht, 1999. http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#21.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950. (reprinted in E.A. Feigenbaum and J. Feldman (eds) *Computers and Thought* McGraw-Hill, New York, 1963, 11–35).

L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953. (2nd edition 1958).

I.P. Wright, A. Sloman, and L.P. Beaudoin. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126, 1996.

R.A. Young. The mentality of robots. *Proceedings of the Aristotelian Society, Supplementary Vol. 68*, pages 199–227, 1994.