Aaron Sloman and Ron Chrisley

Virtual Machines and Consciousness

Abstract: *Replication or even modelling of consciousness in machines requires some clarifications and refinements of our concept of consciousness. Design of, construction of, and interaction with artificial systems can itself assist in this conceptual development. We start with the tentative hypothesis that although the word 'consciousness' has no well defined meaning, it is used to refer to aspects of human and animal information processing; and argue that we can enhance our understanding of what these might be by designing and building virtual-machine architectures capturing various aspects of consciousness. This activity may in turn nurture the development of our concepts of consciousness, showing how an analysis based on information processing virtual machines answers old philosophical puzzles as well enriching empirical theories. This process of developing and testing ideas by developing and testing designs leads to gradual refinement of many of our pre-theoretical concepts of mind, showing how they can be construed as implicitly 'architecture-based' concepts. Understanding how human-like robots with appropriate architectures are likely to feel puzzled about qualia may help us resolve those puzzles. The concept of 'qualia' turns out to be an 'architecture-based' concept, while individual qualia concepts are 'architecture-driven'.*

I: Introduction

The study of consciousness, a long-established part of philosophy of mind and of metaphysics, was banished from science for many years, but has recently re-entered (some would say by a back door that should have been kept locked). Most AI researchers ignore the topic, though people discussing the scope and limits of AI do not. We claim that much of the discussion of consciousness is confused because what is being referred to is not clear. That is partly because 'consciousness' is a *cluster concept*, as explained below.

Correspondence:
School of Computer Science, The University of Birmingham, Birmingham B15 2TT, U.K.
Email: A.Sloman@cs.bham.ac.uk   R.L.Chrisley@cs.bham.ac.uk

Progress in the study, or modelling, of consciousness requires some clarifications and refinements of our concepts. Fortunately, design of, construction of, and interaction with artificial systems can itself assist in this conceptual development. In particular, we start with the tentative hypothesis that although the word 'consciousness' has no well defined meaning, it is used to refer to a cluster of aspects of information processing in humans and other animals. On that basis we can enhance our understanding of what these aspects might be by designing, building, analysing, and experimenting with virtual-machine architectures which attempt to elaborate the hypothesis. This activity may in turn nurture the development of our concepts of consciousness, along with a host of related concepts, such as 'experiencing', 'feeling', 'perceiving', 'believing', 'wanting', 'enjoying', 'remembering', 'noticing' and 'learning', helping us to see them as dependent on an implicit theory of minds as information processing virtual machines. On this basis we can find new answers to old philosophical puzzles as well as enriching our empirical theories. We expect this process to lead to gradual refinement and extensions of many of our pre-theoretical concepts of mind as 'architecture-based' concepts, partly mirroring the development of other pre-scientific concepts under the influence of scientific advances (Sloman, 2002). The result, it is hoped, is that the successor concepts will be free of the many conundra (such as the apparent possibility of zombies) which plague our current, inchoate concept of consciousness.

Specifically, we hope to explain how an interest in questions about consciousness in general and *qualia* in particular arises naturally in intelligent machines with a certain sort of architecture that includes a certain sort of 'meta-management' layer. Explaining the possibility (or near-inevitability) of such developments illustrates the notion of 'architecture-driven' concepts (concepts likely to be generated within an architecture) and gives insightful new explanations of human philosophical questions, and confusions, about consciousness.

We emphasise the importance of the notion of a virtual machine architecture and use that as the basis of a notion of *virtual machine functionalism* (VMF) which is immune to the common attacks on more conventional functionalist analyses of mental concepts which require all internal states and processes to be closely linked to possible input–output relations of the whole system. We propose that science, engineering and philosophy should advance simultaneously, and offer a first-draft, general architectural schema for agent architectures, which provides a useful framework for long-term AI research both on human-like systems and models of other animals, and may also inspire new developments in neuropsychology, and a new understanding of the evolution of mind as well as advancing philosophy.

II: Confused Concepts of Consciousness

Before considering the possibility and details of machine consciousness, we might wonder: what do we mean by 'consciousness'? Let's start with some questions:

- Is a fish conscious?
- Is a fly conscious of the fly-swatter zooming down at it?
- Is a new-born baby conscious (when not asleep)?
- Are you conscious when you are dreaming? Or sleep-walking?
- Is the file protection system in an operating system conscious of attempts to violate access permissions?
- Is a soccer-playing robot conscious? Can it be conscious of an opportunity to shoot?

Not only do different people give different answers to these and similar questions, but it seems that what they understand consciousness to be varies with the question. A central (though unoriginal) motif of this paper is that our current situation with respect to understanding consciousness (and many other mental phenomena, e.g. learning, emotions, beliefs) is similar to the situation described in 'The Parable of the Blind Men and the Elephant'.[1] Six blind men encounter an elephant. Each feels a different part, and infers from the properties of the portion encountered the nature of the whole (one feels the tusk and concludes that he has encountered a spear, another feels the trunk and deduces that he has met a snake, etc.). It is often suggested that we are in the same position with respect to consciousness: different (even incompatible) theories may be derived from correct, but incomplete, views of reality.

As a result (or perhaps because) of this, 'consciousness' is a *cluster concept*, in that it refers to a collection of loosely-related and ill-defined phenomena. It's actually worse than that: the problem is not just that 'consciousness' is a cluster concept, but also that the collection of phenomena which is being gestured towards itself changes with context, e.g. when we think about consciousness in different animals, or in human infants. So it's a 2nd-order cluster concept. Contrast Block's claim (Block, 1995) that 'consciousness' is a *mongrel concept*. By this he means that it confuses a collection of concepts which he believes to be individually well-defined; e.g., 'phenomenal consciousness' and 'access consciousness'. In contrast, not only is it difficult to define precisely any particular cluster concept or the concepts it collects together; it is also difficult to define the concept 'cluster concept' precisely ('cluster concept' may itself be a cluster concept, in which case it would (again) be a 2nd-order cluster concept).[2] A cluster concept is not the same as a *vague* concept (e.g. 'large', 'yellow') for which a boundary cannot be precisely specified along some continuum. A cluster concept is one that is used *in an indeterminate way* to refer to an ill-defined subset of a collection of features or abilities. Some combinations of features are treated as definitely justifying application of the concept, and others definitely fail to justify it, whereas for some intermediate combinations the question whether something is or is not an instance is indeterminate. People may disagree, and the same person may find conflicting reasons for and against applying the concept in those

---

[1] John Godfrey Saxe, 1816–1887; see, e.g., http://www.wvu.edu/~lawfac/jelkins/lp-2001/saxe.html

[2] Yet in a different sense of '2nd-order' than mentioned before; does this make it a 3rd-order cluster concept?

cases. Thus, most people will agree that humans have emotions and that viruses do not, but may disagree as to whether insects, or fish, or unborn human infants do. This is not an *empirical* disagreement, for they cannot agree on what sorts of new evidence would settle the question.

This indeterminacy generates unresolvable disputes about difficult cases. However, some cluster concepts can be refined by basing more determinate variants of them on our architectural theories, as happened to concepts of kinds of matter when we learnt about the architecture of matter. Indeterminate concepts bifurcate into more precise variants (Sloman, 2002). We'll illustrate this below.

1. Evidence for confusion, and partial diagnosis

Our blind (or short-sighted) groping, together with our struggles to treat an indeterminate cluster concept as if it were well defined, have resulted in an astonishing lack of consensus about consciousness, as illustrated in table 1.

Table 1. A Babel of views on the nature of consciousness

Some say consciousness is . . .	While others say consciousness is . . .
Absent when you are asleep	Present when you dream
Absent when you are sleep-walking	Present when you are sleep-walking
Essential for processes to be mental	Not required (there are unconscious mental processes)
Able to cause human decisions and action	Epiphenomenal (causally inefficacious)
Independent of physical matter (i.e. disembodied minds are possible)	A special kind of stuff somehow produced by physical stuff
Just a collection of behavioural dispositions	Just a collection of brain states and processes, or a neutral reality which has both physical and mental aspects
Just a myth invented by philosophers, best ignored	Something to do with talking to yourself
Something you either have or don't have	A matter of degree (of something or other)
Impossible without a public (human) language	Present in animals without language
Present only in humans	Present in all animals to some degree
Located in specific regions or processes in brains	Non-localisable; talk about a location for consciousness is a 'category mistake'
Necessarily correlated with specific neural events	Multiply realisable, and therefore need not have fixed neural correlates
Not realisable in a machine	Realisable in a machine that is (behaviourally; functionally) indistinguishable from us
Possibly absent in something (behaviourally; functionally) indistinguishable from us (zombies)	Necessarily possessed by something which has the same information processing capabilities as humans

Some people offer putative definitions of 'consciousness', for instance defining it as 'self-awareness', 'what it is like to be something', 'experience', 'being the subject of seeming' or 'having somebody home', despite the fact that nothing is achieved by defining one obscure expression in terms of another. We need to find a way to step outside the narrow debating arenas to get a bigger picture. Definitions are fine when based on clear prior concepts. Otherwise we need an alternative approach to expose the conceptual terrain. Hopefully then we'll then see all the sub-pictures at which myopic debaters peer, and understand why their descriptions are at best only part of the truth.

In order to see the bigger picture, it will help to ask why there is a Babel of views in the first place. We suggest that discussion of consciousness is confused for several reasons, explained further below:

- Some people focus on one case: normal adult (academic?) humans, whilst others investigate a wider range of cases, including people with brain damage, infants, and other animals.
- Many thinkers operate with limited ideas about possible types of machines (due to deficiencies in our educational system).
- There is especially a lack of understanding about virtual, information-processing machines resulting in ignorance or confusion about entities, events, processes and states that can exist in such machines.
- Many people are victims of the illusion of 'direct access' to the nature of consciousness, discussed below.
- Some people want consciousness to be inexplicable by science, while others assume that it is a biological phenomenon eventually to be explained by biological mechanisms. (We aim to show how this might be done.)

2. Introspection can be deceptive

A Golden Rule for studying consciousness: Do not assume that you can grasp the full nature of consciousness *simply* by looking inside yourself, however long, however carefully, however analytically.

Introspection, focusing attention inwardly on one's own mental states and processes, is merely one of many types of perception. Like other forms of perception it provides only information that the perceptual mechanism is able to provide! Compare staring carefully at trees, rocks, clouds, stars and animals hoping to discover the nature of matter. At best you learn about a subset of what needs to be explained. Introspection can also give incorrect information, for instance convincing some people that their decisions have a kind of freedom that is incompatible with physical causation, or giving the impression that their visual field is filled with uniformly detailed information in a wide solid angle, whereas simple, familiar experiments show that in the region of the blind-spot there is an undetectable information gap. People can be unaware even of their own strong emotions, such as jealousy, infatuation and anger. Introspection can deceive people into thinking that they understand the notion of two distant events being

simultaneous, even when they don't: simultaneity can be experienced directly, but Einstein showed that that did not produce full understanding of it.
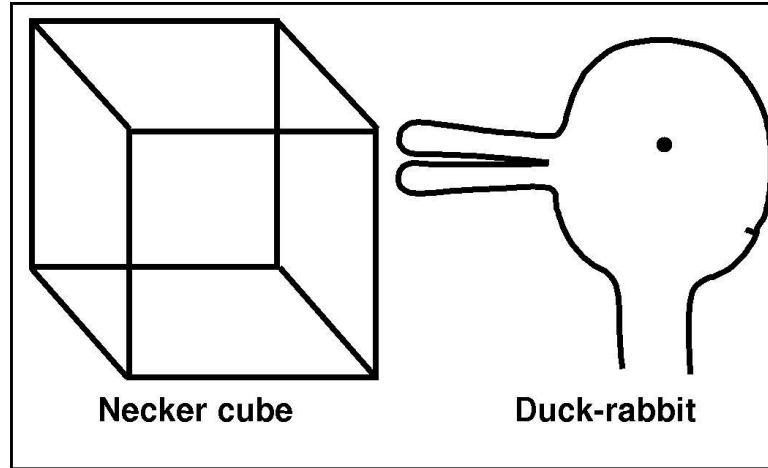

Figure 1. The Necker cube and the duck-rabbit: both are visually ambiguous, leading to two percepts. Describing what happens when they 'flip' shows that one involves only geometrical concepts whereas the other is more abstract and subtle.

3. Introspection can be useful for science

However, introspection is not useless: on the contrary, introspectively analysing the differences between the (probably familiar) visual flips in the two pictures in figure 1 helps to identify the need for a multi-layered perceptual system, described below. When the Necker cube 'flips', all the changes are geometric. They can be described in terms of relative distance and orientation of edges, faces and vertices. When the duck-rabbit 'flips', the geometry does not change, though the functional interpretation of the parts changes (e.g., 'bill' into 'ears'). More subtle features also change, attributable only to animate entities. For example, 'Looking left', or 'looking right'. A cube cannot look left or right. What does it *mean* to say that you 'see the rabbit facing to the right'? Perhaps it involves seeing the rabbit as a *potential mover*, more likely to move right than left. Or seeing it as a *potential perceiver*, gaining information from the right. What does categorising another animal as a perceiver involve? How does it differ from categorising something as having a certain shape?[3] We return to the multiplicity of perception, in explaining the H-CogAff architecture, in section V.5.

These introspections remind us that the experience of seeing has hidden richness, involving a large collection of unrealised, un-activated, but potentially activatable capabilities, whose presence is not immediately obvious. Can we say more about what they are? One way is to learn from psychologists and brain scientists about the many bizarre ways that these capabilities can go wrong. But we

[3] Compare the discussion of 'Seeing as' in Part 2 section xi of Wittgenstein (1953).

can also learn new ways of looking at old experiences: For example, how exactly do you experience an empty space, as in figure 2?

Figure 2. The final frontier?

Humans (e.g. painters, creators of animated cartoons, etc.) can experience an empty space as full of possibilities for shapes, colours, textures, and processes in which they change. How? Can any other animal? Current AI vision systems cannot; what sort of machine could? An experience is constituted partly by the collection of implicitly understood *possibilities for change* inherent in that experience. This is closely related to Gibson's 'affordance' theory (Gibson, 1986; Sloman, 1989;1996) discussed below.[4]

The Iceberg conjecture: Consciousness as we know it is *necessarily* the tip of an iceberg of information processing that is mostly totally inaccessible to consciousness. We do not experience what it is to experience something.

The examples show that when we have experiences there may be a lot going on of which we are normally completely unaware, though we can learn to notice some of it by attending to unobvious aspects of experiences, e.g. unnoticed similarities and differences. So we use introspection to discover things about the phenomenology of experience — contributing both to the catalogue of what needs to be explained and to specifications of the internal states and processes required in a human-like machine.[5]

4. Beyond introspection

Although introspection is useful, we must do more than gaze at our internal navels; we need to collect far more data-points to be explained, e.g. concerning:

- the varieties of tasks for which different sorts of experiences are appropriate — e.g. what sorts of experiences support accurate grasping movements, obstacle avoidance, dismantling and re-assembly of a clock, avoiding a

---

[4] Compare Wittgenstein: 'The substratum of this experience is the mastery of a technique' (op. cit.).

[5] A more robust argument for the Iceberg conjecture, especially in its strong modal form, above, cannot be given here; one can find some support for it in Wittgenstein (1922, cf 2.172: 'The picture, however, cannot represent its form of representation; it shows it forth.') and in Smith (1996, p. 303).

predator, catching fast moving prey, noticing that you are about to say something inconsistent, being puzzled about something, etc.;

- individual differences e.g. experiences at various stages of human development;
- culture-based differences in mental phenomena, e.g. *feeling sinful* is possible in some cultures but not others, and experiencing this text as meaningful depends on cultural training;
- surprising phenomena demonstrated in psychological experiments (e.g. change blindness) and surprising effects of brain damage or disease, including the fragmentation produced by severing the corpus callosum, or bizarre forms of denial;
- similarities and differences between the experiences of different species;
- stages and trends in the evolution of various mental phenomena.

Insofar as different organisms, or children, or people with various sorts of brain damage or disease, have different kinds of mental machinery, different information processing architectures, the types of experiences possible for them will be different. Even if a lion lies down with a lamb, their visual experiences when gazing at the same scene may be different because of the affordances acquired in their evolutionary history.

Understanding our own case involves seeing how we fit into the total picture of biological evolution and its products, including other possible systems on other planets, and also in future robots. It is important to pursue this idea without assuming that states of consciousness of other animals can be expressed in our language. E.g. 'the deer sees the lion approaching' may be inappropriate if it implies that the deer uses something like our concepts of 'lion' and 'approaching'. We must allow for the existence of forms of experience that we cannot describe.[6]

However, 'ontological blindness' can limit the data we notice: we may lack the ability to perceive or conceive of some aspects of what needs to be explained, like people who understand what velocity is but do not grasp that a moving object can have an instantaneous acceleration and that acceleration can decrease while velocity is increasing. *We do not yet have the concepts necessary for fully understanding what the problem of consciousness is.* So effective collection of data to be explained often requires us to refine our existing concepts and develop new ones — an activity which can be facilitated by collecting and attempting to assimilate and explain new data, using new explanatory theories, which sometimes direct us to previously unnoticed phenomena.

We need deeper, richer forms of explanatory theories able to accommodate *all* the data-points, many of which are qualitative (e.g. structures and relationships and changes therein) not quantitative (i.e. not just statistical regularities or functional relationships) and are mostly concerned with what *can* happen or can be done, rather than with predictive laws or correlations.

---

[6] That we cannot *express* such experiential contents in language does not preclude us from using language (or some other tool) to refer to or specify such contents; see Chrisley (1995).

The language of physics (mainly equations) is not as well suited to describing these realms of possibility as the languages of logic, discrete mathematics, formal linguistics (grammars of various kinds) and the languages of computer scientists, software engineers and AI theorists. The latter are languages for specifying and explaining the behaviour of information processing machines. We do not claim that *existing* specialist languages and ontologies are sufficient for describing and explaining mental phenomena, though they add useful extensions to pre-theoretic languages.

5. *Virtual machines and consciousness*

It is widely accepted that biological organisms use information about the environment in selecting actions. Often information about their own state is also used: deciding whether to move towards visible food or visible water on the basis of current needs. Consider this bolder claim:

> **Basic working assumption:** The phenomena labelled 'conscious' involve no magic; they result from the operation of very complex biological information processing machines which we do not yet understand.

Although the first part is uncontentious to anyone of a naturalist bent, the latter half is, of course, notoriously controversial. The rest of this paper attempts to defend it, though a full justification requires further research of the sort we propose.

All biological information processing is based on physical mechanisms. That does not imply that information processing states and processes are physical states and processes in the sense of being best described using the concepts of the physical sciences (physics, chemistry, astronomy, etc.). Many things that are produced by or realised in physical resources are non-physical in this sense, e.g. poverty, legal obligations, war, etc. So we can expect all forms of consciousness to be *based on*, or *realised in*, physical mechanisms, but not necessarily to be physical in the sense of being *describable* in the language of the physical sciences.

One way to make progress is to complement research on physical and physiological mechanisms by (temporarily) ignoring many of the physical differences between systems and focus on higher level, more abstract commonalities. For that we need to talk about what software engineers would call the virtual information processing machines *implemented in* those physical machines. Philosophers are more likely to say the former are *supervenient on* the latter (Kim, 1998): both have a partial, incomplete, view of the same relation. It is important that, despite the terminology, virtual machines are *real* machines, insofar as they can affect and be affected by the physical environment. Decision-making in a virtual machine can be used to control a chemical factory for example.

There are many families of concepts, or 'levels' on which we can think about reality (see figure 3 for some examples). At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions (e.g. poverty can cause crime). But we are not advocating a
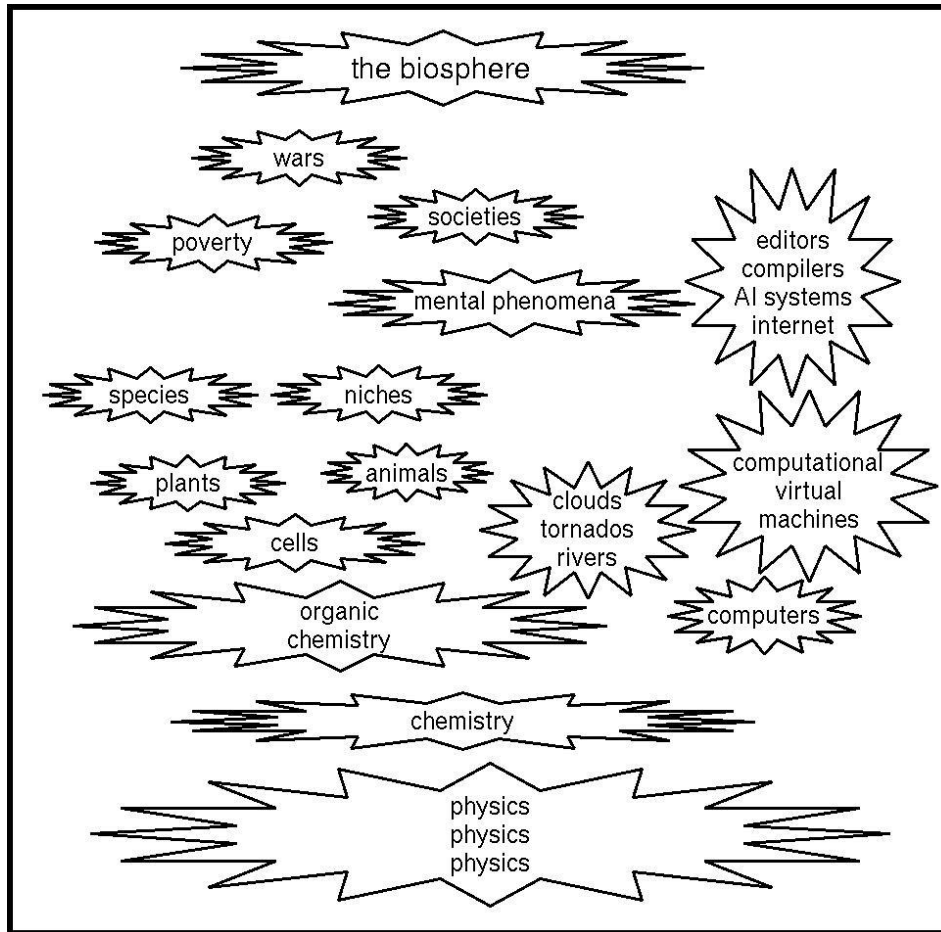
Figure 3. Various levels of reality, most of which are non-physical

'promiscuous' pluralism; the world is one in at least this sense: all non-physical levels are ultimately implemented in physical mechanisms, even if they are not definable in the language of physics. The history of human thought and culture shows not only that we are able to make good use of ontologies that are not definable in terms of those of the physical sciences, but that we cannot cope without doing so. Moreover, nobody knows how many levels of virtual machines physicists themselves will eventually discover.

So when we talk about information processing virtual machines, this is no more mysterious than our commonplace thinking about social, economic, and political states and processes and causal interactions between them.

III: The Concept of an 'Information Processor'

Successful organisms are information processors. This is because organisms, unlike rocks, mountains, planets and galaxies, typically require action to survive, and actions must be selected and initiated under certain conditions. The

conditions do not directly trigger the actions (as kicking a ball triggers its motion): rather organisms have to initiate actions using *internal* energy. Therefore appropriate and timely selection and initiation of action requires, at a minimum, information about whether the suitability conditions obtain.

1. What is information?

Like many biologists, software engineers, news reporters, propaganda agencies and social scientists, we use 'information' not in the technical sense of Shannon and Weaver, but in the sense in which information can be true or false, or can more or less accurately fit some situation, and in which one item of information can be inconsistent with another, or can be derived from another, or may be more general or more specific than another. This is *semantic* information, involving reference to something. None of this implies that the information is expressed or encoded in any particular form, such as sentences or pictures or bit-patterns, neural states, or that it is necessarily communicated between organisms, as opposed to being acquired or used within an organism.[7]

2. We don't need to define our terms

It is important to resist premature demands for a strict definition of 'information'. Compare 'energy' — that concept has grown much since the time of Newton, and now covers forms of energy beyond his dreams. Did he understand what energy is? Yes, but only partly. Instead of defining 'information' we need to analyse the following:

- the variety of *types* of information there are,
- the kinds of *forms* they can take,
- the kinds of *relations* that can hold between information items,
- the means of *acquiring* information,
- the means of *manipulating* information,
- the means of *storing* information,
- the means of *communicating* information,
- the *purposes* for which information can be used,
- the variety of *ways of using* information.

As we learn more about such things, our concept of 'information' grows deeper and richer, just as our concept of 'energy' grew deeper and richer when we learnt about radiant energy, chemical energy and mass energy. It is a part of our 'interactive conceptual refinement' methodology that although we start with tentative 'implicit' definitions, we expect future developments (including, but not limited to, new empirical data) to compel us to revise them. Neither theory

[7]   We have no space to rebut the argument in Rose (1993) that only computers, not animals or brains, are information processors, and the 'opposite' argument of Maturana and Varela summarised in Boden (2000) according to which *only* humans process information, namely when they communicate via external messages. Further discussion on this topic can be found at
http://www.cs.bham.ac.uk/research/cogaff/

nor data has the final say; rather, both are held in a constructive dialectical opposition (Chrisley, 2000).

Like many deep concepts in science, 'information' is *implicitly* defined by its role in our theories and our designs for working systems. To illustrate this point, we offer some examples[8] of processes involving information in organisms or machines:

- external or internal actions triggered by information,
- segmenting, clustering labelling components within a structure (i.e. parsing),
- trying to derive new information from old (e.g. what caused this? what else is there? what might happen next? can I benefit from this?),
- storing information for future use (and possibly modifying it later),
- considering and comparing alternative plans, descriptions or explanations,
- interpreting information as instructions and obeying them, e.g. carrying out a plan,
- observing the above processes and deriving new information thereby (self-monitoring, self-evaluation, meta-management),
- communicating information to others (or to oneself later),
- checking information for consistency.

Although most of these processes do not involve *self*-consciousness, they do involve a kind of awareness, or sentience of something, in the sense of having, or being able to acquire, information about something. Even a housefly has that minimal sort of consciousness. Later we analyse richer varieties of consciousness.

3. *Varieties of information contents*

Depending on its needs and its capabilities, an organism may use information about:

- density gradients of nutrients in the primaeval soup,
- the presence of noxious entities,
- where the gap is in a barrier,
- precise locations of branches in a tree as it flies through them,
- how much of its nest it has built so far,
- which part of the nest should be extended next,
- where a potential mate is,
- something that might eat it,
- something it might eat,
- whether that overhanging rock is likely to fall,
- whether another organism is likely to attack or run,
- how to achieve or avoid various states,
- how it thought about that last problem,
- whether its thinking is making progress.

---

[8] This list, and the other lists we present, are meant only to be illustrative, not exhaustive.

Information contents can vary in several dimensions, for instance whether the information is localised (seeing the colour of a dot) or more wholistic (seeing a tree waving in the breeze), whether it involves only geometric and physical properties (seeing a blue cube) or more abstract properties (seeing a route through shrubbery, seeing someone as angry), whether it refers to something that cannot be directly experienced (electrons, genes, cosmic radiation), whether it refers to something which itself refers to or depicts something else, and so on. Other dimensions of variation include the structure, whether the information involves use of concepts, and objectivity of what is referred to and how it is referred to. (Discussed below in connection with qualia.)

4. Information processing and architecture

What an organism or machine can do with information depends on its architecture. An architecture includes, among other things:

- forms of representation, i.e. ways of storing and manipulating information (Peterson, 1996),
- algorithms,
- concurrently active sub-systems, with different functional roles,
- connections between and causal interactions between sub-systems.

Some architectures *develop* i.e. they change themselves over time so that the components and links within the architecture change. A child's mind and a multi-user, multi-purpose computer operating system in which new interacting processes are spawned or old ones killed, are both examples of changing architectures.

5. What is a machine?

We understand a machine to be a complex whole made of interacting components whose capabilities combine to enable the whole machine to do things. Some of the components may themselves be machines in this sense. There are at least three types of machines:

- Matter manipulating machines: *Diggers, drills, cranes, cookers . . .*
- Energy manipulating machines: *Diggers, drills, cranes, cookers, transformers, steam engines . . .*
- Information manipulating machines: *Thermostats, controllers, most organisms, operating systems, compilers, business organisations, governments . . .*

6. Information processing virtual machines

We are concerned with the third class, the information processing machines. Information-manipulation is not restricted to physical machines, e.g. made of blood, meat, wires, transistors, etc. *Virtual* machines (VM) can also do it. These contain *abstract* non-physical entities, like words, sentences, numbers,

bit-patterns, trees, procedures, rules, etc., and the causal laws that summarise their operation are not the same as the laws of the physical sciences.

It may be true of a chess virtual machine that whenever it detects that its king is threatened it attempts to make a defensive move or a counter-attack, but that is not a law of physics. Without changing any laws of physics, the virtual machine can be altered to play in 'teaching mode' for beginners, so that the generalisation no longer holds. The predictability of a chess virtual machine depends in a complicated way on the fact that the components in which it is implemented obey the laws of physics, though the very same sort of virtual machine could be implemented in different components with different behaviours (e.g. in valves instead of transistors).

The laws that govern the VM are not derivable by pure mathematics or pure logic from a physical description of the components and the physical laws governing their behaviour. 'Bridging laws' relating the states and processes in the virtual machine and those in the physical machine are needed. These cannot be proved by logic alone because the concepts required to define a chess virtual machine (e.g. 'queen', 'check', 'win', 'capture') are not *explicitly definable* in terms of those of physics. Neither are bridging laws empirical, since when an implementation is understood, the connection is seen to be necessary because of subtle and complex structural relations between the physical machine and the virtual machine, despite the different ontologies involved.[9]

7. Evolution of information processing architectures

Exploring the full range of designs for behaving systems requires knowledge of a wide range of techniques for constructing virtual machines of various sorts. Many clues come from living things, since evolution 'discovered' and used myriad mechanisms long before human engineers and scientists existed. The clues are not obvious, however: paleontology shows development of physiology and provides weak evidence about behavioural capabilities, but virtual machines leave no fossils. Surviving systems give clues, however. Some information processing capabilities (e.g. most of those in microbes and insects) are evolutionarily very old, others relatively new (e.g. the ability to learn to read, design machinery, do mathematics, or think about your thought processes.) The latter occur in relatively few species. Perceptual mechanisms that evolved at different times provide very different sorts of information about the environment. An amoeba cannot acquire or use information about the state of mind of a human, though a dog can to some extent. Most organisms, including humans, contain both old and new sub-systems performing different, possibly overlapping, sometimes competing tasks. We need to understand how the new mechanisms associated with human consciousness differ from, how they are built on, and how they interact with the older mechanisms.

---

[9] These are controversial comments, discussed at greater length, though possibly not yet conclusively, in Sloman & Scheutz (2001). Scheutz (1999) argues that the existence of virtual machines with the required structure can be derived mathematically from a description of the physical machine by abstracting from details of the physical machine.
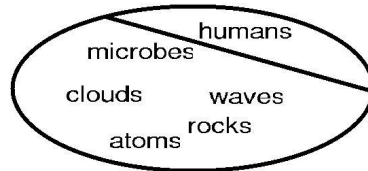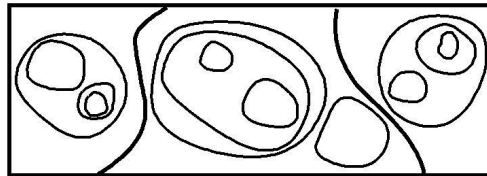
1. A dichotomy (one big division):

2. A continuum (seamless transition):

3. A space with many discontinuities:

Figure 4. Models of conceptual spaces. It is often assumed that the only alternative to a dichotomy (conscious/non-conscious) is a continuum of cases with only differences of degree. There is a third alternative.

In order to understand consciousness as a biological phenomenon, we need to understand the variety of biological information processing architectures and the states and processes they can support. There's no need to assume that a unique, correct architecture for consciousness exists; the belief that there is amounts to believing the conscious/non-conscious distinction is a dichotomy, as in figure 4. Many assume that the only alternative to a dichotomy is a continuum, in which all differences are differences of degree and all boundaries are arbitrary, or culturally determined. This ignores conceptual spaces that have many discontinuities. Examples are the space of possible designs and the space of requirements for designs, i.e. 'niche-space' (Sloman, 2000). Biological changes, being based on molecular structures, are inherently discontinuous. Many of the changes that might be made to a system (by evolution or learning or self-modification) are discontinuous; some examples are: duplicating a structure, adding a new connection between existing structures, replacing a component with another, extending a plan or adding a new control mechanism.

We don't know what sorts of evolutionary changes account for the facts that humans, unlike all (or most) other animals, can use subjunctive grammatical forms, can think about the relation between mind and body, can learn predicate calculus and modal logic, can see the structural correspondence between four rows of five dots and five rows of four dots, and so on. We don't know how many of the evolutionary changes leading to human minds occurred separately in other animals (Hauser, 2001). But it is unlikely to have been either a single massive

discontinuity, or a completely smooth progression. Below we explore some interesting discontinuities (some of which may result from smoother changes at lower levels), but first a methodological detour.

IV: Varieties of Functionalism

What we are proposing is a new kind of functionalist analysis of mental concepts. Functionalism is fairly popular among philosophers, though there are a number of standard objections to it. We claim that these objections can be avoided by basing our analyses on *virtual machine functionalism*.

1. Atomic state functionalism

Most philosophers and cognitive scientists write as if 'functionalism' were a well defined generally understood concept. E.g. Block (1996) writes (regarding a mental state S1):

> According to functionalism, the nature of a mental state is just like the nature of an automaton state: constituted by its relations to other states and to inputs and outputs. All there is to S1 is that being in it and getting a 1 input results in such and such, etc. According to functionalism, all there is to being in pain is that it disposes you to say 'ouch', wonder whether you are ill, it distracts you, etc.

This summary has (at least) two different interpretations, where the second has two sub-cases. On the first reading, which we'll call *atomic state functionalism*, an entity A can have only one, indivisible, mental state at a time, and each state is completely characterised by its role in a state-transition network in which current input to A plus current state of A determine the immediately following output of A and next state of A. This seems to be the most wide-spread interpretation of functionalism, and it is probably what Block intended, since he led up to it using examples of finite-state automata which can have only one state at a time. These could not be states like human hunger, thirst, puzzlement or anger, since these can coexist and start and stop independently.

On the second interpretation of Block's summary, A can have several coexisting, independently varying, interacting mental states. It is possible that Block did not realise that his examples, as ordinarily understood, were of this kind: for instance, the same pain can both make you wonder whether you are ill and distract you from a task, so that having a pain, wondering whether you are ill, having the desire or intention to do the task, and being distracted from it are four coexisting states which need not start and end at the same time. If the pain subsides, you may continue to wonder whether you are ill, and while working on the task (e.g. digging the garden) you might form a new intention to see a doctor later. Coexistence of interacting sub-states is a feature of how people normally view mental states, for instance when they talk about conflicting desires or attitudes.

2. Virtual machine functionalism

If a mind includes many enduring coexisting, independently varying, causally interacting, states and processes then it is a complex mechanism, though not a physical mechanism, since the parts that interact are not physical parts but things like desires, memories, percepts, beliefs, attitudes, pains, etc. This is close to the notion of virtual machine explained earlier, so we'll call this interpretation of Block's summary, allowing multiple, concurrently active, interacting mental states, *virtual machine functionalism* (VMF).

VMF allows that A has many mental sub-states at any time, and each sub-state S of A will typically depend in part on some sub-mechanism in A (e.g. perception sub-systems, action sub-systems, long term memory, short term memory, current store of goals and plans, reasoning sub-systems, etc.), though exactly which states and sub-systems can coexist in humans is an empirical question. We do not assume that functionally distinct sub-systems necessarily map onto physically separable sub-systems.

Each sub-state S is defined by its causal relationships to other sub-states and, in some cases, its causal relations to the environment. (e.g., if S is influenced by sensors or if it can influence motors or muscles). Then a particular type of sub-state S (believing something, wanting something, trying to solve a problem, enjoying something, having certain concepts, etc.) will be defined partly by the kinds of causal influences that state can receive from the environment or from other sub-states and partly by the kinds of causal influences it can have on other sub-states and the environment. The causal influences surrounding a sub-state, or its own internal processes, may cause that state to be replaced by another. For instance A's worry may cause A to remember a previous occasion which causes A to decide that the pain does not indicate illness: so the state of worry causes its own demise. The state of thinking about a problem or imagining how a tune goes will naturally progress through different stages. Each stage in the process will have different causal properties, as Ryle (1949) noted.

Thus, according to VMF, each sub-state S (or persisting process) of agent A may be characterised by a large collection of conditionals of forms like:

> *If A is in sub-state S, and simultaneously in various other sub-states, then if the sub-system of A concerned with S receives inputs I_1, I_2, ... from other sub-systems or from the environment, and if sub-states S_k, S_l, ... exist then S will cause output O_1 to the sub-system concerned with state S_m and output O_2 to the sub-system concerned with state S_n (and possibly other outputs to the environment), and will cause itself to be replaced by state S_2.*

In some cases the causal interactions may be probabilistic rather than determinate, e.g. if part of the context that determines effects of a state consists of random noise or perturbations in lower levels of the system.

Our description is consistent with many different types of causal interactions and changes of state. Some states are representable by numerical values, whereas others involve changing structures (e.g. construction of a sentence or plan). Some causal interactions simply involve initiation, termination, excitation or

inhibition, whereas others are far more complex, e.g. transmission of structured information from one sub-system to another.

So, according to VMF, a functional state S of A depends on sub-systems of A whose states can change, and S is defined in terms of causal connections between inputs and outputs of those sub-systems in the context of various possible combinations of states of other sub-systems of A, i.e. other concurrently active functional states, as well as causal connections with S's own changing sub-states.

3. VMF and architectures

This kind of functionalist analysis of mental states is consistent with the recent tendency in AI to replace discussion of mere *algorithms* with discussion of *architectures* in which several co-existing sub-systems can interact, perhaps running different algorithms at the same time (e.g. Minsky, 1987; Brooks, 1986; Sloman, 1993). Likewise, many software engineers design, implement, and maintain virtual machines that have many concurrently active sub-systems with independently varying sub-states. A running operating system like Solaris or Linux is a virtual machine that typically has many concurrently active components. New components can be created, and old ones may die, or be duplicated. The internet is another, more complex, example. It does not appear that Block, or most philosophers who discuss functionalism, take explicit account of the possibility of virtual machine functionalism of the sort described here, even though most software engineers would find it obvious.

There are two interesting variants of VMF, restricted and unrestricted. *Restricted virtual machine functionalism* requires that every sub-state be causally connected, possibly indirectly, to inputs and outputs of the whole system A, whereas *unrestricted* VMF does not require this.[10] A philosophical view very close to restricted VMF was put forward in Ryle (1949, e.g. in the chapter on imagination), though he was widely misinterpreted as supporting a form of behaviourism.

4. Unrestricted virtual machine functionalism is biologically possible

Unrestricted VMF allows that some sub-state S or continuing sub-process is constantly altered by other sub-states that are connected with the environment even though none of the changes that occur in S affect anything that can affect the environment. An example in a computer might be some process that records statistics about events occurring in other processes, but does not transmit any of its records to any other part of the system, and no other part can access them. Unrestricted VMF even allows sub-systems that beaver away indefinitely without getting any inputs from other parts of the system. For instance, a sub-system might be forever playing games of chess with itself, or forever searching for proofs of Goldbach's conjecture. Thus unrestricted VMF neither requires sensors to be

---

[10] This is also not required by atomic state functionalism as normally conceived, since a finite state machine can, in principle, get into a catatonic state in which it merely cycles round various states forever, without producing any visible behaviour, no matter what inputs it gets.

capable of influencing every internal state of A, nor requires every state and process of A to be capable of affecting the external behaviour of the whole system A.

An interesting intermediate case is also allowed by VMF, namely a process that causally interacts with other processes, e.g. by sending them instructions or answers to questions, but whose internal details do not influence other processes, e.g. if conclusions of reasoning are transmitted, but none of the reasons justifying those conclusions. Then other parts of the system may know what was inferred, but be totally unable to find out why.

After a causally disconnected process starts up in A, it may be possible for new causal links to be created between it and other processes that have links to external transducers; but that is not *a precondition* for its existence. Likewise, a process that starts off with such links might later become detached (as sometimes happens to 'runaway' processes in computers).

It is often supposed that biological information processing systems must necessarily conform to *restricted* VMF, because unless some internal state or process produces external behavioural consequences under some conditions it could not possibly be selected for by evolution. However, this ignores the possibility of evolutionary changes that have side-effects apart from the effects for which they are selected, and also the possibility of evolutionary changes that have no benefits at all, but are not selected out because the environment is so rich in resources. There are some genes with positively harmful effects that continue indefinitely, such as the genes for sickle cell anaemia, which also happen to provide some protection against malaria.

We should at least consider the possibility that at a certain stage in the evolutionary history of an organism a genetic change that produces some biologically useful new information processing capability also, as a side-effect, creates a portion of a virtual machine that either runs without being influenced by or influencing other behaviourally relevant processes, or more plausibly is influenced by other states and processes but has no externally observable effects, except barely measurable increases in energy consumption. Not only is this possible, it might even be biologically useful to later generations, for instance if a later genetic change combines this mechanism with others in order to produce biologically useful behaviours where the old mechanism is linked with new ones that produce useful external behaviour. Various intermediate cases are also possible, for instance complex internal processes that produce no external effect most of the time but occasionally interact with other processes to produce useful external effects. Some processes of idle thinking, without memory records, might be like that. Another intermediate case is a process that receives and transmits information but includes more complex information processing than can be deduced from any of its input-output mappings. Completely disconnected processes might be called 'causally epiphenomenal' the largely disconnected ones and 'partly epiphenomenal'.

An interesting special case is a sub-system in which some virtual machine processes monitor, categorise, and evaluate other processes within A. This internal self-observation process might have no causal links to external motors, so that its

information cannot be externally reported. If it also modifies the processes it observes, like the meta-management sub-systems described later, then it may have external effects. However it could be the case that the internal monitoring states are too complex and change too rapidly to be fully reflected in any externally detectable behaviour: a bandwidth limitation. For such a system experience might be partly ineffable.

5. Detecting disconnected virtual machine states

If everything is running on a multi-processing computer with a single central processing unit (CPU), then detached processes will have some externally visible effect because they consume energy and slow down external responses, though in some cases the effects could be minute, and even if detected, may not provide information about what the detached process is doing. If a detached process D slows others down, D is not *totally* causally disconnected. However this is a purely quantitative effect on speed and energy: the other processes need not be influenced by any of the structural details or the semantic content of the information processing in D. Where the whole system is implemented on large numbers of concurrently operating physical processors with some processor redundancy it may not be possible to tell even that the detached process is running if it slows nothing down, though delicate measurements might indicate an 'unknown' consumer of energy.

In principle it might be possible to infer processing details in such a detached virtual machine by examining states and processes in physical brain mechanisms or in the digital circuitry of a computer, but that would require 'decompiling' — which might be too difficult in principle, as people with experience of debugging operating systems will know. (E.g. searching for a suitable high level interpretation of observed physical traces might require more time than the history of the universe.)

There is a general point about the difficulty of inferring the nature of virtual machine processes from observations of low level physical traces, even when the processes do influence external behaviour. It could be the case that the best description of what the virtual machine is doing uses concepts that we have never dreamed of, for instance if the virtual machine is exploring a kind of mathematics or a kind of physical theory that humans will not invent for centuries to come. Something akin to this may also be a requirement for making sense of virtual machines running in brains of other animals, or brains of newborn human infants.

Despite the difficulties in testing for such decoupled virtual machine processes, a software engineer who has designed and implemented the system may know that those virtual components exist and are running, because the mechanisms used for compiling and running a program are trusted. (A very clever optimising compiler might detect and remove code whose running cannot affect output. But such optimisation can usually be turned off.)

Virtual machine functionalism as defined here permits possibilities that are inconsistent with conventional atomic state functionalism's presumption that

only one (indivisible) state can exist at a time. Likewise VMF goes beyond simple dynamical systems theories which, as noted in Sloman (1993), talk about only one state and its trajectory. A dynamical system that has multiple coexisting, changing, interacting, attractors (like a computer) might be rich enough to support the VM architectures allowed by unrestricted VMF. Any conceptual framework that postulates only atomic (indivisible) global states will not do justice to the complexity either of current computing systems or of biological information-processing systems. We'll see below that it is also inconsistent with the phenomena referred to in talk of qualia.

6. Some VMs are harder to implement than others

The kinds of collections of concurrent interacting processes simultaneously satisfying large collections of conditional descriptions (most of which will be counterfactual conditionals at any time) will be far more difficult to implement than a process defined entirely by a sequential algorithm which always has a single locus of control. For example, Searle's thought experiment (Searle, 1980) in which he simulates a single algorithm in order to give the appearance of understanding Chinese is believable because we imagine him progressively going through the steps, with a finger keeping track of the current instruction.

It would be far more difficult for him to simultaneously simulate a large collection of interacting processes concerned with memory, perception, decision making, goal formation, self-monitoring, self-evaluation, etc., each running at its own (possibly varying) speed, and all involved in causal relationships that require substantial collections of counter-factual conditional statements to be true of all the simulated states and processes.

In other words, although virtual machine functionalism, like other forms of functionalism, allows VM states to be multiply realisable, the constraints on possible realisations are much stronger. For very complex virtual machines there may be relatively few ways of implementing them *properly* in our physical world, since many implementations including Searlian implementations, would not maintain all the required causal linkages and true counter-factual conditionals.

V: Evolvable Architectures

Different sorts of information processing systems are required for organisms with different bodies, with different needs, with different environments — and therefore different niches. Since these are virtual machines, their architectures cannot easily be inspected or read off brain structures. So any theory about them is necessarily at least partly conjectural. However, it seems that the vast majority of organisms have purely reactive architectures. A tiny subset also have deliberative capabilities, though a larger group have reactive precursors to deliberation, namely the ability to have two or more options for action activated simultaneously with some selection or arbitration mechanism to select one ('proto-deliberation'). An even smaller subset of animals seem to have meta-

management capabilities (described below). These different architectural components support different varieties of what people seem to mean by 'consciousness'. Moreover, as we have seen, VMF even allows processes that satisfy the intuition that some mental states (and qualia?) have no necessary connection with perception or action.

1. Reactive architectures

Reactive architectures are the oldest type. A reactive mechanism (figure 5) is one that produces outputs or makes internal changes, perhaps triggered by its inputs and/or its internal state changes, but without doing anything that can be understood as explicitly representing and comparing alternatives, e.g. deliberating about explicitly represented future possibilities.

Many alternative reactive architectures are possible: some discrete and some continuous or mixed; some with and some without internal state changes; some with and some without adaptation or learning (e.g. weight changes in neural nets); some sequential and some with multiple concurrent processes; some with global 'alarm' mechanisms (figure 7), and some without. Some reactions produce external behaviour, while others merely produce internal changes. Internal reactions may form loops. Teleo-reactive systems (Nilsson, 1994) can execute stored plans. An adaptive system with only reactive mechanisms can be a very successful biological machine.

Some purely reactive species have a social architecture enabling large numbers of purely reactive individuals to give the appearance of considerable intelligence (e.g. termites building 'cathedrals'). The main feature of reactive systems is that they lack the core ability of deliberative systems (explained below); namely, the ability to represent and reason about non-existent or unperceived phenomena (e.g., future possible actions or hidden objects). However, we have yet to explore fully the space of intermediate designs (Scheutz & Sloman, 2001).
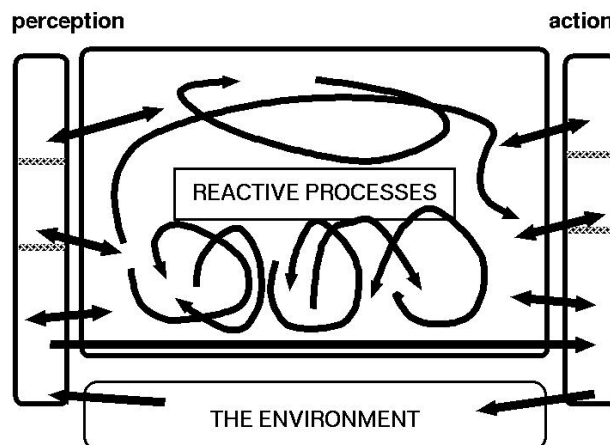

Figure 5. A simple, insect-like architecture. Arrows indicate direction of information flow. Some reactions produce internal changes that can trigger or modulate further changes. Perceptual and action mechanisms may operate at different levels of abstraction, using the same sensors and motors.

In principle a reactive system can produce any external behaviour that more sophisticated systems can produce. However, to do so in practice it might require a larger memory for pre-stored reactive behaviours than could fit into the whole universe. Moreover, the evolutionary history of any species is necessarily limited, and reactive systems can use only strategies previously selected by evolution (or by a design process in the case of artificial reactive systems). A trainable reactive individual might be given new strategies, but it could not produce and evaluate a novel strategy *in advance* of ever acting on it, as deliberative systems with planning capabilities can. Likewise a reactive system could not discover the undesirability of a fatal strategy without acting on it. As Craik (1943) and others have noted, this limitation can be overcome by deliberative capabilities, but evolution got there first.

2. Consciousness in reactive systems

What about 'consciousness' in reactive organisms? Is a fly conscious of the hand swooping down to kill it? Insects perceive things in their environment, and behave accordingly. However, it is not clear whether their perceptual mechanisms produce information states between perception and action usable in different ways in combination with different sorts of information, as required for human-like consciousness. Purely reactive systems do not use information with the same type of flexibility as deliberative systems, which can consider non-existent possibilities. They also lack the architectural sophistication required for self-awareness, self-categorising abilities. A fly that sees an approaching hand probably does not know that it sees — it lacks meta-management mechanisms, described later. So the variety of conscious awareness that a fly has is very different from the kinds of awareness we have by virtue of our abilities to recombine and process sensory information, our deliberative capabilities, and our capacity for reflection. This more elaborate answer, rather than a simple 'yes' or 'no', is the best reply to the question 'is a fly conscious?'

In a reactive system, sensory inputs normally directly drive action-control signals, though possibly after transformations which may reduce dimensionality, as in simple feed-forward neural nets. There are exceptions: e.g. bees get information which can be used either to control their own behaviour or to generate 'messages' later on that influence the behaviour of others. We could define that as a special kind of consciousness.

3. Pressures for deliberative mechanisms

Sometimes planning is useful; in such cases, an architecture such as that depicted in figure 6, containing mechanisms for exploring hypothetical possibilities, as postulated by Craik, is advantageous. This could result from an evolutionary step in which some reactive components are first duplicated then later given new functions (Maynard Smith & Szathmáry, 1999).

Deliberative mechanisms include the ability to represent possibilities (e.g. possible actions, possible explanations for what is perceived) in some explicit
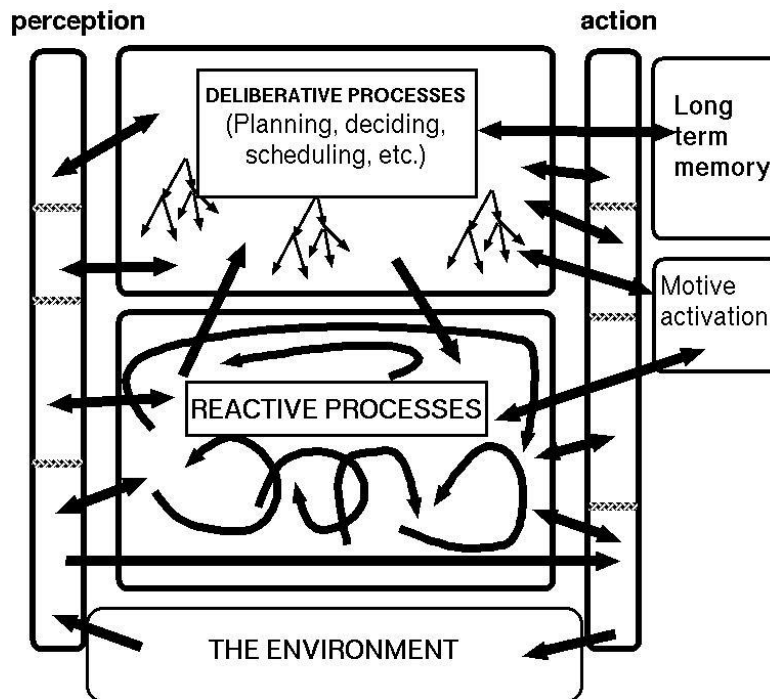
Figure 6. Reactive-deliberative architecture with 'multi-window' perception and action. Higher level perceptual and motor systems (e.g. parsers, command-interpreters) may have 'direct' connections with higher level central mechanisms.

form, enabling alternative possibilities to be compared and one selected. Purely deliberative architectures were employed in many traditional AI systems including Winograd's SHRDLU (Winograd, 1972). Other examples include theorem provers, planners, programs for playing board games, natural language systems, and expert systems of various sorts. In robots moving at speed, deliberative mechanisms do not suffice: they need to be combined with reactive mechanisms, e.g. for dealing with sudden dangers, such as 'alarm' mechanisms.

Deliberative mechanisms can differ in various ways, e.g.:

- the forms of representations used (e.g. logical, pictorial, activation vectors — some with and some without compositional semantics);
- whether they use external representations, as in trail-blazing or keeping a diary;
- the algorithms/mechanisms available for manipulating representations;
- the number of possibilities that can be represented simultaneously (working memory capacity);
- the depth of 'look-ahead' in planning;
- the syntactic depth of descriptions and plans;
- the ability to represent future, past, concealed, or remote present objects or events;
- the ability to represent possible actions of other agents;

- the ability to represent mental states of others (linked to meta-management, below);
- the ability to represent abstract entities (numbers, rules, proofs);
- the ability to learn, in various ways;
- the variety of perceptual mechanisms (see below).

Various forms of learning are possible. Some deliberative systems can learn, and use, new abstract associations, e.g. between situations and possible actions, and between actions and possible effects. In a hybrid reactive-deliberative architecture, the deliberative part may be unable to act directly on the reactive part, but may be able to *train* it through repeated performances.

The kinds of information processing available in deliberative mechanisms can be used to define kinds of consciousness which merely reactive systems cannot have, including, for instance, 'awareness of what can happen', 'awareness of danger avoided'. The perception of possibilities and constraints on possibilities (affordances [Gibson, 1986]) is something that has not yet been adequately characterised or explained (Sloman 1989; 1996). Hybrid reactive-deliberative systems can have more varieties of consciousness, though the kinds in different parts of the architecture need not be integrated (as shown by some kinds of brain damage in humans.)

4. Pressures for multi-window perception and action

Deliberative capabilities provide the opportunity for abstract perceptual and action mechanisms that facilitate deliberation and action to evolve. New *levels of perceptual abstraction* (e.g. perceiving object types, abstract affordances), and support for *high-level motor commands* (e.g. 'walk to tree', 'grasp berry') might evolve to meet deliberative needs — hence the perception and action towers in figure 6. If multiple levels and types of perceptual processing go on in parallel, we can talk about 'multi-window perception', as opposed to 'peephole' perception. Likewise, in an architecture there can be 'multi-window action' or merely 'peephole action'. Later we'll extend this idea in connection with the third, meta-management, layer. Few current AI architectures include such multi-window mechanisms, though future machines will need them in order to have human-like consciousness of more abstract aspects of the environment.

5. Pressures for self-knowledge, self-evaluation and self-control

A deliberative system can easily get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem (one of the causes of stupidity in early symbolic AI programs with sophisticated reasoning mechanisms). One way to prevent this is to have a parallel sub-system monitoring and evaluating the deliberative processes. If it detects something bad happening, then it may be able to interrupt and re-direct the processing. We call this *meta-management* following Beaudoin (1994). (Compare Minsky on 'B brains' and 'C brains' in Minsky, 1987.) It is sometimes called 'reflection' by others though with slightly different
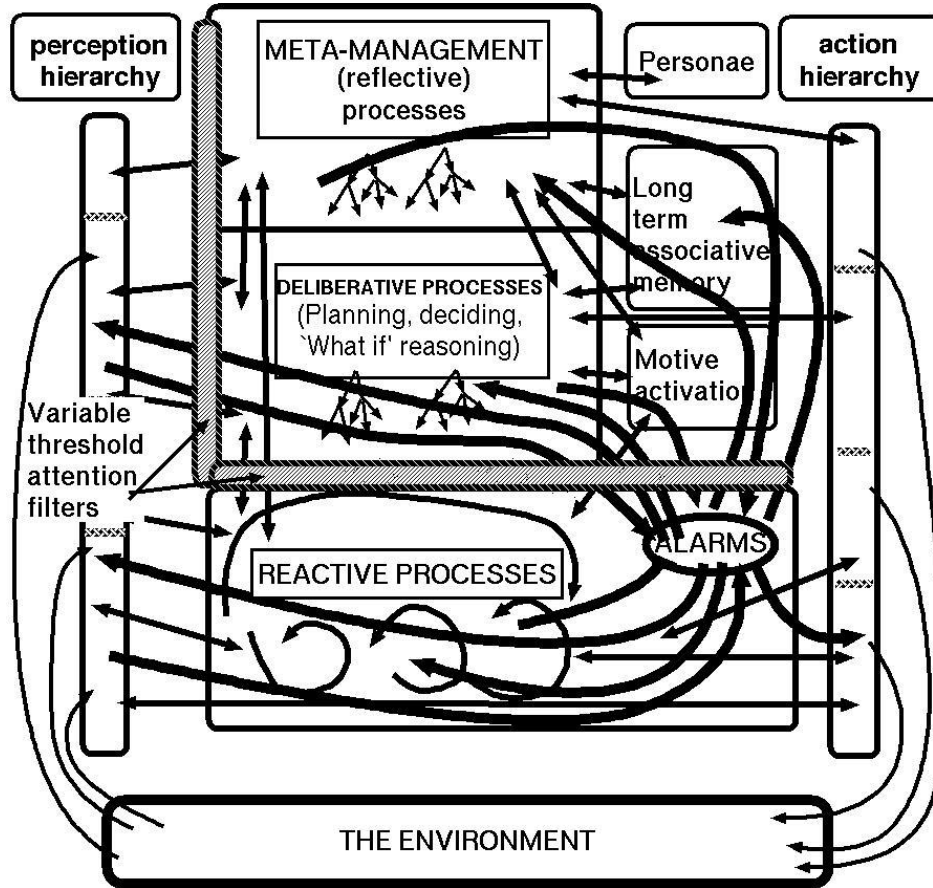
Figure 7. The H-CogAff architecture. The additional central layer supports yet more layers in perception and action hierarchies. Not all possible links between boxes are shown. Meta-management may be able to inspect intermediate states in perceptual layers.

connotations. It seems to be rare in biological organisms and probably evolved very late. This could have resulted from duplication and then diversification of alarm mechanisms, depicted crudely in figure 7. As with deliberative and reactive mechanisms, there are many varieties of meta-management. An interesting early example in AI is described in Sussman (1975). Psychological research on 'executive functions' (e.g. Barkley, 1997) presupposes something like meta-management, often not clearly distinguished from deliberation.

Self monitoring can include categorisation, evaluation, and (partial) control of internal processes, not just measurement. The richest versions of this evolved very recently, and may be restricted to humans, though there are certain kinds of self-awareness in other primates (Hauser, 2001).

Absence of meta-management can lead to 'stupid' reasoning and decision making both in AI systems, and in brain-damaged or immature humans, though

this may sometimes be mis-diagnosed as due to lack of emotional mechanisms, as in Damasio (1994). Both the weaknesses of early AI programs with powerful deliberative capabilities and some effects of brain damage in humans that leave 'intelligence' as measured in IQ tests intact, indicate the need for a distinction between deliberative and meta-management mechanisms. Both will be needed in machines with human-like consciousness.

6. Access to intermediate perceptual data

In addition to monitoring of central problem-solving and planning processes there could be monitoring of intermediate stages in perceptual processes or action processes, requiring additional arrows going from within the perception and action towers to the top layer in figure 7. Examples would be the ability to attend to fine details of one's perceptual experience instead of only noticing things perceived in the environment; and the ability to attend to fine details of actions one is performing, such as using proprioceptive information to attend to when exactly one bends or straightens one's knees while walking. The former ability is useful in learning to draw pictures of things, and latter helps the development of various motor skills, for instance noticing which ways of performing actions tend to be stressful and therefore avoiding them — a problem familiar to many athletes and musicians. All of these processes, consistent with VFM, would need to be replicated in a machine with human-like consciousness.

7. Yet more perceptual and motor 'windows'

We conjecture that, as indicated in figure 7, central meta-management led to opportunities for evolution of additional layers in 'multi-window' perceptual and action systems: e.g. social perception (seeing someone as sad or happy or puzzled), and stylised social action (e.g. courtly bows, social modulation of speech production). This would be analogous to genetically (and developmentally) determined architectural mechanisms for multi-level perception of speech, with dedicated mechanisms for phonological, morphological, syntactic and semantic processing.

In particular, the categories that an agent's meta-management system finds useful for describing and evaluating its own mental states might also be useful when applied to others. (The reverse can also occur.) In summary:

> **Other knowledge from self-knowledge:** The representational capabilities that evolved for dealing with self-categorisation can also be used for other-categorisation, and vice-versa. Perceptual mechanisms may have evolved recently to use these representational capabilities in percepts. Example: seeing someone else as happy, or angry, seeing the duck-rabbit in figure 1 as looking left or right.

Additional requirements for coping with a fast moving environment and multiple motives (Beaudoin, 1994) and for fitting into a society of cognitively capable agents, provide evolutionary pressure for further complexity in the architecture, e.g.:

- 'interrupt filters' for resource-limited attention mechanisms,
- more or less global 'alarm mechanisms' for dealing with important and urgent problems and opportunities, when there is no time to deliberate about them,
- a socially influenced store of personalities/personae, i.e. modes of control in the higher levels of the system.

These are indicated in the figure 7, with extended (multi-window) layers of perception and action, along with global alarm mechanisms. Like all the architectures discussed so far, this conjectured architecture, (which we call 'H-CogAff', for '**h**uman-like architecture for **cog**nition and **aff**ect) could be realised in robots (in the distant future).

8. Other minds and 'philosophical' genes

If we are correct about later evolutionary developments providing high level conceptual, perceptual and meta-management mechanisms that are used both for *self*-categorisation and *other*-categorisation (using 'multi-window perception' as in the duck-rabbit picture, and in perception of attentiveness, puzzlement, joy, surprise, etc., in others), then instead of a new-born infant having to work out by some philosophical process of inductive or analogical reasoning or theory construction that there are individuals with minds in the environment, it may be provided genetically with mechanisms designed to use mental concepts in perceiving and thinking about others. This can be useful for predator species, prey species and organisms that interact socially. Such mechanisms, like the innate mechanisms for perceiving and reasoning about the physical environment, might have a 'boot-strapping' component that fills in details during individual development (see section VIII.2).

Insofar as those mental concepts, like the self-categorising concepts, refer to types of *internal* states and processes, defined in terms of aspects of the architecture that produce them, they will be architecture-based, in the sense defined in section VII, even though the implicitly presupposed architectures are probably much simpler than the actual virtual machine architectures. In other words, even if some animals (including humans!) with meta-management naturally use architecture-based concepts they are likely to be over-simplified concepts in part because they presuppose over-simplified architectures.

Nevertheless, evolution apparently solved the 'other minds problem' before anyone formulated it, both by providing built-in apparatus for conceptualising mental states in others at least within intelligent prey species, predator species and social species, and also by 'justifying' the choice through the process of natural selection, which tends to produce good designs. (Later we'll describe concepts used only to refer to the system's own internal states, in discussing qualia.)

VI: Some Implications

We have specified in outline an architectural framework in which to place all these diverse capabilities, as part of the task of exploring design space. Later we can modify the framework as we discover limitations and possible developments both for the purposes of engineering design and for explanation of empirical phenomena. The framework should simultaneously help us understand the evolutionary process and the results of evolution.

Within this framework we can explain (or predict) many phenomena, some of which are part of everyday experience and some which have been discovered by scientists:

- Several varieties of *emotions*: at least three distinct types related to the three architectural layers, namely, *primary* (exclusively reactive, such as anger), *secondary* (partly deliberative, such as frustration) and *tertiary* emotions (including disruption of meta-management, as in grief or jealousy). Some of these may be shared with other animals, some unique to humans (Wright *et al*., 1996; Sloman & Logan, 2000; Sloman, 2001a).

- Different visual pathways, since there are many routes for visual information to be used (Sloman 1989; 1993; Goodale & Milner, 1992). The common claim that some provide 'what' information and other pathways 'where' information, may turn out to be misleading if the former provides information able to be used in deliberation (and communication) and the other provides information for control of reactive behaviours, e.g. reactive feedback loops controlling grasping or posture. There are probably many more visual pathways to be investigated.

- The phenomena discussed by psychologists in connection with 'executive functions', including frontal-lobe damage and attention disorders (e.g. Damasio, 1994; Barkley, 1997) — meta-management capabilities can be impaired while other things, including deliberative capabilities are intact.

- Many varieties of learning and development. For example, 'skill compilation' when repeated actions at deliberative levels train reactive systems to produce fast fluent actions, and action sequences. This requires spare capacity in reactive mechanisms. Information also flows in the reverse direction as new deliberative knowledge is derived from observation of one's own reactive behaviours, like a violinist discovering that changing the elevation of the elbow of the bowing arm is useful for switching between violin strings and changing the angle of the elbow moves the bow on one string.

- Limitations of self-knowledge. The model does not entail that self monitoring is perfect: so elaborations of the model might be used to predict ways in which self-awareness is incomplete or inaccurate. A familiar example is our inability to see our own visual blind-spots. There may be many forms of self-delusion, including incorrect introspection of what we do or do not know, of how we do things (e.g. how we understand sentences), of what processes influence our decisions, and of when things happen. Experiments

on change-blindness (O'Regan *et al.*, 1999), like many other psychological experiments, assume that people can report what they see. From our point of view they may be able to report only on the seeing processes as they are monitored by the meta-management system. But that need not be an accurate account of *everything* that is seen, e.g. in parts of the reactive layer. This could be the explanation of 'blindsight' (Weiskrantz, 1997): damage to some meta-management access routes prevents self-knowledge about intact (e.g. reactive) visual processes.

- The nature of abstract reasoning. The distinctions provided by the architecture allow us to make a conjecture that can be investigated empirically: mathematical development depends on development of meta-management — the ability to attend to and reflect on thought processes and their structure, e.g. noticing features of your own counting operations, or features of your visual processes.
- Evolvability: further work may help us understand some of the evolutionary trade-offs in developing these systems. (Deliberative and meta-management mechanisms can be very expensive, and require a food pyramid to support them).
- The discovery by philosophers of sensory 'qualia' (see section VIII.1).

VII: Multiple Experiencers: The CogAff Architecture Schema

The multi-disciplinary view of the whole architecture of an organism or system, and the different capabilities, states, processes, causal interactions, made possible by the various components, may lead to a particular model of human information processing. But there are different architectures, with very different information processing capabilities, supporting different states and processes. So we can expect many varieties of mentality, not just one.

Thus, we consider families of architecture-based mental concepts. For each architecture we can specify a family of concepts applicable to states, processes and capabilities supported by the architecture. The use of architecture-based concepts requires an explicit or implicit theory of the architecture supporting the states, processes and capabilities. Just as theories of the architecture of matter refined and extended our concepts of kinds of stuff (periodic table of elements, and varieties of chemical compounds) and of physical and chemical processes, so can architecture-based mental concepts, as explained in Sloman (2002), extend and refine our semantically indeterminate pre-theoretical concepts, leading to clearer concepts related to the mechanisms that can produce different sorts of mental states and processes. Note that the presupposed architectural theory need not be correct or complete. As a theory about an organism's architecture is refined and extended, the architecture-based concepts relying on that theory can also be extended.

This changes the nature of much of philosophy of mind. Instead of seeking to find 'correct' conceptual analyses of familiar mental concepts that are inherently indeterminate, such as 'consciousness', and 'emotion' we explore a space of

more determinate concepts and investigate ways in which our pre-theoretical concepts related to various subsets (as the pre-theoretical concept of 'water' relates to the architecture-based concepts of H_2O and D_2O [deuterium oxide], and old concepts of chemical element relate to newer architecture-based concepts of different isotopes of an element).

New questions then supplant old ones; we can expect to replace old unanswerable questions ('Is a fly conscious?' or 'Can a foetus feel pain?') with new empirically tractable questions (e.g. 'Which of the 57 varieties of consciousness does a fly have, if any?' and 'Which types of pain can occur in an unborn foetus aged N months and in which sense of "being aware" can it be aware of them, if any?').

1. Towards an architecture schema

We have proposed the CogAff schema (Sloman & Logan, 2000; Sloman, 2001b) as a framework for thinking about a wide variety of information processing architectures, including both naturally occurring and artificial ones.

There are two familiar kinds of coarse divisions within components of information processing architectures: one captured in, e.g., Nilsson's 'triple tower' model (Nilsson, 1998), and the other in processing 'layers' (e.g. reactive, deliberative and meta-management layers). These orthogonal functional divisions can be combined in a grid, as indicated in figure 8. In such a grid, boxes indicate possible functional roles for mechanisms. Only a subset of all possible information flow routes are shown; cycles are possible within boxes, but not shown.

We call this superimposition of the tower and layer views the *CogAff architecture schema*, or 'CogAff' for short. Unlike H-CogAff (see section V.5), CogAff is a schema not an architecture: it is a sort of 'grammar' for architectures. The CogAff schema can be extended in various ways: e.g. see figure 9, showing

Figure 8. The CogAff schema: superimposing towers and layers.

Figure 9. Elaborating the CogAff schema to include reactive alarms.

alarm mechanisms added. These deal with the need for rapid reactions using fast pattern recognition based on information from many sources, internal and external, triggering wide-spread reorganisation of processing. An alarm mechanism is likely to be fast and stupid, i.e. error-prone, though it may be trainable. Such mechanisms seem to be involved in several varieties of emotions.

Different organisms, different artificial systems, may have different components of the schema, different components in the boxes, and/or different connections between components. For example, some animals, and some robots, have only the reactive layer (e.g. insects and microbes). The reactive layer can include mechanisms of varying degrees and types of sophistication — some analogue, some digital — with varying amounts of concurrency. The other two layers can also differ between species.

The schema is intended to illustrate our methodology, rather than to be a final and definitive framework for architectures. It may turn out that there are better ways of dividing up levels of functionality, or that more sub-divisions should be made — e.g. between analog and discrete reactive mechanisms, between reactive mechanisms with and without chained internal responses, between deliberative mechanisms with and without various kinds of learning, or with various kinds of formalisms; and between many sorts of specialised 'alarm' mechanisms.

2. CogAff and varieties of consciousness

Different architectures subsumed by the CogAff schema support different kinds of mental processes connected more or less closely with our normal notion of 'consciousness'. For example, all support some form of 'sentience', i.e. awareness of something in the environment, including the fly's awareness of your hand swooping down to catch it. If two perceptual pathways are affected when the fly detects motion of the hand, e.g. one relatively slow normal behavioural control pathway, and a rapid reaction pathway involving an 'alarm' mechanism, then the fly has two sorts of awareness of the hand. But that does not imply that it is aware of that awareness. Compare LeDoux (1996) on emotions in rats.

As mentioned earlier, architectural change can occur not only over evolutionary timescales, but also within an individual. Learning can introduce new architectural components, e.g. the ability to read music, the ability to write programs. Development of skill (speed and fluency) through practice can introduce new connections between modules, e.g. links from higher-level perceptual layers to specialist reactive modules, for instance in learning to read fluently, or developing sophisticated athletic skills. Highly trained skills can introduce new 'layer-crossing' pathways. For example, in vision, recognition of a category originally developed for deliberation can, after training, trigger fast reactions. So the architectural approach can accommodate the intuitive claim that the varieties of consciousness that are possible within an individual can develop over time. The same would happen in human-like machines.

3. Some sub-species of the CogAff schema

The schema supports many varieties of architectures that vary according to which of the 'boxes' contain mechanisms, what the mechanisms do and how they are connected.

An example sub-category of the CogAff schema is what we call 'Omega architectures'. Here only some of the possible routes through the system are used, forming roughly the shape of an Omega Ω, made up of a pipeline, with 'peephole' perception and action, as opposed to 'multi-window' perception and action (see section V.4). For examples see Albus (1981) and Cooper & Shallice (2000).

Another sub-species of CogAff is the subsumption architecture, containing several layers all within the reactive layer (Brooks, 1991). This sort of architecture is useful for understanding or designing certain relatively primitive sorts of organisms (e.g. insects, fish, crabs?) and robots.

By locating various architectures within a common schematic framework, we facilitate the task of comparing and contrasting the various forms of consciousness which the architectures support. Comparing designs and analysing trade-offs is a better way to proceed than arguing endlessly about which architecture is 'correct', or 'sufficient for consciousness'. Instead we replace our single cluster concept with a variety of more refined, architecture-based concepts about which more productive empirical investigations are possible. Of course, conceptual indeterminacy may never be eliminated completely, because of the possibility of unanticipated subdivisions with surprising combinations of features.

VIII: Some Objections

1. An architecture-based explanation of qualia?

At this point, some readers will be wondering how these architectural notions make any headway on the infamous problems of consciousness, e.g. the problem of qualia - the private, ineffable way things seem to us. We have planted pointers to our answer at various points in preceding sections. In section VI, we suggested that an architecture-based explanation of qualia is possible. Primarily we explain qualia by providing an explanation of *the phenomena that generate philosophical thinking of the sort found in discussions of qualia*. Note that we are not talking merely about explaining behaviour, for we have repeatedly discussed explanations of how internal, possibly externally undetectable, states and processes can occur in certain virtual machine architectures.

The concept of 'qualia' arose out of philosophical discussions of our ability to attend to aspects of internal information processing (internal self-awareness). That possibility is inherent in any system that has the H-CogAff architecture (see section V.5), though different varieties of the phenomenon will be present in different architectures, depending on the forms of representation and modes of monitoring available to meta-management. Some forms will provide the ability to attend not only to what is perceived in the environment, but to also features of the *mode of perception* that are closely related to properties of intermediate

Figure 10. Noticing two perspectives on the same object is one route to concerns about qualia.

sensory data-structures, as indicated by some of the arrows in figure 7 from the perceptual tower to the centre.

Consider perceiving a table. Most adults (though not young children) can attend not only to the table and its fixed 3-D shape, but also to the 2-D *appearance* of the table in which angles and relative lengths of lines change as you change your viewpoint (or the table is rotated; see figure 10). The appearance can also change as you squint, tap your eyeball, put on coloured spectacles, etc. This is exactly the sort of thing that led philosophers (and others) to think about qualia (previously called 'sense data') as something internal, non-physical, knowable only from inside, etc. If meta-management processes have access to intermediate perceptual states, then this can produce self-monitoring of sensory contents, leading robot philosophers with this architecture to discover 'the problem(s) of qualia'. And the same would go for *anything* which has that architecture: six robots with the H-CogAff architecture discussing various aspects of their experience of the same table seen from different viewpoints could get bogged down discussing consciousness, just like six blind philosophers.

> **What qualia are:** qualia are what humans or future human-like robots refer to when referring to the objects of internal self observation.

2. *Architecture-based and architecture-driven concepts*

We have used 'architecture-based' to describe the concepts used by us to refer ('from outside') to possible states and processes in an information processing machine, where our concepts are defined in terms of what components of the architecture can do. An individual can also use such concepts to refer to its own states, as suggested earlier. But there is another way in which an information processing system can refer to its own states, which explain some aspects of notions of qualia.

Concept formation is a huge topic, but, for now, consider the likelihood that in many organisms there are processes of concept formation which emerge from interactions between a self-organising classification system and the information fed into it. A well known example of a mechanism that can achieve this is a

Kohonen net (Kohonen, 1989). We describe as 'architecture-driven' sets of concepts created within an architecture as part of the individual history of the organism or machine. If individual A1 develops its own concepts used to describe internal states of another agent A2 on the basis of assumptions about the information processing architecture of A2, then the concepts are architecture-driven in relation to A1, and architecture-based in relation to A2, or any other system with the assumed architecture. It is possible for A1 to use architecture-driven architecture-based concepts to refer to itself. Architecture-driven concepts can refer to many different sorts of things, e.g. colours and shapes of objects in the environment, tastes, tactile qualities of objects, etc., if the concepts are developed by an organism on the basis of perceptual experience of those objects.

3. The privacy and ineffability of qualia

Now suppose that an agent A with a meta-management system, as in figure 7, uses a self-organising process to develop concepts for categorising its own internal virtual machine states as sensed by internal monitors. These will be architecture-driven concepts, but need not be architecture-based if the classification mechanism does not use an implicit or explicit theory of the architecture of the system it is monitoring, but merely develops a way of organising its 'sensory' input data. If such a concept C is applied by A to one of its internal states, then the only way C can have meaning for A is in relation to the set of concepts of which it is a member, which in turn derives only from the history of the self-organising process in A. These concepts have what Campbell (1994) refers to as 'causal indexicality'.

   This means that if two agents A and B have each developed concepts in this way, then if A uses its concept Ca, to think the thought 'I am having experience Ca', and B uses its concept Cb, to think the thought 'I am having experience Cb' the two thoughts are intrinsically private and incommunicable, even if A and B actually have exactly the same architecture and have had identical histories leading to the formation of structurally identical sets of concepts. A can wonder: 'Does B have an experience described by a concept related to B as my concept Ca is related to me?' But A cannot wonder 'Does B have experiences of type Ca', for it makes no sense for the concept Ca to be applied outside the context for which it was developed, namely one in which A's internal sensors classify internal states. They cannot classify states of B.

   When different agents use architecture-driven concepts, produced by self organising classifiers, to classify *internal states of a virtual machine*, and are not even partly explicitly defined in relation to some underlying causes (e.g. external objects or a presumed architecture producing the sensed states), then there is nothing to give those concepts any user-independent content, in the way that our colour words have user-independent content because they refer to properties of physical objects in a common environment. Thus self-referential architecture-driven concepts used by different individuals are strictly non-comparable: not only can you not know whether your concepts are the same as mine, the question

is *incoherent*. If we use the word 'qualia' to refer to the virtual machine states or entities to which these concepts are applied, then asking whether the qualia in two experiencers are the same would then would be analogous to asking whether two spatial locations in different frames of reference are the same, when the frames are moving relative to each other. But it is hard to convince some people that this makes no sense, because the question is grammatically well-formed. Sometimes real nonsense is not *obvious* nonsense.

We have now indicated how the process of coming to think about and ask questions about qualia is explained by the nature of the architecture of the thinker. The process arises when architecture-driven concepts produced by a self-monitoring sub-system refer internally. In talking about hese concepts we are using architecture-based concepts.

Is this a new kind of explanation? Perhaps, although it seems to have some similarities with one of Hume's explanations of the concept of 'causation'. Hume attempted to explain what a cause was by looking at what it was *about us* that made us look at the world in terms of that concept. Kant attempted to reformulate this by arguing that having a concept of causation was *necessary* for having experiences of an objective external world.

Becoming interested in and puzzled by qualia is a product of sophisticated biological mechanisms, whose nature explains why some questions about qualia can have no answers. We do not say, as some positivist philosophers would, that all talk about qualia is nonsensical: rather we try to show what can and cannot be sensibly said about them.

Some robots with our information processing architecture will discover qualia and be puzzled about them. The more intelligent ones should accept our explanation of how that happens.

4. Is something missing?

There will always be people who are convinced that this sort of project inevitably fails to answer the questions about consciousness which *they* think are the real ones. Often these are people who say of consciousness some of the things we listed in table 1, e.g.

- It's indefinable, knowable only through having it;
- It's what it is like to be something (hungry, in pain, happy, a bat . . .);
- It's possibly absent in something (behaviourally, functionally) indistinguishable from us (zombies).

The fact that many people do think like this is *part of what needs to be explained* by any adequate theory of consciousness. Our explanation is that it is a side-effect of some of the processes made possible by the existence of a meta-management layer which allows an information processing system to attend to aspects of its own internal functioning, e.g. some of the intermediate states in its sensory mechanisms, and use architecture-driven concepts to describe them. Therefore, we offer yet another conjecture:

The inevitability of consciousness talk: When robots have suitably rich internal information processing architectures some of them will also feel inclined to talk about consciousness, and qualia, in a way similar to the way we do.

This isn't a particularly new idea; science fiction writers thought of this long ago. It implies that even if philosophical theories about qualia, about 'what it is like to be something', involve much confusion and even error, it does not follow that they are completely wrong. They are based on a correct *partial* view of how minds work.

5. Zombies

Many will be *convinced* that something is left out, and will express this conviction in the form of the 'zombie' argument: a robot could have all the information processing capabilities described here (albeit sketchily) and still lack 'this' — said attending inwardly, using human meta-management capabilities. That is, the robot might be a *zombie* (for a survey of arguments see Chalmers, 1996). Usually such arguments fail to distinguish the different forms of functionalism we have distinguished. It is true that a system could have all the *externally observable* behaviours, i.e. the same input–output mappings as humans have, without having the sort of consciousness discussed here. However the claim that a robot can have all the *internal* information processing capabilities humans have, and can have internal sub-states satisfying all the same sets of causal relationships and counterfactual conditionals, and still lack something we have, becomes incoherent if nothing specific can be said about what is missing.

When internal processing of a human-like virtual machine is described *in great detail*, including the meta-management abilities involved in thinking about qualia, including, where appropriate the 'disconnected' and non-reportable states and processes, it is not clear that anything intelligible is left over: the description of a zombie as being just like us in all its capabilities yet unlike us in experiencing qualia becomes incoherent.

6. Are we committed to 'computationalism'?

It is important to distinguish two questions:

(1) Is there any information processing virtual machine architecture that is sufficient to produce mental states and processes like ours?
(2) Which, if any, of these virtual machines can be implemented on a computer (of the sort that we currently know how to build)?

It is often assumed, wrongly, that a negative answer to 2 implies a negative answer to 1. That's because many people do not appreciate that the general notion of an information processing machine is not defined in terms of computers — computers just happen to be the best tools we currently have (and even they often include sub-systems that do not conform to the standard notion of a computer, e.g. analogue components and various transducers. In future we may

invent new kinds of information processing engines, as different from computers as computers are from mechanical calculators. It might turn out that certain sorts of virtual machine architectures are adequate for the implementation of all typical adult human mental phenomena, but that no digital computer is able to support them all. We might find other kinds of previously unknown computers in biological systems.

Finding out answers requires us first to clarify meanings of the questions and the available answers. We know no better method than the method outlined here.

7. Falsifiability? Irrelevant.

What we have proposed isn't directly confirmable or falsifiable as a scientific hypothesis. Some, perhaps armed with a superficial reading of Popper (1934), might take this lack of falsifiability as grounds for rejecting our proposal. But to ask: 'Is the theory falsifiable?' is to ask the wrong question. Within the proposed framework we can make simultaneous progress in science (several different sciences) and philosophy, including investigating relationships between brain mechanisms and the virtual machine architectures described here. What's more important than (immediate) falsifiability is the ability to generate large numbers of different, non-trivial consequences about what is possible, e.g. implications about possible types of learning, about possible forms of perception, about possible types of emotions.

You can't empirically refute statements of the form 'X can happen'. But they can open up major new lines of research and unify old ones. Following Popper and Lakatos we need to ask whether this will turn out to be a *progressive* or a *degenerative* research programme. We hope we have given reason to believe that it is the former.

References

Albus, J. (1981), *Brains, Behaviour and Robotics* (Peterborough: Byte Books, McGraw Hill).
Barkley, R.A. (1997), *ADHD and the Nature of Self-control* (New York: The Guildford Press).
Beaudoin, L. (1994), 'Goal processing in autonomous agents', PhD thesis, School of Computer Science, The University of Birmingham (http://www.cs.bham.ac.uk/ research/ cogaff/).

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

Block, N. (1996), 'What is functionalism?', http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.html (from *The Encyclopedia of Philosophy Supplement*, Macmillan, 1996).

Boden, M.A. (2000), 'Autopoesis and life', *Cognitive Science Quarterly*, **1** (1), pp. 115–43.

Brooks, R. (1986), 'A robust layered control system for a mobile robot', *IEEE Journal of Robotics and Automation* RA-2, pp. 14–23.

Brooks, R.A. (1991), 'Intelligence without representation', *Artificial Intelligence*, **47**, pp. 139–59.

Campbell, J. (1994), *Past, Space and Self* (Cambridge, MA: MIT Press)

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

Chrisley, R. (1995), 'Non-conceptual content and robotics: Taking embodiment seriously', in *Android Epistemology*, ed. K. Ford, C. Glymour, P. Hayes (Cambridge, MA: AAAI/MIT Press).

Chrisley, R. (2000), 'Transparent computationalism', in *New Computationalism: Conceptus-Studien 14*, ed. M. Scheutz (Sankt Augustin: Academia Verlag).

Cooper, R. and Shallice, T. (2000), 'Contention scheduling and the control of routine activities', *Cognitive Neuropsychology*, **17** (4), pp. 297–338.

Craik, K. (1943), *The Nature of Explanation* (Cambridge: Cambridge University Press).

Damasio, A. (1994), *Descartes' Error, Emotion Reason and the Human Brain* (New York: Grosset/Putnam Book).

Gibson, J. (1986), *The Ecological Approach to Visual Perception* (, Hillsdale, NJ: Lawrence Erlbaum Associates; originally published in 1979).

Goodale, M. and Milner, A. (1992), 'Separate visual pathways for perception and action', *Trends in Neurosciences*, **15** (1), pp. 20–25.

Hauser, M. (2001), *Wild Minds: What Animals Really Think* (London: Penguin).

Kim, J. (1998), *Mind in a Physical World* (Cambridge, MA: MIT Press).

Kohonen, T. (1989), *Self-Organization and Associative Memory* (Berlin: Springer-Verlag).

LeDoux, J. (1996), *The Emotional Brain* (New York: Simon & Schuster).

Maynard Smith, J. and Szathmáry, E. (1999), *The Origins of Life: From the Birth of Life to the Origin of Language* (Oxford: Oxford University Press).

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

Nilsson, N. (1994), 'Teleo-reactive programs for agent control', *Journal of Artificial Intelligence Research*, 1, pp. 139–58.

Nilsson, N. (1998), *Artificial Intelligence: A New Synthesis* (San Francisco: Morgan Kaufmann).

O'Regan, J., Rensink, R. and Clark, J. (1999) , 'Change-blindness as a result of "mudsplashes"', *Nature*, **398** (6722), p. 34.

Peterson, D. (ed. 1996), *Forms of Representation: An Interdisciplinary Theme for Cognitive Science* (Exeter: Intellect Books).

Popper, K. (1934), *The Logic of Scientific Discovery* (London: Routledge).

Rose, S. (1993), *The Making of Memory* (London: Bantam Books).

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

Scheutz, M. (1999), 'The missing link: Implementation and realization of computations in computer and cognitive science', PhD thesis, Indiana University (University of Michigan Microfiche).

Scheutz, M. and Sloman, A. (2001), 'Affect and agent control: Experiments with simple affective states', in *Intelligent Agent Technology: Research and Development* (New Jersey: World Scientific).

Searle, J. (1980), 'Minds brains and programs', *The Behavioral and Brain Sciences*, **3**, pp. 417–57 (with commentaries and reply by Searle).

Sloman, A. (1989), 'On designing a visual system (Towards a Gibsonian computational model of vision)', *Journal of Experimental and Theoretical AI* **1** (4), pp. 289–337.

Sloman, A. (1993), 'The mind as a control system', in *Philosophy and the Cognitive Sciences*, ed. C. Hookway & D. Peterson (Cambridge: Cambridge University Press).

Sloman, A. (1996), 'Actual possibilities', in *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, ed. L. Aiello S. Shapiro (Boston, MA: Morgan Kaufmann Publishers).

Sloman, A. (2000), 'Interacting trajectories in design space and niche space: A philosopher specu-
    lates about evolution', *Parallel Problem Solving from Nature — PPSN VI, Lecture Notes in
    Computer Science, No 1917* (Berlin: Springer-Verlag).
Sloman, A. (2001a), 'Beyond shallow models of emotion', *Cognitive Processing: International
    Quarterly of Cognitive Science*, **2** (1), pp. 177–98.
Sloman, A. (2001b), 'Varieties of affect and the CogAff architecture schema', in *Proceedings Sym-
    posium on Emotion, Cognition, and Affective Computing AISB'01 Convention', York*, ed.
    C. Johnson, pp. 39-48.
Sloman, A. 2002, 'Architecture-based conceptions of mind', in *In the Scope of Logic, Methodol-
    ogy, and Philosophy of Science (Vol II)*, pp. 403–27 (Dordrecht: Kluwer; Synthese Library Vol.
    316).
Sloman, A. and Logan, B. (2000), 'Evolvable architectures for human-like minds', in *Affective
    Minds*, ed. G. Hatano, N. Okada and H. Tanabe (Amsterdam: Elsevier).
Sloman, A. and Scheutz, M. (2001), 'Tutorial on philosophical foundations: Some key questions',
    in *Proceedings IJCAI-01', AAAI, Menlo Park, California,* pp. 1–133.
    http://www.cs.bham.ac.uk/~axs/ijcai01.
Smith, B. (1996), *On the Origin of Objects* (Cambridge, MA: MIT Press).
Sussman, G. (1975), *A Computational Model of Skill Acquisition* (American Elsevier).
Weiskrantz, L. (1997), *Consciousness Lost and Found* (Oxford: Oxford University Press).
Winograd, T. (1972), 'Procedures as a representation for data in a computer program for under-
    standing natural language', *Cognitive Psychology* (later published as a book *Understanding
    Natural Language*, Academic Press, 1972).
Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*, trans. C.K. Ogden (London: Routledge
    and Kegan Paul).
Wittgenstein, L. (1953), *Philosophical Investigations* (Oxford: Blackwell; 2nd edition 1958).
Wright, I., Sloman, A. and Beaudoin, L. (1996), 'Towards a design-based analysis of emotional
    episodes', *Philosophy Psychiatry and Psychology*, **3** (2), pp. 101–26.
    Reprinted in *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, ed.
    R.L.Chrisley (London: Routledge 2000).