Invited talk: *11th International Congress of Logic, Methodology and Philosophy of Science,*
Cracow, Poland 1999, August 20-26 Section 12. Philosophy of Cognitive Sciences and Artificial Intelligence.
in *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II),* Synthese Library Vol. 316,
Eds. P. Gärdenfors, K. Kijania-Placek & J. Woleński Kluwer, Dordrecht, pp 403–427, 2002.

Architecture-Based Conceptions of Mind

Aaron Sloman
School of Computer Science, University of Birmingham
Birmingham, B15 2TT, UK
A.Sloman@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~axs/

Abstract

It is argued that our ordinary concepts of mind are both implicitly based on architectural presuppositions and also cluster concepts. By showing that different information processing architectures support different classes of possible concepts, and that cluster concepts have inherent indeterminacy that can be reduced in different ways for different purposes we point the way to a research programme that promises important conceptual clarification in disciplines concerned with what minds are, how they evolved, how they can go wrong, and how new types can be made, e.g. philosophy, neuroscience, psychology, biology and artificial intelligence.

1 Introduction

We seem to have direct access to mental phenomena, including thoughts, desires, emotions and, above all our own consciousness. This familiarity leads many people to believe they know exactly what they are talking about when they engage in debates about the nature of mind, and refer to consciousness, experience, awareness, the ‘first-person viewpoint’, and so on.

However, this conviction is at odds with the diversity of opinions expressed about the nature of the phenomena, and especially the widely differing definitions offered by various types of psychologists, cognitive scientists, brain scientists, AI theorists and philosophers, when they attempt to define concepts like ‘emotion’ and ‘consciousness’.

The confusion has several roots, one of which is the *hidden* complexity of both the phenomena and the architectural presuppositions we unwittingly make when we use such concepts.

Another is the common error of believing that we have a clear understanding of concepts just because they refer to phenomena that we experience directly. This is as mistaken as thinking we fully understand what simultaneity is simply because we have direct experience of seeing a flash and hearing a bang simultaneously. Einstein taught us otherwise. That we can recognise some instances of a concept does not imply that we know what is meant *in general* by saying that something is or is not an instance. Endless debates about where to draw boundaries are a symptom that our concepts are confused, whether the debates are about which animals have consciousness, whether machines can be conscious, whether unborn infants have experiences, or whether certain seriously brain-damaged humans still have minds.

Such questions cannot be resolved by empirical research when there is so much disagreement about what sort of evidence is relevant. Does wincing behaviour in a foetus prove that it feels pain and is therefore conscious, or is it a mere physiological reaction? How can we decide? Does a particular type of neural structure prove that the foetus (or some other animal) is conscious, or is the link between physical mechanisms and consciousness too tenuous to prove anything?

This paper shows how the hidden complexity of our concepts and the phenomena they refer to explain why there is so much confusion and disagreement and indicates how we can begin to make progress beyond sterile debates.

Many of our concepts are implicitly architecture-based and different thinkers attend to different aspects of the architecture. They are also ‘cluster concepts’, referring to ill-defined clusters of capabilities supported by the architecture, and different views favouring different clusters. If we understand this we can see how to define different families of more precise concepts, on the basis of which answerable questions can be formulated. Which definitions are *correct* is a pointless question.

2 Architecture-based concepts

We can deepen our understanding of these concepts, and, where necessary, repair their deficiencies, by seeking an explanatory theory which accounts for as many phenomena as possible and then use it as a framework for systematically generating concepts. A common error is believing that we have to define our concepts before we seek explanatory theories. Typically it is only after we have a theory that we can understand the concepts describing the phenomena to be explained. So it is to be expected that we shall not be able to give good definitions of most of our mental concepts until we have good explanatory theories.

This does not imply that our pre-theoretical concepts are completely wrong. Our existing concepts of mind work well enough for ordinary conversational purposes (e.g. when we ask ‘When did he regain consciousness?’, ‘Are you still angry with me?’, etc.). So a good theory of the architecture underlying mental states and processes should generate concepts which *extend* and *refine* our previous concepts, rather than replacing or eliminating them.

New theories of the sub-atomic architecture of matter extended and revised our concepts of kinds of elements, kinds of chemical compounds, and kinds of physical and chemical processes. We still talk about iron, carbon, water, etc., though we also now know about isotopes and new sorts of elements and compounds, and many new kinds of processes involving previously known kinds of physical stuff. We still talk about solids, liquids and gases though we also know about other states of matter supported by the architecture.

2.1 Architecture-based cluster-concepts

Muddles in our pre-theoretical concepts of mind surface when we try to ask philosophical or scientific questions, e.g. ‘How did consciousness evolve?’ ‘What are its neural correlates?’ ‘Which animals have it?’ What we normally refer to as consciousness involves the exercise of a large, diverse, ill-defined cluster of capabilities (many of them unconscious!) supported by our information processing architectures. If there is no well-defined subset of capabilities which are necessary or sufficient for consciousness, then some of our apparently meaningful questions, like many questions involving cluster concepts, may be ill-defined. Many mental concepts share this semantic indeterminacy, e.g. ‘emotion’, ‘intelligence’, ‘understanding’, ‘pleasure’, etc.

The idea that there are cluster concepts, that various kinds of indeterminacy or, what has been called *open texture*, pervades ordinary language is very old, e.g. in the writings of Wittgenstein (1953), Waismann (1965) and many others. I shall attempt to explain how it comes about that ordinary mental concepts have that feature, and what to do about it.

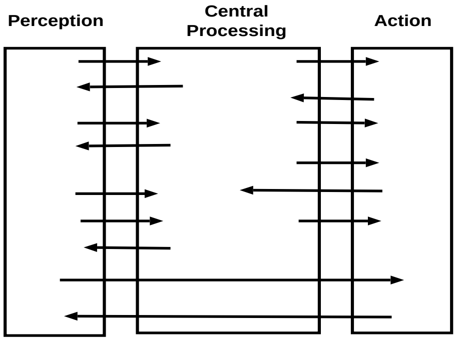

Figure 1: (a)

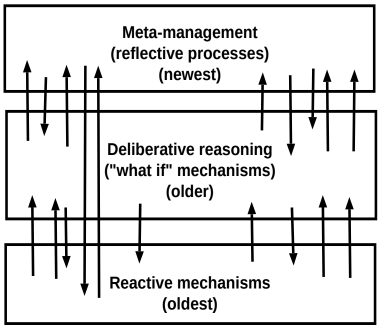

Figure 1: (b)

Two ways of partitioning an organism's architecture. (a) shows information flow through the system via sensory, central and motor mechanisms, where each 'tower' may comprise simple transducers or may have extremely sophisticated multi-functional layered mechanisms. (b) indicates differences in processing layers in the architecture which evolved at different times. Reactive mechanisms are highly parallel, digital or analog, and respond automatically to triggering conditions. Their actions may be internal or external. Deliberative mechanisms can do 'what if' reasoning, considering possible futures and forming plans. A meta-management layer can monitor, evaluate, and to some extent redirect processes in other mechanisms. Processing in different sub-systems may be concurrent and asynchronous. Higher levels need not always dominate lower levels. E.g. in 'tertiary emotions' the third layer partly loses control of attention. Some animals may have only the bottom or bottom and middle layer.

2.2 Multiple architectures generate multiple families of concepts

The analogy suggested above between the way theories of the architecture of matter extend and refine ordinary concepts of kinds of stuff and the way a new theory of the architecture of mind could illuminate concepts of mentality, is only partial, because there is only one physical reality and one architecture for physical matter (although it may have many levels of abstraction), whereas there are many kinds of minds with different architectures.

Figures 1 (a) and (b) illustrate two typical architectural decompositions of an intelligent organism, software system, or robot. Figures 2 combines the two views and add further detail. Figures 3(a) and (b) elaborate further. Organisms with simpler architectures have fewer architectural layers, and simpler perceptual or motor subsystems. They would then support simpler collections of processes, and different concepts would be applicable to them. If insects have only the reactive layer, that will limit the set of concepts applicable to them. Awareness in a flea and awareness in the chimpanzee bitten by the flea are very different.

By contrast, there is one physical universe and its unique architecture generates a *unique* collection of concepts of types of stuff and types of processes involving matter. Of course, different concepts are relevant to different levels of complexity, e.g. sub-atomic physics, chemistry, mechanics, cosmology, etc., But they all refer to the same physical architecture described at different emergent levels. Physicists may disagree as to what the unique architecture of physical reality is, just as two ethologists may disagree about the architecture of a gorilla's mind, but that is not like ethologists studying the minds of many distinct species.

Different kinds of minds have different architectures which will each generate *separate* families of concepts for types of mental states and processes. We should not expect that concepts relevant to describing mental states in an adult human are applicable to a simple robot, a software agent in a computer game, a new born human infant, or a bat (Nagel 1981).

Researchers with different foci of interest can be expected to produce different definitions. In some cases they are studying different systems, e.g. when a brain-scientist studying rats discusses emotions with a social psychologist studying adult humans. Rats and adult humans do not necessarily have the same architecture, though there will be many similar sub-mechanisms, on account

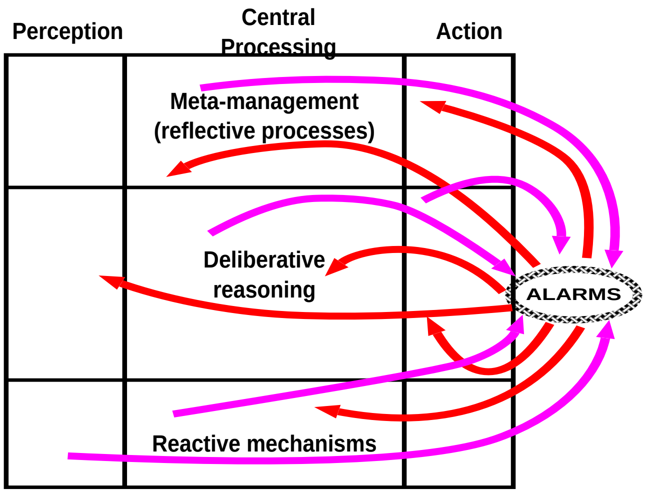

Figure 2:
This figure combines the triple tower and triple layer views, and adds a global alarm system, receiving inputs from all the main components of the system and capable of sending control signals to all the components. The various components have functions defined by their relationships with other parts of the system. Perceptual and motor towers have evolved different processing layers, operating at different levels of abstraction, and feeding into and controlled by the different central layers. The alarm mechanism is an extension of the reactive layer. Since an alarm system needs to operate quickly when there are impending dangers or short-lived opportunities, it cannot use elaborate inferencing mechanisms, and must be pattern based. Global alarm mechanisms are likely therefore to make mistakes at times, though they may be trainable.

of our common evolutionary history and some common needs. Rats may be lacking in the meta-management layer and therefore in types of self awareness and self control. So researchers may (possibly unwittingly) use different concepts with different presuppositions.

Even when two scientists study humans, they may have different research interests which cause them to attend to phenomena generated by different aspects of the architecture. They may be unaware of this, because they are studying different parts of a system that is more complex than any of them realise, like the proverbial collection of blind people each trying to say what an elephant is on the basis of what they individually can feel. It is not hard to convince a blind man that he is in contact with only a small region of a large structure. It is much harder to convince people producing theories of mind that they are attending to a tiny part of a huge system. The rest of this paper attempts to explain what sort of system.

3 Information-processing architectures

Scientists study different kinds of machines: e.g. machines that manipulate matter (e.g. by rearranging it geometrically or transforming its state, for instance by cooking it), machines that manipulate energy (storing, transforming, using it), and machines that manipulate information (acquiring, storing, transforming, interpreting, deriving, communicating and using it). The same physical object can simultaneously implement more than one type of machine.

Machines need not be artefacts: all these sorts of machines have existed in various natural forms for millions of years. The sun is a machine which (among other things) transforms mass energy into electromagnetic energy, and all living organisms are more or less sophisticated information processing machines. Individual organisms use information about the environment in order to obtain nourishment, avoid harm, find mates, etc. and their reproductive mechanisms manipulate and use information about the ‘design’ of the organism.

The study of machines that manipulate matter and energy is very old, but only recently have we begun to understand information processing machines. Our understanding is still very limited. E.g.

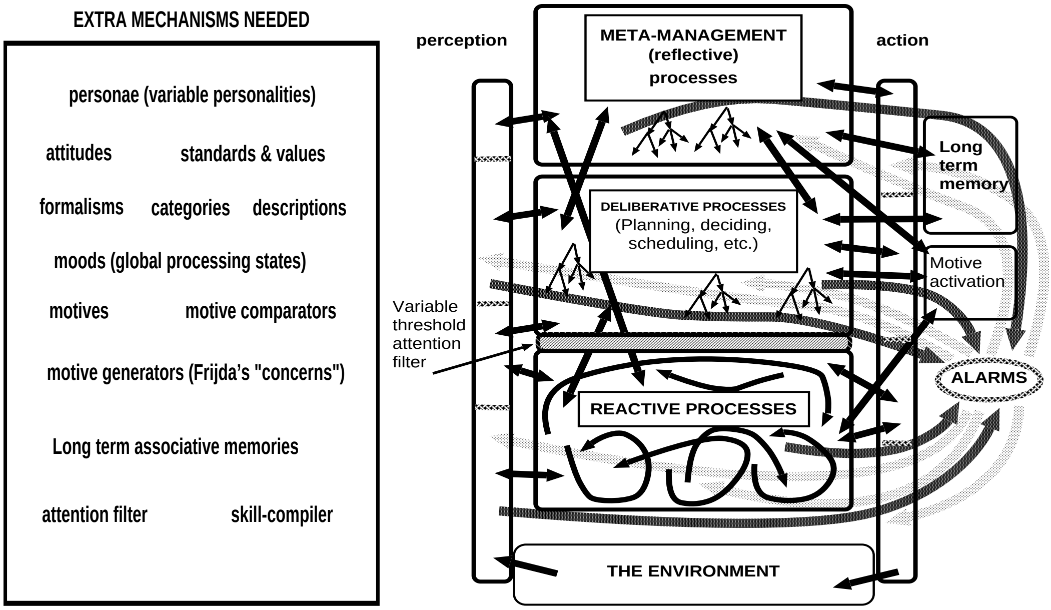
EXTRA MECHANISMS NEEDED
personae (variable personalities)
attitudes
standards & values
formalisms
categories
descriptions
moods (global processing states)
motives
motive comparators
motive generators (Frijda's "concerns")
Long term associative memories
attention filter
skill-compiler
perception
META-MANAGEMENT (reflective) processes
action
Long term memory
DELIBERATIVE PROCESSES (Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Figure 3: (a) Figure 3: (b)

In (a) we list some additional components required to support processing of motives, 'what if' reasoning capabilities in the deliberative layer, and aspects of self-control. It is conjectured that there is a store of different, culturally influenced, 'personae' which take control of the top layer at different times, e.g. when a person is at home with family, when driving a car, when interacting with subordinates in the office, in the pub with friends, etc. In (b) relations between some of the components are shown along with a global alarm system, receiving inputs from everywhere and sending interrupt and redirection signals everywhere. It also shows a variable-threshold interrupt filter, which partly protects resource-limited deliberative and reflective processes from excessive diversion and redirection.

it is not easy to define 'information' – another partly indeterminate cluster concept. Nevertheless, in the last half century we have learnt about some of the variety of types of information processing machines and their architectures.

3.1 Virtual machines

Machines are generally complex systems with an architecture, which is defined by the number and types of interacting components and their contributions to the capabilities of the larger system. The components may also be machines with their own architectures.

Some machines are physical systems, whose architecture involves a physical decomposition into spatially distinct parts, e.g. the architecture of a building, a bulldozer or a bottle-filling machine. However, software engineers have built many information processing machines that are not physical machines but *virtual* machines. Familiar examples include chess playing programs, word processors, email systems, operating systems, compilers, the prolog virtual machine which operates on lists, predicates, variables numbers, strings, etc.

Virtual machines are implemented in (or, in the jargon of philosophers, supervenient on) physical machines, but do not have any simple relationship with those physical machines (Sloman 1994*b*). For instance, in a virtual machine in a computer, component A can be part of B while B is part of A, which is impossible for physical machines. The architectural decomposition of a virtual machine, or of an information structure it creates, need not be tree-like: it can include containment loops, unlike the structural decomposition of physical systems.

There does not have to be any simple correlation or mapping between components or events in virtual machines and components or events in the physical implementation machine. The same virtual machine could be implemented on very different physical machines. A virtual machine known to us today might have totally unexpected physical implementations in the future. Even within a single machine, scheduling and memory management can cause the relationships between virtual and physical entities to change frequently. A certain sequence of bit-manipulation events in the CPU could implement quite different processes in different virtual machines, depending on the context in which it occurs. (This sort of thing could make a nonsense of the search for so-called ‘Neural Correlates of Consciousness’.)

The expression ‘physical symbol system’ often used in discussing computational models of mind (Newell 1980) is very misleading, since almost all the symbols of interest will not be *physical* symbols even if they are implemented in physical machines. It would be more accurate to refer to ‘physically implemented symbol systems’.

The internal processing in a virtual machine may be far richer than the externally visible behaviour, partly because of the limited bandwidth of the expressive mechanisms available and partly because most internal processes have no direct external connections. This defeats all behaviourist analyses of mental concepts and also functionalist analyses that require mental phenomena ultimately to be analysed in terms of input-output dispositions.

Another common mistake is thinking that everything in an information processing system is implemented entirely in the physical machine which contains the system, implying, for instance, that all human mental states are implemented in brain states. Mental states in a person and information processes in a computer have semantic content, and can refer to objects in the environment, e.g. London, or the Eiffel tower. The semantic content of symbols used in such cases is only *partly* determined by internal mechanisms. In part, as Strawson (1959) showed, the content is determined, by the agent’s spatial and causal relationships to the referent.

So, part of the implementation of such mental processes may be in the *environment*, unlike simple cases such as a Prolog virtual machine which refers only to internal operations and symbols, unless there are portions referring to file systems, or input/output devices. If an email system or internet browser can refer to remote entities then it will be partly implemented in the mechanisms linking it to those entities. However, simple email programs have no knowledge of other sites, and will need to interact with other programs that include richer semantic content.

3.2 Virtual machines have real causal powers

Virtual machines can include non-physical states, processes and causal relations that are as real as poverty, crime, economic inflation, inheritance of property, and acquisition of voting rights, all of which have real causal powers despite being non-physical phenomena. Anyone who thinks poverty cannot cause theft insofar as poverty and theft are not physical states, will have a hard time solving social problems. These are not physical phenomena: they don’t obey or violate laws of physics because their ontology is not describable using concepts of physics.

It is often argued that if the underlying physics is causally complete, virtual machine events are ‘epiphenomenal’ and cannot have any physical effects. This argument is implicitly based on a model of causation which assumes that causality is some sort of fluid obeying conservation laws. If causal powers are analysed in terms of truth of various sorts of conditionals, then causal completeness at a physical level does not rule out other kinds of causation being real: there can be alternative sets of sufficient conditions for something to happen. Some of the components of these alternative sets may also be necessary or sufficient for components of the other sufficient sets. E.g., some physical events may necessitate some virtual machine events, and vice versa.

The physical design of a pentium or sparc processor ensures that when the physical machine is switched on and working a pentium or sparc *virtual* machine exists. In such a design virtual machine events (e.g. addition of two numbers, fetching contents of memory, or invocation of a procedure) can cause physical as well as virtual machine events to occur. The relationship between physical and virtual levels, however, is not at all easy to analyse (Chalmers 1996; Scheutz 1999a, 1999b). The connection between the physical machine and the virtual machine is not logical or mathematical since the concepts needed to specify the virtual machine and its laws of behaviour are not definable in terms of those of physics. For instance, ‘checkmate’ is not a physical concept and is not definable in terms of physical concepts, and the laws of chess are not derivable from those of physics. Chess could change while physics remained the same.

Something like the ‘conserved fluid’ model of causation leads many to assume that only *physical* events in the computer can ‘really’ cause physical events. But tangled circular causal relationships are typical of relations between virtual machines in computers and lower level virtual machines or electronic machines in which they are implemented. We should not be surprised if biological evolution produced similar tangles in the relationship between minds and brains.

There is still much philosophical work to be done: our conceptual tools need sharpening. Later we may be in a better position to explain some of the phenomena of supervenience, generally discussed by philosophers as if virtual machines in computers had never been invented.

3.3 Virtual machines: summary

(a) Functions in a virtual machine, such as deriving new information, making a plan, taking a decision, choosing a chess move, or multiplying two numbers are not physical functions.
(b) Virtual machine events and processes require a physical infrastructure. Some of the processes are implemented within the body, but not all, for instance when a true belief about a remote object becomes false without the agent knowing.
(c) Causal relationships can hold between virtual machine events.
(d) They can hold in either direction between physical and virtual machine events.
(e) This does not presuppose causal incompleteness at the physical level.
(f) The (software) engineering concept of ‘implementation’ and the philosophical concept of ‘supervenience’ are closely related.
(g) Sometimes the existence of a (working and switched on) physical machine guarantees the existence of a (working) virtual machine of a certain sort, and this is not just a matter of an arbitrary interpretation of the physical processes.
(h) Although physical states, events, and processes can determine virtual machine phenomena, the latter are not logically or mathematically derivable from the former, if the concepts required to specify the virtual machine are not definable in terms of those of physics, and their laws of behaviour are not logically derivable from those of physics (even when supplemented with physical descriptions of the implementation machine).

4 Minds are information processing systems

If we are looking for a good explanatory theory we need to be sure which phenomena are to be explained, and we need to have good knowledge of types of architectures and mechanisms capable of explaining various kinds of phenomena. Otherwise we are likely to come up with over-simple theories.

Humans, like other organisms, need information about themselves and their environment in order to select between alternative possible behaviours. Sometimes the information is used as soon as it is acquired, and sometimes it is stored for later use. Sometimes it is used in more or less the form in which it was acquired, whereas in other cases information is abstracted, combined, used to derive new information, etc. via processes of varying sophistication and length. Some of the information (e.g. gained by senses or body monitors) is about what is the case. Some of it is about what to do or avoid. Some of it is control information, which when activated produces some kind of behaviour, e.g. blinking or abandoning a problem.

Mental states and processes involve both actual processing of information, and also, as Gilbert Ryle pointed out in 1949, *dispositions* to process information.

On a narrow concept of functionalism, all the information processed is assumed to be about the environment and the dispositions are assumed to be concerned with external behaviour in that environment. However as any software engineer knows, information processing systems often acquire, interpret, store or use information about their own *internal* states (both hardware states and virtual machine states). Many of their actions are internal, for instance finding edge features in an image received via a camera, or changing the priority of processes handled by a scheduler, or modifying a stored plan after discovering that under certain conditions it gets stuck in a loop. In all these cases the action involves detecting and reacting to features of internal virtual machine structures and either modifying them or creating new ones, some of which will cause further internal processing to occur.

Where there are mechanisms (e.g. programs, neural nets) that are ready to react to certain internal conditions by changing something internal, these constitute dispositions to produce internal behaviour. Complex software systems can also support dispositions to produce new dispositions or to modify old ones. So if minds are information processing systems we can expect to find networks of dispositions within them, many with little or no direct connection to external behaviour, e.g. when the architecture involves different concurrently active layers operating on information at different levels of abstraction, as depicted in the figures. The diagrams above schematically present outlines of a family of types of architectures that might account for many aspects of mental processes in humans, other animals, and future robots and software agents. Detailed investigation of such families of architectures and the systems of concepts which they support is a long term project.

I doubt that Ryle knew much about computer programs in 1949, but what he wrote about the nature of mind (unlike what many other philosophers write) fits the general characteristics of sophisticated information processing systems very well, including his acknowledgement of the richness of 'internal' behaviour, for instance in his chapter on imagination.

Unfortunately, many who write about the mind body-problem, about supervenience, about consciousness, etc. are unfamiliar with the variety of types of software architectures, or the variety of relationships between software virtual machines and the underlying hardware.

4.1 What are we trying to explain?

There is a more subtle deficit in many discussions about minds and consciousness: namely insofar as they say anything about what needs to be explained what they say is often far too vague and general. For example, we need far more detailed analyses of the phenomenology of various types of visual and other sensory experiences instead of coarse-grained characterisations such as 'experiencing a circular red patch'. There are very many different ways of experiencing a red patch, depending on which spatial information processing capabilities are available. An organism that lacks mechanisms for detecting and using symmetry will not experience the circular patch as symmetrical even if it is.

Similarly a human can experience the patch as a spatial region across which something *could* move, even though nothing is actually moving. Whether other organisms, or new-born human infants can do that is not at all obvious. That ability to experience spatial structures as inherently susceptible of changes of various kinds is part of the infrastructure of human intelligence. It is used by someone who looks at an unfamiliar window catch and attempts to understand how to open the window. It is part of the ability to plan spatial actions in advance of doing them. It is relevant to predicting the actions or movements of other things in the environment. It is used in playing board games which require the ability to think about possible changes in the configuration of pieces. It is also deeply implicated in mathematical reasoning about spatial structures (e.g. (Jamnik, Bundy & Green 1999)). Without this ability Euclid's geometry would have been impossible.

Many AI researchers investigating visual capabilities or spatial reasoning abilities assume that they have to be implemented in mechanisms that can manipulate 2-D arrays of some kind, representing information about the contents of the visual field (an internal information structure) and viewpoint-relative information about the external environment in registration with it. For examples, see (Narayanan 1993, Glasgow, Narayanan & Chandrasekaran 1995). However it is not obvious that these forms of representation and image manipulation adequately support the perception and use of affordances and more generally the grasp of possibilities for change inherent in all spatial configurations (Gibson 1986, Sloman 1989, Sloman 1996).

This is a topic requiring much further research. Many clues can come from surprising combinations of capabilities found as a result of brain damage which reveals skills we did not realise were part of our ordinary experience until we find people deprived of them. Often there is knowledge available, but ignored by researchers. For instance many philosophers and psychologists think that a necessary feature of having emotions is being aware of them, yet novelists, playwrights and gossips know well that a person can be angry, infatuated, or jealous without being aware of the fact, even when it is obvious to others.

A richer and more accurate catalogue of capabilities displayed in human mental processes could guide the search for explanatory theories about mechanisms they require. We can then investigate empirically which organisms have them, and be able to explain which types of robots will have them.

Until then we can expect endless debates about whether minds could be implemented in computers or whether brains need quantum mechanisms to support consciousness. Most arguments on either side of such debates are at a gross level of abstraction and fail to explain specific detailed phenomena, such as a geometer's ability to find interesting geometrical proofs or a child's ability to build things out of meccano components, or your ability to have an itch or find a story funny.

5 Towards an architecture for adult human minds

My colleagues and I are exploring properties of a family of virtual machine architectures that make use of (at least) three different concurrently active architectural layers (Figures 2 and 3) which probably evolved at very different times, which we share with some other animals to varying degrees, and which, along with additional supporting modules, account for different cognitive and affective states and processes. We refer to these layers as reactive, deliberative and meta-management (or reflective) layers. In Albus (1981, Ch 7) a related view of the architecture is described in terms of reptilian, old mammalian and new mammalian layers. Layered architectures are commonplace in AI, e.g. (Nilsson 1998), though different researchers propose different numbers of layers, and differ in how they construe the relationships between layers, e.g. whether information flows serially up and down the hierarchy (as Albus and Nilsson seem to suggest) or whether processing in different layers is concurrent as in our model and (Brooks 1991).

Although all the conjectured layers are ultimately implemented in reactive mechanisms, they have different processing capabilities, as indicated in the figure captions, and require different internal architectures and connections to other layers, along with different supporting mechanisms (Fig. 3(a)). All the layers are subject to interference from the others and from one or more fast but stupid partly trainable 'global alarm' mechanisms, which can rapidly redirect processing when triggered by patterns indicating actual or potential threats or opportunities.

Our previous work uses this architectural framework to distinguish three categories of processes which appear to correspond to familiar types of emotions. The first two labelled 'primary' and 'secondary', emotions were described in (Damasio 1994, Goleman 1996, Picard 1997). The third category 'tertiary' emotions was originally suggested by familiar human experiences such as grief or excited anticipation.

Primary emotions, such as many kinds of fear responses, require only a reactive layer, possibly with an alarm system which can take global control under certain circumstances. Secondary emotions, such as apprehension and relief, require in addition a deliberative layer, with mechanisms supporting 'what if' reasoning capabilities. These internal processes can trigger disturbances without anything perceived being responsible. Some sub-classes of secondary emotions include external and peripheral bodily changes, whereas others do not. Tertiary emotions require the third layer performing meta-management tasks such as monitoring, evaluating and possibly redirecting some of the reactive and deliberative processes. Certain forms of disruption of these high level control mechanisms, for instance in grief, discussed at length in (Wright, Sloman & Beaudoin 1996) produce tertiary emotions. Other examples are infatuation, humiliation, and thrilled anticipation.

The third layer, which we suggest can 'adopt' different 'personalities' for controlling thinking and behaviour in different contexts, seems also to be crucial to absorption of a culture and various kinds of mathematical, philosophical and scientific thinking, all of which involve reflective capabilities.

The variable-threshold attention filter depicted somewhat misleadingly in Fig. 3(b) can explain how alarms and new signals from the reactive layer sometimes do and sometimes do not distract attention in the top two layers, depending on the urgency and importance of their current tasks. In intense emotional states, like grief and infatuation, the filter may fail to suppress unwanted internally generated distractions. Further details can be found in our previous work (Sloman & Croucher 1981, Beaudoin & Sloman 1993, Beaudoin 1994, Wright 1977, Sloman 1998*a*, Sloman 2000).

Further subdivisions can be made according to how the emotions develop, which secondary effects they have, which additional dispositions are present, which kinds of evaluations and emotions are involved, how the emotion subsides, whether or not the agent is aware of what is happening, and so on.

Having defined different classes of emotions in terms of the information processing mechanisms involved we can then see that organisms whose architecture does not include a meta-management layer cannot have tertiary emotions, and those which do not include a deliberative layer (insects?) cannot have secondary emotions.

As with emotions we can also define different kinds of sentience or awareness in terms of different sorts of architectural underpinnings. An organism with only a reactive architecture (insects?) will have a type of awareness of its environment and certain aspects of its own body state which is sufficient to trigger appropriate internal and external reactions. An organism or robot which also has deliberative mechanisms will be aware of possible future events and perhaps also possible past events (what might have been), though the extent of such awareness can differ widely, depending on the mechanisms available and the size and sophistication of the temporary workspace required. Evolution of the third layer provides organisms with an additional kind of self-awareness including the ability to classify and evaluate current internal states of many kinds. If it also has access to some of the internal databases used by perceptual mechanisms, that will produce sensory qualia.

If we define different sorts of consciousness in terms of the capabilities involved and the architectures they require, we can then replace interminable philosophical debates about which animals are conscious with a collection of different empirical questions which it may be possible (though difficult) to answer.

A multi-layered architecture of the sort proposed could give a robot many human-like mental states and processes, including qualia (when the meta-management layer 'attends' to particular kinds of internal phenomena). Such robots could fall into old philosophical confusions about consciousness and begin to wonder whether humans are conscious.

5.1 Virtual machine functionalism

As we learn more about animal brains and about engineering design requirements for sophisticated robots, these crude models and distinctions will turn out to be grossly over-simplified.

Nevertheless, considering descriptive and explanatory concepts generated by a *virtual machine information processing architecture* which has a physical implementation, leads to a deeper and more comprehensive explanatory theory than is normally found in scientific or philosophical studies of mind, partly because it can explain more of the fine structure of the phenomena than previous theories and partly because it allows more levels of analysis to be integrated, in something like the way that computing science produces integrated theories concerning many levels of virtual machine and their relationships to underlying digital systems and the physical hardware used to implement them.

Because virtual machine (VM) functionalism allows part of the implementation to be in the environment and does not claim that the physical implementation includes *only* the agent's body, it does not lead to solipsism and is immune against 'twin-earth' arguments, found in the writings of Putnam and others.

VM functionalism differs from most varieties of functionalism because it is not about input-output relationships: the vast majority of functions in an information processing virtual machine typically involve no externally observable behaviour. If they provide any biological advantage it must be very indirect.

There is also no requirement for mental mechanisms and the states and processes they produce all to be the result of evolutionary selection for their beneficial effects.

On the contrary, every complex design for a working system involves tradeoffs, and there are bound to be many aspects of the system which are merely side-effects, sometimes harmful side-effects, of mechanisms originally selected for other functions, or even mechanisms produced by random mutations, without contributing to fitness.

It is also easy for such systems to have software bugs, even when there is no physical damage or deficiency. If an individual organism has to bootstrap itself through interaction with a rich and complex environment after birth, it can easily get things wrong, including forming erroneous generalisations about the environment, acquiring harmful tastes and preferences, and learning far from optimal plans or algorithms. There is plenty of evidence for this, in the form of addictions and many kinds of self-harming behaviour in social misfits. Where several such agents interact, as in human society, the scope for positive feedback loops which cause disastrous consequences for all concerned is very evident through all human history.

A corollary of all this is that there is no requirement of *rationality* contrary to much philosophical writing (or requirements of Newell's 'Knowledge Level'). For instance most of the reactive mechanisms are not rational or irrational: they simply react.

5.2 Zombies

Architecture-based concepts generated in the framework of virtual machine functionalism can subvert familiar philosophical thought experiments about zombies. People who reject the functional analysis of familiar mental concepts often try to describe so-called zombies which have all the relevant functional dispositions but don't necessarily have what we normally think of as experience, qualia, etc.

However, attempts to specify in sufficient detail a zombie with *all* of the appropriate kinds of virtual machine functionality but nevertheless lacking human-like experiences, degenerate into incoherence when spelled out in great detail.

When you have *fully* described the internal states, processes, dispositions and causal interactions within a zombie whose information processing functions are alleged to be *exactly* like ours, the claim that something else might still be missing becomes incomprehensible.

Of course, those who simply do not wish to believe that humans are information processing systems for theological or ethical reasons will not be convinced by any of this. Neither will those who think they can tell by introspection that these arguments are wrong. Some future robots will fall into the same trap.

6 Varieties of minds

The information processing framework generates a rich field of theoretical and empirical research into many types of minds, and many families of mental concepts.

A common mistake in theorising about mind and consciousness is to consider only adult human minds (often implicitly restricted to the minds of academic researchers!), ignoring people from other cultures, infants, people with brain damage or disease, insects, birds, chimpanzees and other animals (Kohler 1927), as well as robots and software agents in synthetic environments. By broadening our view, we find evidence for diverse information processing architectures, each supporting and explaining a specific combination of mental capabilities.

We can be sure that future designs for artificial systems, possibly based on new types of physical hardware for information processing engines, will contain many surprises and stretch our ideas about the sorts of architectures that we should be looking for in natural systems.

However it is already clear that there are different classes of mental capabilities in different animals, different people, and the same person at different stages of development, from infancy to old age. Since different architectures provide different collections of capabilities, different sets of concepts will be applicable to them. If a computing system lacks a file manager then one cannot ask what levels of file protections it provides or how efficiently it defragments its filestore. We cannot ask how many interrupt priority levels there are in an operating system which always runs every program to completion once begun.

Concepts describing mental states and processes in one animal or machine may be inappropriate when describing another. Likewise, concepts relevant to normal adult humans may be inappropriate not only for describing bat minds, but also for describing mental states of new-born infants or victims of Alzheimer's disease.

The information processing mechanisms in a new-born child probably render it totally incapable of wondering how long it has existed, considering whether there is evidence for other minds, feeling humiliated or having a strong ambition to be elected president.

If the architecture of a new born infant is still in a very primitive state, ready to begin 'bootstrapping' itself through interaction with its environment, then most mental states of such an infant are not describable in the language we normally use to talk about adults, and *vice versa*. For instance, if a human infant, or an ant, lacks the architectural underpinnings of what we normally call 'suffering' then it may be incapable of suffering, or even of being contented or happy.

This suggestion may seem surprising, and even obnoxious. But at which stage after fertilisation of an egg does suffering, or pleasure, become possible? If there is no clear answer it may be because we don't really know what question we are asking.

That could be remedied by clarifying what sorts of suffering, or pleasure, are possible in various sorts of architectures, in various stages of development.

6.1 Towards answerable questions

We can expect to find that there are interestingly different types of minds, based on different information processing architectures. In that case we can investigate the *empirical* question whether neonates typically have architectures of type A1, A2, or A3, and how these develop into architectures of type B1, B2, B3, etc.

We could formulate new empirical questions, such as whether infants, or children at a certain stage of development are capable of having pain of type P5, or P27, or P34, etc., instead of endlessly debating the totally indeterminate question: 'Can a new born (or even unborn) infant experience pain?' where the debates are fruitless because different people (often with different ethical or religious agendas) choose to regard different kinds of evidence as relevant, for instance whether a foetus moves in response to noise, or whether it has certain neural structures.

When we have specified which concepts are supported by a type of information processing architecture it becomes an empirical question whether a particular organism or machine has that architecture, so that the concepts are applicable to it.

For instance, primary, secondary and tertiary emotions require different architectural layers, as explained above. Whether an animal does or does not have a particular layer is then an empirical question which can be used as part of the evidence whether it is capable of types of emotions or other mental states that depend on the layer.

Such questions may be hard to answer, since, as discussed previously, neither the underlying physical architecture nor the observable behaviour will always provide definitive clues regarding the virtual machine architecture. So the methodology of checking hypotheses about the virtual machine architecture in an organism, and therefore about which mental states it can have, is very subtle and complex, and there are no guaranteed tests for correctness.

That, however, is no different in principle from any other deep scientific research problem. Only in shallow correlational science can hypotheses be checked easily by doing experiments (Sloman 1978, Ch 2).

7 Mapping design space

A proper understanding of all this requires comparative analysis in design space. We understand a particular architecture better if we know what differences would arise out of various sorts of design changes: which capabilities would be lost and which would be added. I.e. understanding any *particular* sort of architecture requires comparison with a variety of *alternative* designs, in a ‘neighbourhood’ in design space (Sloman 1993).

This involves going beyond the majority of AI projects and psychological investigations in two ways: (a) considering designs for *complete* agents and (b) doing *comparative* analysis of different sorts of designs. We also need to analyse different sets of requirements for behaving systems (niche space). In biological evolution, changes in *designs* for some species or parts of an organism lead to changes in *requirements* for those or other species or parts, which in turn can lead to new changes in designs. Co-evolution of different sorts of organisms and also co-evolution of parts of the same type of organism can therefore be seen as involving multiple feed-back loops in design space and niche space, as discussed in (Sloman 1994*a*, Sloman 1998*b*, Popper 1976, p. 173ff).

The space of possible designs is enormous. Many different designs are capable of explaining similar capabilities. However, we can constrain our theories using considerations such as: (1) trade-offs that influence evolutionary developments, (2) what is known about our evolutionary history, (3) what is known about human and animal brains, the effects of brain damage, and other results of empirical study, (4) what we have learnt in AI about the scope and limitations of various information processing architectures, mechanisms and representations, (5) introspective evidence, such as my knowledge that I considered and evaluated alternative ways of travelling to a conference. (Constraints on theories about biological designs are discussed in (Sloman & Logan 2000).)

8 Biology abhors a continuum

The diversity of architectures and types of minds does not imply that the phenomena form a *continuum* of cases. Jumping to the conclusion that there is a continuum without any interesting divisions is another common mistake, made frequently after people have tried in vain to find a single major division between things with and things without minds (or consciousness, or emotions, etc.)

Many types of discontinuous change can occur during evolution of a complex organism. So an alternative to a single major discontinuity is a large number of lesser discontinuities, as various discrete features of an architecture evolve.

It is often forgotten that Darwinian evolution has to be discrete partly because genes (and DNA) cannot vary continuously. Further, without Lamarckian inheritance there can only be a finite number of evolutionary steps in any time period, no matter how long it is, since there can only be a finite number of generations.

Similarly, discrete changes can occur during individual development, as more and more components of a cluster of capabilities are acquired through adaptation, learning, or genetically driven development. This can include formation of a new modules, creation of new links between pre-existing modules, transfer of control of some processing from one module to another, introduction of a new step or a new test in a stored plan, and so on.

Continuous changes can occur in quantitative parameters represented by analog mechanisms, whereas structural changes are inherently discrete: it is not possible to have half or quarter of a conditional instruction in a working program or an arbitrarily small portion of a data-structure.

So if development of new algorithms, grammars, strategies, information structures etc. is part of what happens during human development, then that too is likely to involve many discontinuous steps, e.g. learning a new algebraic formula, or part of a Beethoven piano sonata, or becoming acquainted with a new person, or town.

Selection among the steps available during learning and development is likely to be extremely variable across individuals, so the result of individual development can be enormous individual diversity, within a discrete space of possibilities, just as in evolution of organisms.

8.1 Different clusters of capabilities, and indeterminate concepts

Consider concepts which connote not dichotomies or matters of degree, but complex, ill-defined clusters of discrete capabilities. Then, if different subsets are present at different stages of development of a species or an individual, it may not make sense to ask at which stage those concepts, e.g. ‘understanding’, become applicable. Very different subsets may be found in different species, and different subsets may remain after brain damage or degeneration, without any prior principle for regarding some of those subsets as constituting understanding.

So it is pointless asking whether chimps or ten month old humans, or certain stroke victims understand language, as if the boundaries between doing so and not doing so were well defined.

In such cases we should not expect our pre-theoretic concepts, such as ‘intelligence’, ‘awareness’, ‘emotion’, etc. to be capable of making sharp distinctions. If a brain damaged person loses a subset of the capabilities involved in interpreting language, does that person still understand or not? At one extreme we would all agree that no understanding is left and at another extreme we would all agree that there is ordinary understanding. But for many intermediate cases where some but not all linguistic capabilities are lost, it may not be clear whether the person should be described as understanding or not: and arguing about it may be as silly as arguing over whether an isotope of carbon really is carbon, or whether a circle is an ellipse or not: it has some of the properties in common with a typical ellipse but not others. For mathematicians it is useful to classify it as an ellipse, but not for wheel-makers.

It may, however, be useful to define different kinds of understanding, different kinds of linguistic capabilities, different kinds of awareness, corresponding to different intermediate cases: for then we can consider different treatments, different implications for other people (e.g. relatives) and different ethical questions. If one patient still knows what is happening around him but has lost the ability to care about it, that is different from one who cares but lacks the ability to know.

Even when we know what sorts of components are implicitly referred to by our pre-theoretic ‘cluster concepts’ we cannot expect the clusters to have precise, generally agreed boundaries. An individual language user may not have any basis for deciding how to classify some of the intermediate cases. Some of the component capabilities may also be described using cluster concepts. The indeterminacy is not a result of smooth variation along a continuum of cases where no boundaries exist, but rather a result of there being many discontinuities and with no principled criteria for deciding which ones mark the conceptual boundaries we all think we already understand.

There need not even be an implicit commitment to particular boundaries. If there has never previously been any reason to decide whether this or that subset of capabilities suffices for attributing understanding, or consciousness there need not be any basis for taking such a decision now. Later we may wish to make our concepts more precise for scientific or philosophical or legal purposes. But different purposes may require different boundaries.

If we can specify the architectures which generate different subsets of information processing capabilities we shall at least have a principled collection of capabilities to consider grouping in various ways. For instance if we find that a particular subset enables a certain kind of social structure we can then ask which animals have that subset, and not bother to argue whether that means they understand language or not.

A theory of such an architecture enables us to ask new, deeper, more precise questions not only about the development of individuals but about the evolution of mentality in different species, replacing old indeterminate, unanswerable questions, such as when consciousness or emotions evolved, or what their function is.

8.2 ‘Private’ ontologies

Self-monitoring reflective processes require the use of an ontology of types of mental states and some sort of formalism for describing and evaluating those states. If the organism includes learning mechanisms which can develop their own ontology and formalisms (as some neural network learning systems are capable of doing on the basis of presentation of many examples, and some symbolic learning systems do using problem-driven processes), we can then envisage an individual which finds ways of describing its own internal states which make sense only to that individual.

If it belongs to a species which has developed a rich communication language, individuals may learn ways of conveying certain aspects of their ontologies for self-description. But there may always be subtle aspects which can be communicated only between individuals who happen, by chance, to have developed similar ontologies.

Since many aspects of ontologies will depend on how internal states relate to the shared environment, to shared needs and goals (food, mates, etc.), there may be evolutionary pressures for individual ontologies to develop a large shared component, enabling quite rich communication about mental states. However there is likely always to be a residue that is individual and incommunicable. For instance, a colour-blind person’s experiences may not be accurately describable to one with normal vision.

9 Conclusion: There’s a long way to go

The previous sections have outlined a research methodology which involves investigating design space and niche space: possible types of information-processing virtual machines and the various types of requirements which they can satisfy. Within that framework we can show that, for a particular type of architecture, concepts can be defined which refer to classes of states, events and processes supported by the architecture, and that in general different architectures support different families of concepts. This allows many old concepts referring to mental states and processes to be extended, refined, sub-divided into special cases, and used to formulate new answerable empirical questions about the mental life of humans, other animals and possible future robots.

Many of our old concepts of mind are cluster concepts with considerable amounts of indeterminacy, allowing them to be refined and extended by removing some of the indeterminacy and making their ill-defined presuppositions more explicit. In some cases new concepts will be required, for describing phenomena we had not previously thought of, as has already happened in computer science with a host of new concepts including memory management, garbage collection, layered protocols, interrupt priorities, event handlers, various kinds of variable scoping, incremental compilation, handshaking, virtual machine, and many more (Minsky 1987).

The exploration of such architectures is an extension of the old philosophical goal of trying to understand what is and is not possible and how it is possible. But it cannot be done as an arm-chair philosophical activity: it is far too difficult. First of all it is essential to learn about phenomena and constraints discovered in other disciplines, such as psychology, neuroscience, psychiatry, ethology, anthropology, computer science and software engineering, which suggest both requirements and designs that would not occur even to the most gifted arm-chair thinker (though science-fiction writers sometimes manage surprisingly well).

Secondly as we consider more and more complex architectures specified in more and more detail we'll find that we cannot simply imagine what their properties will be, nor be certain that our designs are coherent and workable. For that reason philosophy of the kind espoused here has to be done in part by designing and implementing working models, e.g. using the tools and techniques of AI. Often the process of trying to implement a design shows that it is flawed long before the implementation is complete. And later, when it is complete and demonstrable. new limitations will turn up showing that re-design is needed.

If philosophers wish to take part in these design-based explorations, they will have to learn to use the languages, tools and techniques developed for artificial intelligence research. No doubt they will be able to make major contributions to those tools.

These ideas must not be judged by current AI systems, which are still very primitive: whether based on symbolic mechanisms or neural nets or dynamical systems theory, whether implemented using physical robots or only software simulations they are all lacking in most of the capabilities of reptiles, birds and mammals, though some of them are beginning to display insect-like capabilities. The visual and motor capabilities of current artificial systems are nowhere near those of a squirrel, monkey or nest-building bird.

There are some systems which have a very specialised human capability, and may even outperform humans, e.g. in playing chess or proving theorems, but that is *all* they do. They don't enjoy or get bored by it. They don't care about succeeding or failing. They lack the architectural underpinnings for (human-like) wanting and caring even though they may be programmed to give the superficial appearance of wanting and caring. They can't use what they have already learnt in order to master some totally different activity.

In particular the attempts which have been made to simulate emotions so far fail to address the various ways in which emotions emerge from interactions within an architecture, and fail to distinguish emotions requiring different sorts of architectural layers.

Getting beyond such limitations requires deeper analysis of the requirements for doing so. That is a multi-disciplinary research task involving philosophical analysis along with research in brain science, psychology, AI, ethology.

Architectures proposed in future will generate new concepts that may help us in our everyday thinking about ourselves and our loved ones, and also lead to better forms of therapy, deeper theories of learning, improved educational procedures, a new view of the relations between mind and brain, and a better appreciation of how we relate to other animals and the ever more intelligent artefacts that will be developed in centuries to come.

Notes

Many colleagues and students especially at the University of Birmingham have helped with the development of these ideas. Their work is reported in papers in the Cognition and Affect web directory:
http://www.cs.bham.ac.uk/research/cogaff/
Our software tools are available in: http://www.cs.bham.ac.uk/research/poplog/

To help with these investigations we have implemented software tools to support explorations of possible architectures, in the Sim_agent toolkit (Sloman & Poli 1996, Sloman & Logan 1999).

References

Albus, J. S. 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/).

Beaudoin, L. & Sloman, A. 1993 , A study of motive processing and attention, *in* A. Sloman, D. Hogg, G. Humphreys, D. Partridge & A. Ramsay, eds, 'Prospects for Artificial Intelligence', IOS Press, Amsterdam, pp. 229–238.

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.

Damasio, A. R. 1994 , *Descartes' Error, Emotion Reason and the Human Brain*, Grosset/Putnam Books.

Gibson, J. 1986 , *The Ecological Approach to Visual Perception*, Lawrence Earlbaum Associates.

Glasgow, J., Narayanan, H. & Chandrasekaran, eds 1995 , *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, MIT Press, Cambridge, Massachusetts.

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

Jamnik, M., Bundy, A. & Green, I. 1999 , 'On automating diagrammatic proofs of arithmetic arguments', *Journal of Logic, Language and Information* **8**(3), 297–321.

Kohler, W. 1927 , *The Mentality Of Apes*, Routledge & Kegan Paul, London.

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.

Narayanan, E. N. 1993 , 'The imagery debate revisited', *Special issue of* Computational Intelligence **9**(4), 303–435.

Newell, A. 1980 , 'Physical symbol systems', *Cognitive Science* **4**, 135–183.

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

Picard, R. 1997 , *Affective Computing*, MIT Press, Cambridge, Mass, London, England.

Popper, K. 1976 , *Unended Quest*, Fontana/Collins, Glasgow.

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

Scheutz, M. 1999*a* , The Missing Link: Implementation and Realization of Computations in Computer and Cognitive Science, PhD thesis, Indiana University. (University of Michigan Microfiche).

Scheutz, M. 1999*b* , ‘When physical systems realize functions...’, *Minds and Machines* **9**, 161–196. 2.

Sloman, A. 1978 , *The Computer Revolution in Philosophy*, Harvester Press (and Humanities Press), Hassocks, Sussex.

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 , Prospects for AI as the general science of intelligence, *in* A. Sloman, D. Hogg, G. Humphreys, D. Partridge & A. Ramsay, eds, ‘Prospects for Artificial Intelligence’, IOS Press, Amsterdam, pp. 1–10.

Sloman, A. 1994*a* , Explorations in design space, *in* A. Cohn, ed., ‘Proceedings 11th European Conference on AI, Amsterdam, August 1994’, John Wiley, Chichester, pp. 578–582.

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

Sloman, A. 1996 , Actual possibilities, *in* L. C. Aiello & S. C. Shapiro, eds, ‘Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR ‘96)’, Morgan Kaufmann Publishers, pp. 627–638.

Sloman, A. 1998*a* , Damasio, Descartes, alarms and meta-management, *in* ‘Proceedings International Conference on Systems, Man, and Cybernetics (SMC98)’, IEEE, pp. 2652–7.

Sloman, A. 1998*b* , The “semantics” of evolution: Trajectories and trade-offs in design space and niche space, *in* H. Coelho, ed., ‘Progress in Artificial Intelligence, 6th Iberoamerican Conference on AI (IBERAMIA)’, Springer, Lecture Notes in Artificial Intelligence, Lisbon, pp. 27–38.

Sloman, A. 2000 , Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?), *in* K. Dautenhahn, ed., ‘Human Cognition And Social Agent Technology’, Advances in Consciousness Research, John Benjamins, Amsterdam, pp. 163–195.

Sloman, A. & Croucher, M. 1981 , Why robots will have emotions, *in* ‘Proc 7th Int. Joint Conference on AI’, Vancouver, pp. 197–202.

Sloman, A. & Logan, B. 1999 , ‘Building cognitively rich agents using the Sim_agent toolkit’, *Communications of the Association of Computing Machinery* **42**(3), 71–77.

Sloman, A. & Logan, B. 2000 , Evolvable architectures for human-like minds, *in* ‘Proceedings 13th Toyota Conference, on Affective Minds Shizuoka, Japan, Nov-Dec 1999’, Elsevier.

Sloman, A. & Poli, R. 1996 , Sim_agent: A toolkit for exploring agent designs, *in* M. Wooldridge, J. Mueller & M. Tambe, eds, ‘Intelligent Agents Vol II (ATAL-95)’, Springer-Verlag, pp. 392–407.

Strawson, P. F. 1959 , *Individuals: An essay in descriptive metaphysics*, Methuen, London.

Waismann, F. 1965 , *The Principles of Linguistic Philosophy*, Macmillan, London.

Wittgenstein, L. 1953 , *Philosophical Investigations*, Blackwell, Oxford.

Wright, I. P. 1977 , Emotional agents, PhD thesis, School of Computer Science, The University of Birmingham. (Available online at http://www.cs.bham.ac.uk/research/cogaff/).

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