

Virtual Machines and Consciousness

Aaron Sloman and Ron Chrisley

School of Computer Science, The University of Birmingham, UK

<http://www.cs.bham.ac.uk/research/cogaff/>

A.Sloman@cs.bham.ac.uk

R.L.Chrisley@cs.bham.ac.uk

April 2, 2002

DRAFT version of a paper to be submitted to the *Journal of Consciousness Studies*, special issue on machine consciousness.

[[Choose another title? Abstract to be re-written?]]

Abstract

Replication or even modelling of consciousness in machines requires some clarifications and refinements of our concept of consciousness. Fortunately, design of, construction of, and interaction with artificial systems can itself assist in this conceptual development. In particular, if we start with the tentative hypothesis that consciousness is a form of information processing, we can build virtual machine architectures which attempt to capture various aspects of consciousness. This activity may in turn nurture the development of our concepts of consciousness in such a way that we can at last understand why consciousness is a form of information processing. This process leads gradual refinement of many of our pre-theoretical concepts of mind, showing how they can be best construed as “architecture-based” concepts.

Contents

1	Introduction	3
1.1	Confused concepts of consciousness	3
2	Partial diagnosis of the confusion	5
2.1	The limits of introspection	6
2.2	Beyond introspection	8
2.3	Virtual machines and consciousness	9

3	Information processing Machines	10
3.1	Types of machines	10
3.2	Information processing virtual machines	10
3.3	How to think about non-physical levels in reality	11
3.4	Virtual machine functionalism	12
3.5	Causal powers of inaccessible components	13
3.6	Could decoupled components evolve?	13
3.7	Global states vs rich ontologies	14
3.8	Some limits to self-monitoring	15
4	Organisms as information processors	15
4.1	Organisms need to process information	15
4.2	We don't need to define our terms	16
4.3	Information processing and architecture	17
4.4	The space of designs is discontinuous	18
5	Evolvable architectures	19
5.1	Reactive architectures	19
5.2	Consciousness in reactive systems	20
5.3	Pressures for deliberative mechanisms	21
5.4	Pressures for multi-window perception and action	22
5.5	Pressures for self-knowledge, self-evaluation and self-control	23
5.6	Access to intermediate perceptual data	24
5.7	Yet more perceptual and motor "windows"	24
5.8	Further steps to a human-like architecture	25
5.9	Other minds and "philosophical" genes	25
6	Some Implications	27
7	Multiple elephants: The CogAff architecture schema	28
7.1	Towards an architecture schema	29
7.2	CogAff and consciousness	30
7.3	Some sub-species of the CogAff schema	31
8	Summary of our proposal so far	31
9	Some objections	32

9.1	An architecture-based explanation of qualia?	32
9.2	Is something missing?	34
9.3	Zombies	34
9.4	Are we committed to “computationalism”?	35
9.5	The causation problem: Epiphenomenalism	35
9.6	Falsifiability? Irrelevant.	36
10	Acknowledgements	36

1 Introduction

The study of consciousness 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). It has always, however, been a part of philosophy of mind and of metaphysics. 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.

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, if we start with the tentative hypothesis that consciousness is a form of information processing, we can design, build, analyse, and experiment with virtual machine architectures which attempt to capture various aspects of consciousness. This activity may in turn nurture the development of our concepts of consciousness in such a way that we can at last understand why consciousness is a form of information processing.

This process leads to gradual refinement of many of our pre-theoretical concepts of mind, showing how they can be best construed as “architecture-based” concepts.

It also leads to an explanation of how an interest in questions about consciousness in general and *qualia* in particular can arise in machines with a certain sort of architecture that includes a “meta-management” layer. We suggest that that is the explanation 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* which is immune to some of the attacks on more conventionalist functional analyses of mental concepts. Thus we simultaneously attempt to advance science and philosophy, as well as providing a first draft general architectural schema for agent architectures. which offers a useful framework for long term AI research on human-like systems.

1.1 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:

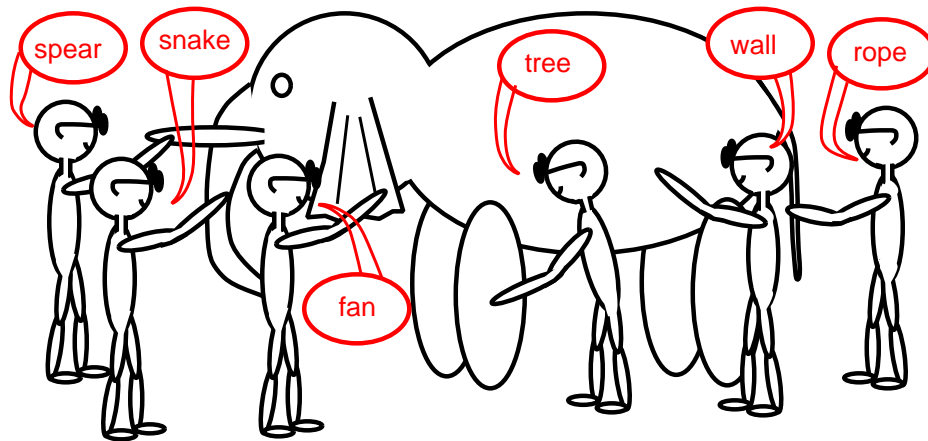


Figure 1: “The Parable of the Blind Men and the Elephant”

- 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?
- 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 not particularly original) 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 In Saxe’s poem, six blind men encounter what is, unbeknown to them, a small part of an elephant. Each feels a different part, and infers (incorrectly) from the properties of the portion encountered the nature of the whole (one feels the tusk and concludes that he has encountered something very much like a spear, another feels the trunk and deduces that he has met a snake, etc.). See figure 1.

We are in the same position with respect to consciousness. None of us can see the “whole elephant” of consciousness. What’s worse, it may be that there are actually several “elephants” involved: consciousness is probably a “cluster concept” sometimes referring to one collection of phenomena, sometimes to another collection, e.g. when we think about consciousness in different animals, or in human infants and adults.

Our blind groping has resulted in an astonishing lack of consensus on any of the important features of consciousness, as displayed in table 1.

Can we do something about this babel? We need to find a way to step outside the narrow debating arenas to get a bigger picture. 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, and at worst products of muddle, confusion, ignorance and prejudice.

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

Some say consciousness is...	While others say consciousness is...
Indefinable, knowable only through having it	What it is like to be something, e.g., hungry, in pain, happy, a bat... (Nagel 1981) ^a
Absent when you are asleep	Present when you dream
Essential for processes to be mental	Not required (there are mental processes inaccessible to consciousness)
Causes 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)
Requires 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

Table 1: A babel of views of the nature of consciousness

^aCompare http://www.cs.bham.ac.uk/~axs/misc/like_to_be_a_rock/

2 Partial diagnosis of the confusion

In order to see the bigger picture, it will help to ask why there is a babel of views in the first place. Blind groping is only part of the story; much scientific and philosophical discussion of consciousness is confused for several reasons, including:

- There is much conceptual confusion (caused partly by unwitting use of 'cluster concepts'²)

²It is not easy to define "cluster concept" precisely. This is not the same as the notion of a *vague* concept (e.g. "large", "orange") for which a boundary cannot be precisely specified along some continuum of values. A cluster concept connotes a collection of features or abilities in such a way that some combinations of those features definitely justify application of the concept, and others definitely fail to justify it, while 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. Thus, most people will agree that humans have emotions and that viruses do not, but may disagree as to whether insects do or fish do. The phrase seems to have been coined by D. Gasking and has been in intermittent use since the mid 20th century. Closely related notions are "family resemblance concept" (Wittgenstein 1953), "open texture" (Waismann 1965) and Minsky's notion of a "suitcase concept" used in his draft book on Emotions online

- There is an excessive focus on one case: normal adult (academic?) humans
- 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 (even computer scientists do not all grasp the generality and importance of the idea of such virtual machines, though they use instances of the concept every day).
- Many are victims of the illusion of “direct access” to the nature of consciousness (we experience it so directly that there is no room for mistaken beliefs about it – but simultaneity is also experienced directly and that did not prevent confusions about it.)

2.1 The limits of introspection

As a partial remedy to this situation, we propose the following guideline:

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 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 (that’s a tautology!). Compare staring carefully at trees, rocks, clouds, stars, birds and beasts hoping to discover the nature of matter. At best you learn a subset of what needs to be explained, like perceiving only the elephant’s trunk. This is not to say that introspection is useless: on the contrary, introspectively analysing the differences in the (probably familiar) visual flips in figure 2 helps to identify the need for a multi-layered perceptual system, described below.

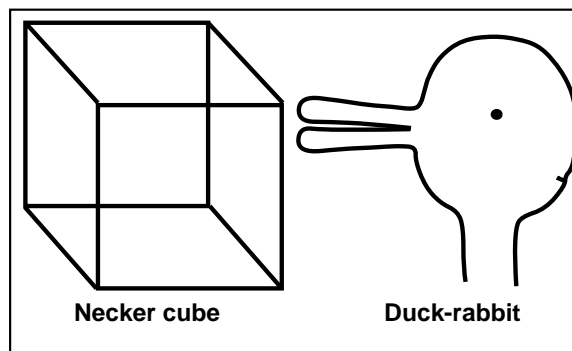


Figure 2: The Necker cube and the duck-rabbit: both are visually ambiguous, leading to two percepts. Describing what happens when they ‘flip’ shows that one involves only geometrical concepts whereas the other is more abstract and subtle.

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: The functional interpretation of the parts changes (e.g., “bill”

at <http://www.media.mit.edu/~minsky/>. Compare Ch. XI in (Cohen 1962). We argue here and in (Sloman to-appear) that by construing cluster concepts as “architecture-based” we can reduce their indeterminacy by improving our architectural theories. See: <http://www.cs.bham.ac.uk/research/cogaff/sloman-lmpsfinal.pdf>

into “ears”). More subtle features change, attributable only to animate entities. For example, “Looking left”, or “looking 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? Is this seeing, or only inferring? The differences between different experiences of the same ambiguous picture are visual, not simply inferential, which is why the examples occur in textbooks on *vision*, not *reasoning* (Sloman 1989).³ Another example is seeing a face as happy or sad. We return to the perception of other agents in the discussion of “multi-window” perception, in explaining the H-Cogaff architecture, in section 5.8.

These examples remind us that the experience of seeing has hidden richness. What constitutes our experience at any time includes a large collection of unrealised, un-activated, but potentially activatable capabilities, in addition to a large collection that we unaware of activating. Can we say more about what those unacknowledged capabilities are? One way is to learn from psychologists and brain scientists about the many bizarre ways that these capabilities can go wrong. But we can also learn new ways of looking at old experiences: For example, how exactly do you experience an empty space? See figure 3.

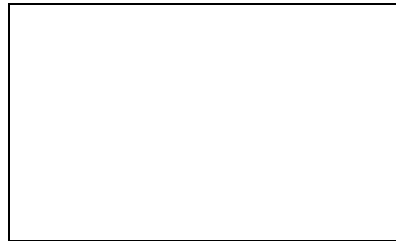


Figure 3: The final frontier?

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? (A possible answer is: Harold Cohen’s remarkable AARON program on display at <http://www.kurzweilcyberart.com/>. An early version is discussed in (Boden 1990).)

These conclusions about the nature of perceptual experience are partly based on introspection: introspective analysis of experience can be useful. But doing it well can be very difficult, since most of what makes an experience what it is is not part of the experience. An experience is constituted partly by the collection of implicitly understood *possibilities for change* inherent in that experience – a sort of “grammar” for experienced processes. This is closely related to Gibson’s “affordance” theory (Gibson 1986, Sloman 1989, Sloman 1996). We need a more precise characterisation of the grasp of what is and is not possible in order to constrain the search for explanatory mechanisms and forms of representation. (Are modal logics plausible candidates? What alternatives are there?)

The examples show that when we have experiences there may also be a lot going on of which we are completely unaware. *We do not experience what it is to experience something*. This leads us to make a second conjecture:

³Compare the discussion of ‘Seeing as’ in Part 2 section xi of (Wittgenstein 1953)

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.

2.2 Beyond introspection

Instead of merely gazing at our internal navels, we need to collect far more data-points, 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 predator, catching fast moving prey, etc.
- differences between humans at various stages of development;
- differences between mental phenomena in different cultures;
- unobvious aspects of conscious experiences, (including unconscious aspects);
- surprising effects of brain damage or disease;
- similarities and differences between different species;
- stages and trends in evolution.

Like its counterpart, introspection, mere data collection is not enough. In particular “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. So effective data collection 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, especially by building working models.

We also 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 laws or correlations.

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 (including languages which specify machines that interpret new languages). The latter are languages for specifying and explaining the behaviour of information processing machines. As we try to show below, such languages may have a special connection to consciousness, even if no *existing* programming language or design language is adequate for our purposes.

So there’s far more besides the actual contents of experience that needs to be understood and explained in explaining what experiencing is. In particular, we need to understand what sort of machinery makes possible the implicit grasp of myriad possibilities for change inherent in human consciousness of something unchanging. Insofar as different organisms, or chil-

dren, or people with various sorts of brain damage or disease, have different kinds of mental machinery, the types of experiences possible for them will be different.⁴

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 robot labs. There are many more elephants than meet the eye – wherever you look, or grope.

2.3 Virtual machines and consciousness

To make a fresh start at explaining consciousness, let's make explicit a fundamental premise:

Basic working assumption: Consciousness is not magic; it results from the operation of (very complex) information-processing machines which we do not yet understand.

Consciousness started as a biological phenomenon. It was produced by evolution, somehow using physical resources. But that does not make it itself a physical phenomenon, 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, etc.

Nor is consciousness one thing: there are many varieties of consciousness in biological organisms. We need a new conceptual framework for thinking about how these varieties differ and how they are similar. But despite their differences, we claim that there is a common core: they all depend on the fact that biological organisms are information processors. We can also abstract away from some of the specifics of evolved life on earth to explore more varieties of information processors, as long as we take care to learn from the biological subset. Suitable non-biological machines should, in principle, be able to replicate most aspects of biological forms of consciousness. Future human-like machines will re-discover all the puzzles of consciousness that have befuddled humans – if they have the same or very similar mental features. (NB: not all humans are alike).

This approach, encompassing many physical forms of organism and machine is made tractable by (temporarily) ignoring many of the physical differences between systems and focusing 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). We believe they both have a partial view of the same relation. (Another elephant.)

⁴Even if a heavenly leopard lies down with a kid or a lion with a lamb their visual experiences when gazing at the same scene may be different because of the affordances required in their evolutionary history.

3 Information processing Machines

3.1 Types of machines

Central to the approach we are advocating here is the concept of a machine. 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...*

We are concerned with the third class of machines, the information processing machines. The mechanisms that make information manipulation possible are not just physical machines, e.g. made of blood, meat, etc. They include also *virtual* machines. In computer science, software engineering and AI we have learned the importance of virtual machines, e.g. Lisp, Prolog and Java virtual machines; chess virtual machines; spreadsheets; neural nets; etc. Mechanisms that operate on rapidly changing complex information structures are typically virtual machines rather than physical machines; for example, parsers, compilers, structure matchers, search engines, planners, etc. Physical machines cannot change their structure fast enough, though virtual machines *implemented* in physical machines can.

3.2 Information processing virtual machines

What makes them *virtual* machines is the fact that the entities within them are *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. In fact, without changing any laws of physics it may be possible for the same virtual machine to play in “teaching mode” for beginners, in such a way that that generalisation no longer holds (as can a human chess player). Its (partial) predictability as a chess player depends on the fact that the components in which it is implemented obey the laws of physics, but the very same virtual machine could be implemented in different components subject to different laws (e.g. valves instead of transistors, or some new kind of computing machinery).

Moreover, the laws that govern the virtual machine are not derivable by pure mathematics or pure logic from the physical laws governing the components: “bridging postulates” will be needed to link the states and processes in the virtual machine to those in the physical machine. They cannot be linked 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. It’s a different ontology. But the ontology is built by engineers,

not by magic, and no mysterious spiritual machines are needed.⁵

If we are to explore the full range of designs for behaving systems, we need to be familiar with a wide range of techniques for constructing virtual machines of various sorts. Many clues to the variety of possible types of virtual machines come from living things.

Evolution “discovered” and used many things long before human engineers and scientists thought of them. Paleontology shows the development of physiology and provides some weak evidence about behavioural capabilities. But there is very little direct evidence regarding early forms of information processing: *virtual machines leave no fossils*. But surviving forms of information processing give clues, and we can test theories in working models.

Some of the forms 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.) We return to questions about the powers and functions of different systems, and how they fit together below in section 4.

3.3 How to think about non-physical levels in reality

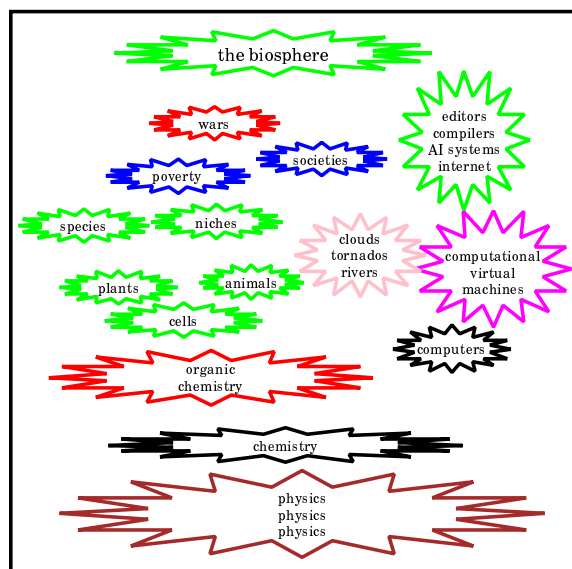


Figure 4: Various levels of reality, most of which are non-physical

There are many families of concepts, or “levels” on which we can think about reality (see figure 4 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 “promiscuous” pluralism; the world is one in at least this sense: all non-physical levels are ultimately implemented in physical mechanisms. However, nobody knows how many levels of virtual machines physicists themselves will eventually discover. In any case, the history of human thought and culture shows not only that we are

⁵These are controversial comments, discussed at greater length, though possibly not yet conclusively, in papers here: <http://www.cs.bham.ac.uk/research/cogaff/>. (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.

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.

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.

3.4 Virtual machine functionalism

Functionalist analyses of mental concepts are fairly popular, though there are a number of difficulties (e.g. the “zombie” problem discussed below). We claim that these difficulties can be overcome by basing our analyses on *virtual machine functionalism*.

Most philosophers and cognitive scientists write as if ‘functionalism’ were a well defined generally understood concept. E.g. (Block 1996) says

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.

I.e. according to Block a functional state S of a machine or animal A is defined in terms of sets of possible inputs and outputs to A and other functional states of A. The state S is identified by what happens for each possible input of A: what output will A produce in response (if any) and what state transitions will occur in A. So a functional state of A is defined in terms of causal connections between inputs and outputs of A and other functional states. On this view there could not be part of the system that was never affected by inputs and itself never affected any outputs, for instance a sub-process running in a computer that did nothing but explore mathematical proofs, using only internal storage mechanisms.

In contrast, what we call “virtual machine functionalism” refers to virtual machines whose states and events can have causal connections only with other such states, and without requiring any links to external transducers. These states *may* be linked to external transducers, but they need not be. Virtual machine entities, states and processes can exist and interact without any *external* causal connections. An example would be a process running a program that neither reads nor prints anything, but merely explores theorems derivable from a set of axioms, or which repeatedly plays games of chess with itself.

After such a process starts up, 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 become detached (as sometimes happens to “runaway” processes in computers).

In principle it might be possible to infer what is happening 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 physical traces might require more time than the history of the universe.)

3.5 Causal powers of inaccessible components

Although events in such a decoupled component of a virtual machine are not externally detectable by modifying inputs and checking outputs, nevertheless they support true counterfactual conditional statements of the form “If such and such a connection were made to the speech understanding and speech generating mechanisms then this individual would answer questions about theorems it had proved”. But the truth of that counterfactual does not require the sub-machine to be *actually* connected to input or output modules. Moreover, it is possible that the bandwidth of the available output devices may be too limited to express full information about the virtual machine states.⁶ Despite the difficulties in testing, 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 be turned off.)

In Block’s definition, quoted above, we see a hint of all this, for he allows internal events to have two kinds of effects, external (e.g. saying ‘ouch’) and internal (e.g. wondering whether you are ill, being distracted). It is not clear, however, whether he realised that some of the states and processes allowed as functional might be completely disconnected from inputs and outputs. The very label “functional” is often taken to connote having some kind of potential survival value for an organism or robot, which disconnected internal processes could not have.

3.6 Could decoupled components evolve?

That is a problem for “virtual machine functionalism” if it is proposed as a framework for analysing mental phenomena humans and other animals. For it is unlikely that biological evolution could produce such disconnected virtual machines: if they have no behavioural effects that influence biological fitness they would not be selected. The answer to this is that selected mechanisms can have side-effects. A mutation that produced biologically useful brain processes might also produce disconnected virtual machine components. Or a harmless mutation might produce a disconnected copy of a connected virtual machine. The additional machine would require additional energy to be consumed to no purpose, but the difference might be too small to be selected against in a lush environment.

From our viewpoint the various components of a virtual machine can themselves be seen as evolving sub-systems influenced by co-evolving subsystems: a mind can be viewed as an eco-system (Sloman & Logan 2000).

The upshot of all this is that we can talk about architectures for virtual machines, where the components of the architectures are defined primarily in terms of causal interactions purely

⁶We claim that a careful reading of (Ryle 1949), especially his chapter on imagination, shows a view very similar to virtual machine functionalism, although he did not have our computational concepts. He is often wrongly characterised as a behaviourist.

within the virtual machines, like two chess programs repeatedly playing chess with each other. We can then regard any direct or indirect links to external transducers as *contingent* additions, not logically necessary, features. A consequence of this is to reduce the significance of psychological experiments as a way of studying psychology. It also undermines “zombie” arguments against functionalism as explained below.

3.7 Global states vs rich ontologies

Another feature of virtual machine functionalism is that it is not restricted to referring to states of the *whole* system. When a virtual machine is running, for instance a chess-playing virtual machine, there are virtual machine entities that endure across events that change the machine’s state, for instance the board, positions on the board, pieces on some of the positions, etc. Their states do not necessarily all change in synchrony. In addition there may be some entities that exist for only a short time, with components and relationships that vary rapidly. For instance while the machine is considering a move it may temporarily build a partial lookahead tree which it examines, and then extends, and then shortens, and then extends in another way. So while some nodes of the tree are created and destroyed others continue to exist. Later almost all of the tree may be removed just before a move is selected. So the notion that an information processing machine has only one global but atomic (unanalysable) state at a time, as indicated in state transition tables, does not do justice to the complexities and subtleties that have been found necessary for the purposes of AI and software engineering.

Even if all the states of varying complexity can be mapped onto a collection of such global states with transitions between them (e.g. by treating every change of sub-state, no matter how minor, as a change of state of the whole machine), this fact gives no insight into what really happens when most interesting computer programs run. There is therefore no reason to believe that the notion of a unitary state-transition table will be helpful in thinking about human minds either.

So instead of thinking of a virtual machine only in terms of transitions between possible states of the whole system we need to think of it as having a rich ontology of entities of different kinds and different degrees of complexity, which coexist but endure over different time-scales, being created and destroyed as required, with changing properties of and relations between those entities, and changing internal structure and complexity in some of the entities.⁷

Moreover, even if some of the main features of such a virtual machine are capable of being tested for while it is running, there can be many details in the temporary data-structures that have no links to external sensors or effectors, and could not be revealed by the program. For some other internal processes it may be possible to reveal them in summary form, but not in complete detail. For others, any attempt to report their existence, or their features and relationships, will interfere with the running of the machine, as many students of programming have found if they attempt to turn on trace-printing for the printing sub-routines. This kind of complexity is commonplace in programming, yet ignored in most philosophical discussions.

⁷This is also an objection to standard presentations of dynamical systems theory.

3.8 Some limits to self-monitoring

Later, we'll introduce the notion of a system in which some virtual machine processes monitor, categorise, evaluate and perhaps also modify other processes. This kind of internal self-observation can itself be a process of a kind that, in some cases, cannot be externally reported, or in some cases can only be reported partially or inaccurately. Thus a system may have information about itself that is not accessible externally. But some internal processes may not be self-accessible, either because suitable connections do not exist or because there is no suitable internal formalism for recording what happens in those processes, or because the information routes within the system do not have sufficient band-width to allow all the details to be recorded, or because recording all details would require the memory capacity to grow too fast.

4 Organisms as information processors

Successful organisms are information processors (in this discussion, we use "information" in the everyday sense, not the Shannon/Weaver technical sense). This is because survival for organisms, unlike rocks, mountains, planets and galaxies, typically requires action (exceptions may be certain dormant spores that are able to survive indefinitely without doing anything), and actions must be selected and initiated under certain conditions. Therefore successful selection and timed initiation requires at a minimum information about whether those conditions obtain. See figure 5.

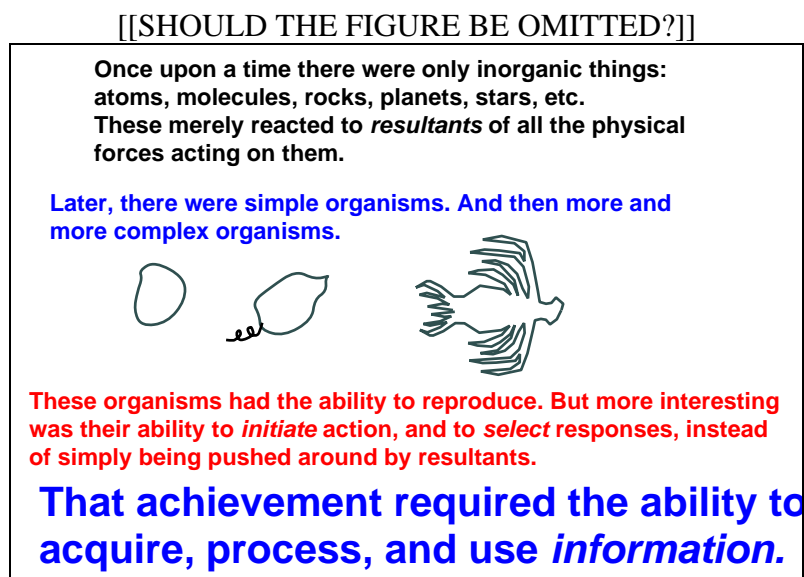


Figure 5: A fairy tale?

4.1 Organisms need to process information

Acting or selecting requires information, which in an organism may include the following: 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 you fly through
- how much of your nest you have built so far
- which part of the nest should be extended next
- where a potential mate is
- something that might eat you
- something you might eat
- what that thing over there is likely to do next
- how to achieve or avoid various states
- how you thought about that last problem
- whether your thinking is making progress

and much, much more. Although most of these processes do not involve *self*-consciousness, they do involve a kind of consciousness, or sentience, defined as the ability to acquire information about something. We shall see that some organisms have richer varieties of consciousness.

4.2 We don't need to define our terms

It is important that we resist the urge prematurely to ask 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? 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 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. Like many deep concepts in science, it is *implicitly* defined by its role in our theories and our designs for working systems. To illustrate this point, we offer this partial analysis of *things an organism or machine can do with information*:

- Reacting immediately (it can trigger immediate action, either external or internal)
- Segmenting, clustering labelling of components within a complex information structure (i.e. parsing)

- Trying to derive new information from it (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- Storing it for future use (and possibly modifying it later)
- Considering alternative possibilities, e.g. in planning
- Interpret it as as containing instructions, and obeying them, e.g. carrying out a plan
- Observing the process of doing all the above and deriving new information from it (self-monitoring, meta-management)
- Communicating it to others (or to oneself later)
- Checking it for consistency, either internal or external.

... and many more. Different forms of representation may be used for different purposes.

4.3 Information processing and architecture

What an organism or machine can do with information depends on its architecture – not just its physical architecture, but also its virtual machine information processing architecture. An architecture includes:

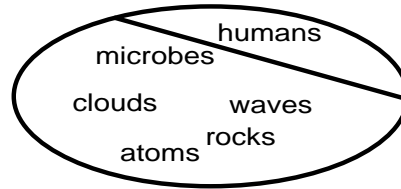
- *forms of representation,*
- *algorithms,*
- *concurrently processing sub-systems,*
- *connections between them,*

In addition 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 computer operating system are both examples. The various kinds and uses of information processing did not all evolve at the same time. Not all of them occur in all animals. For example, perceptual mechanisms that evolved at different times provide very different sorts of information about the environment:

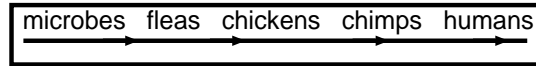
- Some information is very localised and simple (here's a dot, there's some motion);
- Other information may be far more holistic (e.g. recognising a scene as involving a forest glade);
- Some may be very abstract (the weather looks fine; it looks as if a fight is about to break out in that crowd) compare figure 2;
- Some perceptual mechanisms involve only general-purpose knowledge about the geometry and topology of static and moving shapes;
- Others require specific knowledge about things that are relevant only in a particular part of the world, or a particular type of activity. For example, seeing text, hunting fast-moving prey, seeing geological formations, looking at exposed brains.

These different biological information processing functions require different kinds of mechanisms, often operating on different forms of representation and different forms of long and

1. A dichotomy (one big division):



2. A continuum (seamless transition):



3. A space with many discontinuities:

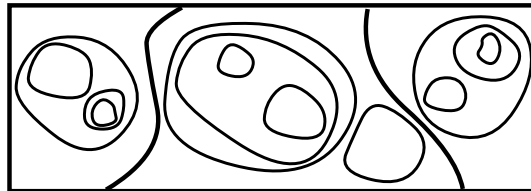


Figure 6: Models of conceptual spaces. It is often assumed that the only alternative to a dichotomy (conscious/non-conscious) is a continuum of cases with only differences of degree. There is a third alternative.

short term storage. Sometimes they require different sub-mechanisms working together (perceiving, learning, using prior knowledge, deciding what to do, constructing plans, executing plans, etc.). But there is always an overall architecture containing all the mechanisms and the processes they produce.

Some of the more sophisticated mechanisms and architectures evolved only relatively recently, and are in very few species (e.g. deliberative capabilities, described below). We need to understand how the new mechanism differ from, how they are built on, and how they interact with the much older, more wide-spread mechanisms. The same organism, e.g. a human, may include both very ancient commonplace mechanisms and very new rare mechanisms, in many sub-systems.

If what we have said so far is correct, then in order to understand consciousness we need to understand the space of information processing architectures and the states and processes they can support – including the varieties of types consciousness. There's no unique, correct architecture; the belief that there is amounts to believing the conscious/non-conscious distinction is a dichotomy, as in figure 6. Many assume that the only alternative is a continuum, in which all divisions are arbitrary and all differences are differences of degree. Both are wrong, since conceptual spaces may have many discontinuities. Examples are the space of possible designs and the space of requirements for designs (niches) (Sloman 2000).

4.4 The space of designs is discontinuous

The view that consciousness is just a matter of *degree* (see figure 6) ignores the fact that many evolutionary and developmental changes in biology are inherently discontinuous, involve many changes of kind, and can be big or small, for example, structural changes, and

development of new capabilities. There are two reasons for this discreteness: (a) molecular structures and molecular changes in DNA are discrete, and (b) there can be only a finite number of generations between any two time points, which rules out a continuum of stages. Likewise niches, and sets of requirements, can change discontinuously, depending on how surrounding designs change. Some discontinuities may be big, others small, so not every space is either a continuum or a dichotomy. All of this is not only compatible with, but implied by Darwinian evolution (Sloman 2000).

There are many different types of designs, and many ways in which designs can vary. Some variations are continuous (getting bigger, faster, heavier, etc.). But some variations are discontinuous. For examples changes between two software designs are discontinuous.

Many of the changes that might be made to a system (by evolution or design) are discontinuous:

- duplicating a structure,
- adding a new connection between existing structures,
- replacing a component with another,
- extending a plan.
- 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.⁸

5 Evolvable architectures

Evolution has obviously produced an enormous variety of physical designs for organisms. 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 seem to have deliberative capabilities in addition. An even smaller subset seem to have meta-management capabilities (described below). These different architectural components support different varieties of consciousness.

5.1 Reactive architectures

A reactive mechanism (figure 7) 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 considering and comparing alternatives or deliberating about explicitly represented possibilities.

⁸Can a chimpanzee do any of those? Some apes can do amazing things (Hauser 2001).

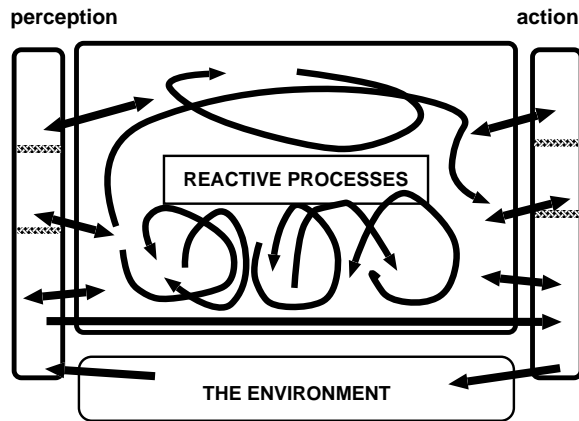


Figure 7: A simple, insect-like architecture. Arrows indicate direction of information flow. Some reactions produce internal changes that can trigger or modulate further changes. Perceptual and action mechanisms may operate at different levels of abstraction, using the same sensors and motors.

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 15), 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 can 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).

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 cannot use strategies not previously selected by evolution (or by a designer in the case of artificial reactive systems). These limitations can be overcome by deliberative capabilities, as evolution seems to have discovered, for some organisms.

5.2 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. How-

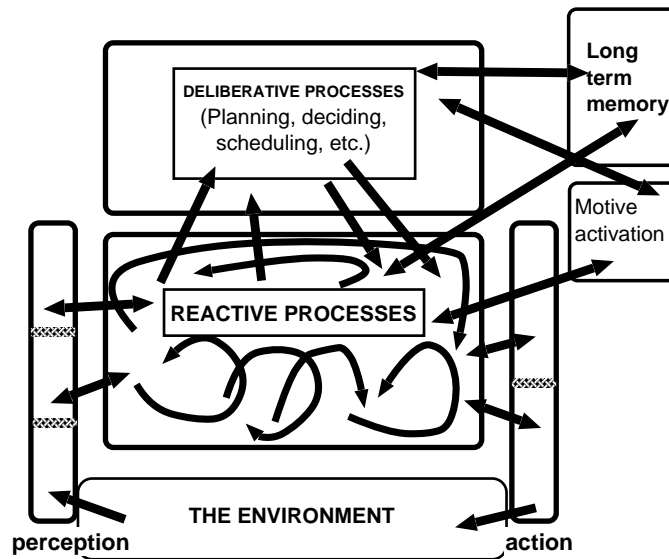


Figure 8: An architecture that is both reactive and deliberative.

ever, 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. (Compare ways of using information that a table is in the room.) Rather, it seems that their sensory inputs directly drive action-control signals, though possibly after transformations which may reduce dimensionality, as in simple feed-forward neural nets. There may be 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.

Typically purely reactive systems do not use information with the same type of flexibility as a deliberative system which can consider non-existent possibilities. They also lack 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. It is this more elaborate answer, rather than a simple “yes” or “no”, which is the best reply to the question “is a fly conscious?”

5.3 Pressures for deliberative mechanisms

Sometimes the ability to plan is useful; in such cases, an architecture such as that depicted in figure 8 is an advantage. The laminar structure in the vertical dimension need not correspond to spatially distinct aspects of the system; more usually, the higher layers will be realised in (some subset of) the physical aspects of the system which realise the lower layers. This could result from an evolutionary step in which some reactive component is first duplicated then later given a new function (Maynard Smith & Szathmary 1999).

Deliberative mechanisms provide the ability to represent possibilities (e.g. possible actions, possible explanations for what is perceived). Purely deliberative architectures were usually employed in traditional AI (except in robotics, since robots need reactive mechanisms in

addition to planning capabilities). A famous example was 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. Deliberative mechanisms can differ in various ways, e.g.:

- the forms of representations used (often data-structures in virtual machines, e.g. logical, pictorial, activation vectors – some with and some without compositional semantics).
- whether they use external representations, as in trail-blazing or message-sending
- 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 ability to represent future, past, 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)

Some deliberative capabilities require the ability to learn new abstract associations, e.g. between situations and possible actions, between actions and possible effects. In a hybrid reactive-deliberative architecture, the deliberative part may be unable to act directly on the deliberative 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”. The perception of possibilities and constraints on possibilities (affordances) is something that has not yet been adequately characterised or explained (Sloman 1996). Hybrid reactive-deliberative systems can have more varieties of consciousness, though the kinds in different parts of the architecture need not be integrated.

5.4 Pressures for multi-window perception and action

Deliberative capabilities may provide the opportunity for more 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 “taller” perception and action towers in figure 9. 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.

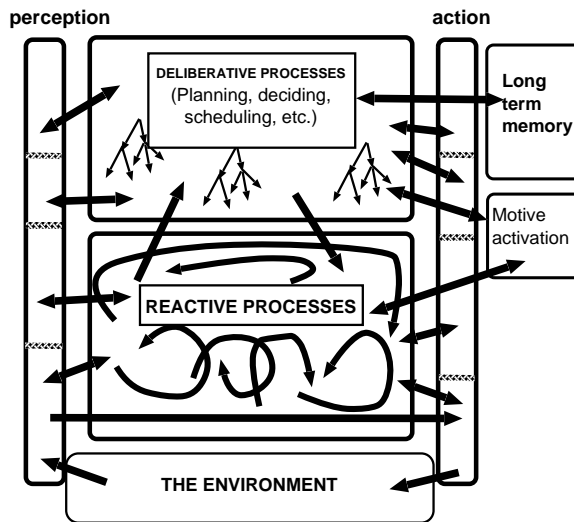


Figure 9: Reactive-deliberative architecture with “multi-window” perception and action. Higher level perceptual and motor systems (e.g. parsers, command-interpreters) may have “direct” connections with higher level central mechanisms.

5.5 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 main causes of stupidity in many 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 connotations. It seems to be rare in biological organisms and probably evolved very late. Much research on ‘reflective’ AI systems is in progress. A simplified version is depicted in figure 10. As with deliberative and reactive mechanisms, there are many forms of meta-management and we’ll discuss more elaborate versions below.

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). Among psychologists and psychiatrists it is fairly common to refer to “executive” functions. We believe that this concept as normally used does not adequately distinguish deliberative from meta-management functions. 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 sub-division.

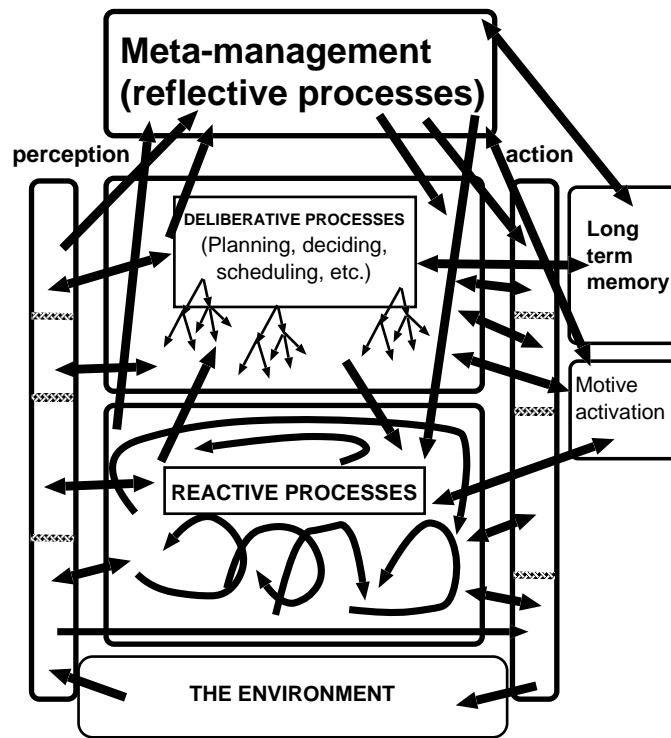


Figure 10: A reflective architecture

5.6 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 10. 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 with development 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. A full catalogue of uses for such internal monitoring mechanisms has not yet been produced.

5.7 Yet more perceptual and motor “windows”

We conjecture that, as indicated in figure 11, 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 process can also occur, during social learning, e.g. when people learn to think of their own thoughts as 'sinful', a powerful tool of social control. So:

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 (cf figure 12).

5.8 Further steps to a human-like architecture

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,
- 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 11, 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 **c**ognition and **a**ffect) could be realised in robots (in the distant future).

5.9 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 (in multi-window perception illustrated in figures 2 and 12) 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, for prey species and for organisms that interact socially. Such mechanisms, like the innate mechanisms for perceiving and reasoning about the physical environment might have a "bootstrapping" component that fills in many details during individual development through a process of interacting with the environment.

Insofar as those mental concepts, like the self-categorising concepts, refer to internal states and processes they will be architecture-based even though the naturally developed, implicitly pre-supposed architectures are probably much simpler than the actual virtual machine

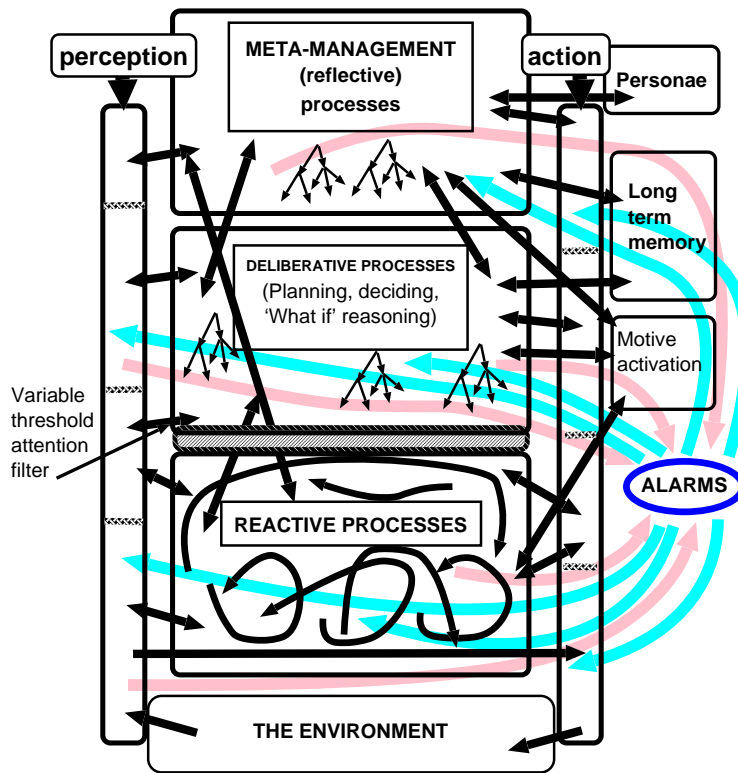


Figure 11: The H-CogAff architecture. As before, the addition of a central layer provides pressure to evolve yet more layers in perceptual and motor sub-systems, with direct links to the new layer.

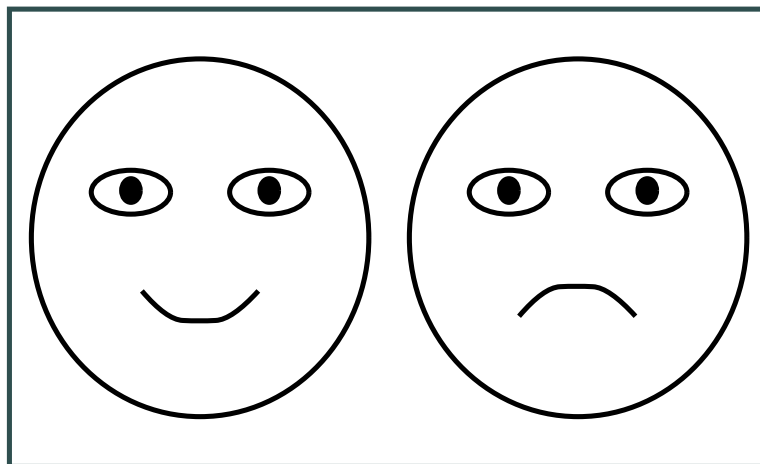


Figure 12: Seeing a state of mind when looking at a face is an example of perceptual processes linked to concepts used by a meta-management system. Compare figure 2.

architectures of other agents. In other words, even if some animals 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, by providing built-in apparatus for conceptualising mental states in others at least within intelligent prey species, predator species and social species.

6 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 layers: *primary* (exclusively reactive, such as anger), *secondary* (partly deliberative, such as frustration) and *tertiary* emotions (including disruption of meta-management, such as grief). Some of the explained emotions might be shared with other animals, some will be unique to humans (Wright, Sloman & Beaudoin 1996, Sloman & Logan 2000);
- Discovery of *different visual pathways*, since there are many routes for visual information to be used (Sloman 1989, Goodale & Milner 1992, Sloman 1993) — though we can expect later research to distinguish many more perceptual pathways as we come to understand the varying architectural requirements for perceptual input (e.g. detecting various kinds of affordances and functioning in various kinds of feedback control loops);
- Many possible *types of brain damage* and their effects, e.g. frontal-lobe damage interfering with meta-management (Damasio 1994);
- *Blindsight*: damage to some meta-management access routes prevents self-knowledge about intact (reactive?) visual processes (Weiskrantz 1997);
- 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, (perhaps corresponding, in the human case, to structures in the cerebellum). 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. We can also analyse development of the architecture in infancy, including development of personality as the architecture grows;
- 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 timing: people may take part in experiments that indicate when they become aware that they have decided, without realising that their decisions were actually taken slightly before the awareness of the decision was brought about.

[[XXX REFERENCE to Libet's work??]]

Experiments on change-blindness [[XXX ref O'Regan]] and many others assume that people can tell 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.

- 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;
- 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.)
- Discovery by philosophers of sensory ‘qualia’. We can see how philosophical thoughts (and confusions) about consciousness are inevitable in intelligent systems with partial self-knowledge (discussed further below in section 9.1).

7 Multiple elephants: 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, presents a (fairly) complete elephant. (At least more complete than normal). But there are different architectures, with very different information processing capabilities, supporting different states and processes. For example, fleas, fishes, and philosophical humans. So there are many elephants – not just one.

Thus, we consider families of architecture-based mental concepts. For each architecture we can specify a family of concepts of types of virtual machine information processing states, processes and capabilities supported by the architecture. 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. Likewise, architecture-based mental concepts can extend and refine our semantically indeterminate pre-theoretical concepts, leading to much clearer concepts related to the mechanisms that can produce different sorts of mental states and processes.

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 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

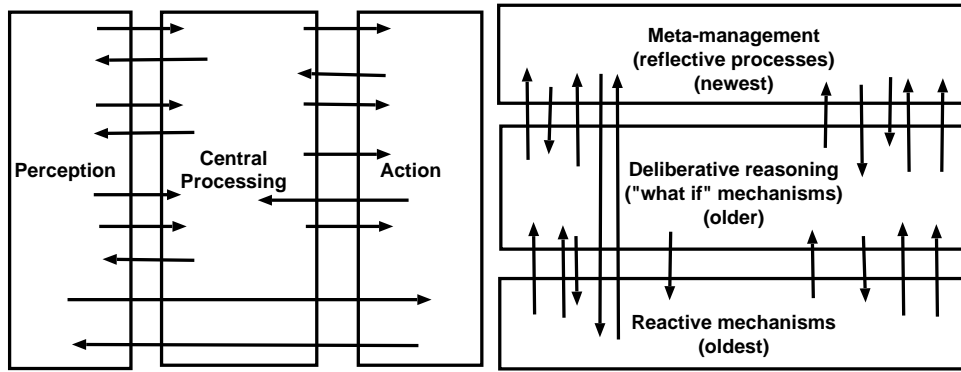


Figure 13: The vertically and horizontally separated components of architectures: towers and layers.

architecture-based concepts of H_2O and D_2O 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?”).

7.1 Towards an architecture schema

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

Two coarse divisions within components of information processing architectures are familiar, as depicted in Nilsson’s “triple tower” model (Nilsson 1998) and models with ‘layers’ (e.g. reactive, deliberative and meta-management layers), as in figure 13. These orthogonal functional divisions can be combined in a grid, as indicated in figure 14. 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 5.8), CogAff is a schema not an architecture: it is a sort of ‘grammar’ for architectures. 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, microbes). The reactive layer can include mechanisms of varying degrees and types of sophistication, some analog, some digital, with varying amounts of concurrency. The other two layers can also differ between species.

The CogAff schema can be extended in various ways: e.g., see figure 15. 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

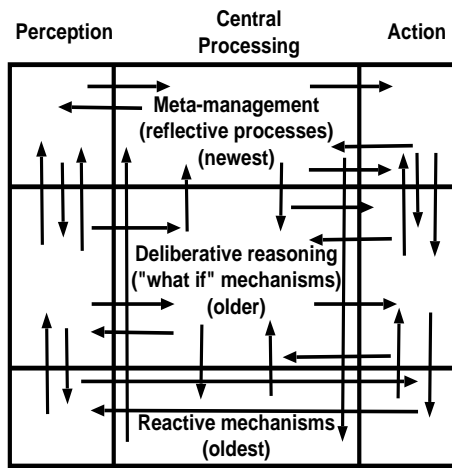


Figure 14: The CogAff schema: superimposing towers and layers.

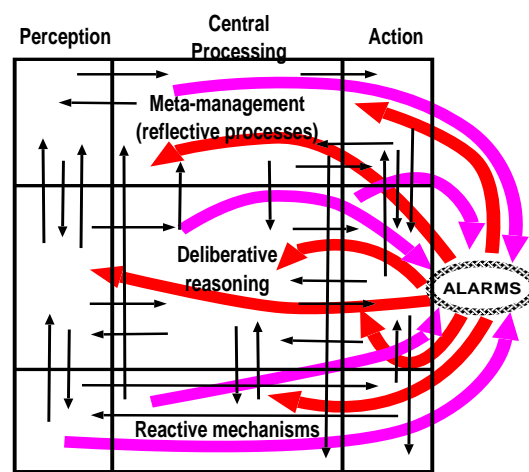


Figure 15: Extending the CogAff schema to include alarms.

with and without various kinds of learning, or with various kinds of formalisms; and between many sorts of specialised “alarm” mechanisms. Alarm mechanisms deal with the need for rapid reactions using fast pattern recognition based on information from many sources, internal and external. An alarm mechanism is likely to be fast and stupid, i.e. error-prone, though it may be trainable. The CogAff schema is still a draft, likely to evolve.

7.2 CogAff and 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.

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 varieties of consciousness that are possible within an individual can develop over time.

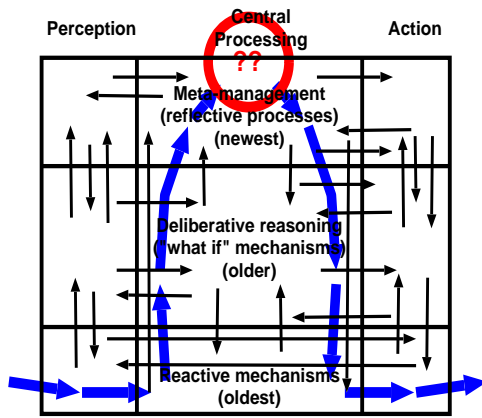


Figure 16: Omega architectures: a common sub-species of the CogAff schema.

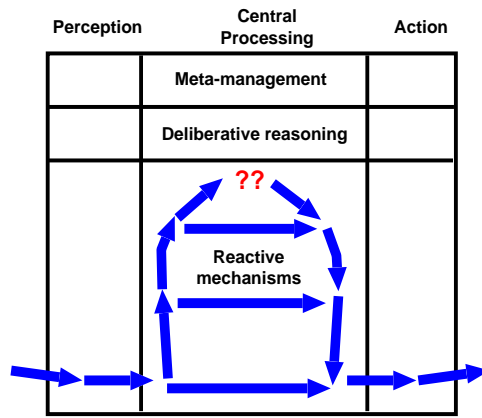


Figure 17: Subsumption architectures: another sub-species.

7.3 Some sub-species of the CogAff schema

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 (see figure 16). This is just a pipeline, with “peephole” perception and action, as opposed to “multi-window” perception and action (see section 5.4). For examples see (Albus 1981, Cooper & Shallice 2000).

Another sub-species of CogAff is the subsumption architecture (Brooks 1991); see figure 17. This 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 is “correct”, or “sufficient for consciousness”.

8 Summary of our proposal so far

The solutions and dissolutions of muddles about consciousness that we offer are based on:

- Virtual machine functionalism – emphasising internal causal powers, states, interactions, implemented in physical mechanisms, but not themselves physical (like poverty);
- Comparative studies of minds of many kinds (infants, toddlers, children humans, healthy, damaged, disturbed, many kinds of animals, many kinds of machines);
- Investigation of the (huge) space of virtual machine architectures, including both evolved and designed architectures;
- Refining and extending (not replacing) existing confused concepts with several families of architecture-based concepts (many ‘elephants’) (different information processing architectures support different varieties of consciousness – and different varieties of learning, motivation, beliefs, emotions, intentionality, etc.);

- Separation of conceptual questions (how to distinguish different concepts of experience) from empirical issues (which types does a new born human infant, or a mouse, have);
- Integration of multiple disciplines (for example, Philosophy, psychology, ethology, neuroscience, evolution, artificial intelligence, software engineering and computer science);

This is a complex, long term research programme. Our approach is a mixture of science, engineering and philosophy. The science and engineering include the study of evolvable virtual information processing architectures as part of a larger study of:

- the space of possible designs;
- the space of possible niches;
- relations between those spaces;
- trajectories within those spaces;
- the dynamics of interacting trajectories in those spaces.

On the basis of such explorations we develop a general conceptual framework for defining different types of consciousness in terms of the information processing architectures that support them and the kinds of states, events and processes that can occur in those architectures. That can then lead to empirical investigations to find out which animals have particular precisely defined sorts of consciousness: a deeper and more rewarding quest than arguing about the presence or absence of one ill-defined sort.

9 Some objections

9.1 An architecture-based explanation of qualia?

At this point, some readers might be at the point of exasperation, wondering how these architectural notions could make any headway on some infamous problems of consciousness, e.g., the problem of qualia – the private, ineffable way things seem to us. In section 6, we suggested that an architecture-based explanation of qualia is possible. Such an approach doesn't explain qualia by saying what they *are*. Instead we explain qualia by *explaining the phenomena that generate philosophical thinking of the sort found in discussions of qualia*.

The discovery of the concept of 'qualia' is a consequence of having the ability to attend to aspects of internal information processing (internal self-awareness), and then trying to express the results of such attention. That possibility is inherent in any system that has the H-CogAff architecture (see section 5.8), though different versions will be present in different architectures, depending, e.g., on the forms of representation and modes of monitoring available to meta-management. A meta-management system could give an agent 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 sensory data-structures.

Consider perceiving a table. You 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

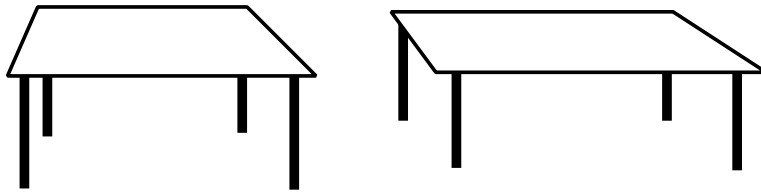


Figure 18: Noticing two perspectives on the same object is one route to concerns about qualia.

change as you change your viewpoint (or the table is rotated; see figure 18). 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 reflective robots discussing their experience of the same table seen from different viewpoints could get bogged down discussing consciousness, just like six blind philosophers.

Moreover, if the precise set of concepts used by each individual is in part a product of that individual’s learning history, for instance if the concepts are produced by a self-organising neural network, then the concepts of qualia, or sense-data, used by different individuals will be strictly non-comparable: not only can you know which of your qualia are the same as mine, the question is incoherent, for the nature of the qualia will be determined mainly by how they fit into a system of relationships to other qualia that can exist in the same system. 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.

Note that we are not saying what qualia are. We rather explain how the process of coming to think and worry about qualia is explained by the nature of the architecture of the thinker. (It is an “architecture-driven”) concept (contrasted with architecture-based concepts which are not necessarily used within the architecture that supports them).

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.

Again, we are not saying: “this is what qualia are...” Instead, we offer (conjectured) *sufficient* conditions for an information processing system to go through the very same processes as led humans to start thinking about sensory (and other kinds of) qualia. Becoming interested in and puzzled by qualia is a side-effect of sophisticated biological mechanisms. In cases such as the problem of qualia, we think that the problem needs not a solution, but a resolution. We do not offer a direct answer, and we do not say the problem is nonsensical as some positivist philosophers would. Such a resolution is provided by explaining the mental mechanisms that generate an interest in thinking and talking about qualia. Robots with our information processing architecture would do the same. The more intelligent ones should accept our

explanation of how that happens.

9.2 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 in 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.

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 the nature of mind. A meta-management system can develop its own conceptual framework for categorising its own internal states and processes, which may have features which it cannot possibly communicate (reliably) to others. This could lead robot philosophers to raise unanswerable questions about whether others have the same experiences as they do.

We take a different approach: When a question has no answer because it is based on confusions, the next best thing to giving an *answer* is to present a *theory* of the architectures and mechanisms which lead, in humans, to the question being formulated, and would also do in machines with a similar information-processing architecture. But it appears to be much easier to persuade six blind men that they are feeling a small part of a much larger object.

9.3 Zombies

When people feel that the kind of explanation being offered here cannot suffice since they are *convinced* that something is left out, this often takes the form of the 'zombie' argument: a robot could have all the information processing capabilities described here 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 this is based on a confusion between:

1. a robot having all the *externally observable* behaviours humans have, without being conscious, and
2. a robot having all the *internal* information processing capabilities humans have, without being conscious.

Case 2 is hard to understand even if you are familiar with the design and operation of virtual machines and much harder if you are not! When internal processing of a human-like virtual machine is described *in great detail*, including the meta-management abilities involved in thinking about qualia, 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.

9.4 Are we committed to “computationalism”?

It is important to distinguish two questions:

1. Is any information processing virtual machine architecture 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. In the next century 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. Finding out the answers requires us first to clarify meanings of the questions and the available answers. We know no better method than the method outlined here.

9.5 The causation problem: Epiphenomenalism

A problem not discussed here is how it is possible for events in virtual machines to have causal powers. It is sometimes argued that since (by hypothesis) virtual machines are fully implemented in physical machines, the only causes really operating are the physical ones. This leads to the conclusion that virtual machines and their contents are “*epiphenomenal*”, i.e. lacking causal powers. If correct, this would imply that if mental phenomena are states,

processes or events in virtual information processing machines, then mental phenomena (e.g. desires, decisions) have no causal powers.

A similar argument, if correct, would refute many assumptions of everyday life, e.g. ignorance can cause poverty, poverty can cause crime, etc. Dealing with this issue requires a deep analysis of the notion of ‘cause’, probably the hardest unsolved problem in philosophy. In the absence of a full account of causation, we can say this: virtual machine states are only as epiphenomenal as any other state existing on a level of description higher than the lowest level of physics (if any such level exists). So on our account, mental states will be just as real as anything else: molecules, trees, interest rates, justice, poverty, war, etc.

9.6 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 (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.

10 Acknowledgements

This work is partly funded by grant F/94/BW from the Leverhulme Trust, for research on ‘Evolvable virtual information processing architectures for human-like minds’. The ideas presented here were inspired by work of many well known philosophers, and developed with the help of Matthias Scheutz and also many students, colleagues and friends, including Margaret Boden, Pat Hayes, Steve Allen, John Barnden, Luc Beaudoin, Catriona Kennedy, Brian Logan, Riccardo Poli and Ian Wright. We have also learnt much from the empirical work of colleagues in the School of Psychology at the University of Birmingham, including Glyn Humphreys, Jane Riddoch and Alan Wing.

References

Albus, J. (1981), *Brains, Behaviour and Robotics*, Byte Books, McGraw Hill, Peterborough, N.H.

- Beaudoin, L. (1994), Goal processing in autonomous agents, PhD thesis, School of Computer Science, The University of Birmingham. (Available at <http://www.cs.bham.ac.uk/research/cogaff/>).
- Block, N. (1996), 'What is functionalism?'. <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionality.html>, (Originally in *The Encyclopedia of Philosophy Supplement*, Macmillan, 1996).
- Boden, M. A. (1990), *The Creative Mind: Myths and Mechanisms*, Weidenfeld & Nicolson, London.
- Brooks, R. A. (1991), 'Intelligence without representation', *Artificial Intelligence* **47**, 139–159.
- Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, New York, Oxford.
- Cohen, L. (1962), *The diversity of meaning*, Methuen & Co Ltd, London.
- Cooper, R. & Shallice, T. (2000), 'Contention scheduling and the control of routine activities', *Cognitive Neuropsychology* **17**(4), 297–338.
- Damasio, A. (1994), *Descartes' Error, Emotion Reason and the Human Brain*, Grosset/Putnam Books, New York.
- Gibson, J. (1986), *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, Hillsdale, NJ. (originally published in 1979).
- Goodale, M. & Milner, A. (1992), 'Separate visual pathways for perception and action', *Trends in Neurosciences* **15**(1), 20–25.
- Hauser, M. (2001), *Wild Minds: What Animals Really Think*, Penguin, London.
- Kim, J. (1998), *Mind in a Physical World*, MIT Press, Cambridge, Mass.
- LeDoux, J. (1996), *The Emotional Brain*, Simon & Schuster, New York.
- Maynard Smith, J. & Szathmáry, E. (1999), *The Origins of Life: From the Birth of Life to the Origin of Language*, Oxford University Press, Oxford.
- Minsky, M. L. (1987), *The Society of Mind*, William Heinemann Ltd., London.
- Nagel, T. (1981), What is it like to be a bat, in D. Hofstadter & D.C.Dennett, eds, 'The mind's I: Fantasies and Reflections on Self and Soul', Penguin Books, pp. 391–403.
- Nilsson, N. (1994), 'Teleo-reactive programs for agent control', *Journal of Artificial Intelligence Research* **1**, 139–158.
- Nilsson, N. (1998), *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, San Francisco.
- Popper, K. (1934), *The logic of scientific discovery*, Routledge, London.

- Ryle, G. (1949), *The Concept of Mind*, Hutchinson, London.
- 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. & Sloman, A. (2001), *Affect and Agent Control: Experiments with Simple Affective States.*, in N. Z. *et al*, ed., 'Intelligent Agent Technology: Research and Development', World Scientific Publisher, New Jersey, pp. 200–209.
- Sloman, A. (1989), 'On designing a visual system (Towards a Gibsonian computational model of vision)', *Journal of Experimental and Theoretical AI* **1**(4), 289–337.
- Sloman, A. (1993), *The mind as a control system*, in C. Hookway & D. Peterson, eds, 'Philosophy and the Cognitive Sciences', Cambridge University Press, Cambridge, UK, pp. 69–110.
- Sloman, A. (1996), *Actual possibilities*, in L. Aiello & S. Shapiro, eds, 'Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)', Morgan Kaufmann Publishers, Boston, MA, pp. 627–638.
- Sloman, A. (2000), *Interacting trajectories in design space and niche space: A philosopher speculates about evolution*, in M. et al., ed., 'Parallel Problem Solving from Nature – PPSN VI', Lecture Notes in Computer Science, No 1917, Springer-Verlag, Berlin, pp. 3–16.
- Sloman, A. (2001), *Varieties of Affect and the CogAff Architecture Schema*, in C. Johnson, ed., 'Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention', York, pp. 39–48.
- Sloman, A. (to-appear), *Architecture-based conceptions of mind*, in 'Proceedings 11th International Congress of Logic, Methodology and Philosophy of Science', Kluwer, Dordrecht, pp. 397–421. (Synthese Library Series).
- Sloman, A. & Logan, B. (2000), *Evolvable architectures for human-like minds*, in G. Hatano, N. Okada & H. Tanabe, eds, 'Affective Minds', Elsevier, Amsterdam, pp. 169–181.
- Waismann, F. (1965), *The Principles of Linguistic Philosophy*, Macmillan, London.
- Weiskrantz, L. (1997), *Consciousness Lost and Found*, Oxford University Press, New York, Oxford.
- Winograd, T. (1972), 'Procedures as a Representation for Data in a Computer Program for Understanding Natural Language', *Cognitive Psychology*. (Later published as a book *Understanding Natural Language*, Academic Press, 1972).
- Wittgenstein, L. (1953), *Philosophical Investigations*, Blackwell, Oxford. (2nd edition 1958).
- Wright, I., Sloman, A. & Beaudoin, L. (1996), 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology* **3**(2), 101–126. Repr. in R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000.