Presented at 13th Toyota Conference, on "Affective Minds"
Nagoya Japan, Nov-Dec 1999
To appear in a volume published by Elsevier, Ed. Giyoo Hatano, 2000, Copyright Elsevier.

EVOLVABLE ARCHITECTURES FOR HUMAN-LIKE MINDS

Aaron Sloman and Brian Logan*
School of Computer Science, The University of Birmingham, UK
http://www.cs.bham.ac.uk/ {~axs/ ~bsl/}

1 Abstract

There are many approaches to the study of mind, and much ambiguity in the use of words like 'emotion' and 'consciousness'. This paper adopts the design stance and attempts to understand human minds as information processing virtual machines with a complex multi-level architecture whose components evolved at different times and perform different sorts of functions. A multi-disciplinary perspective combining ideas from engineering as well as several sciences helps to constrain the proposed architecture. Variations in the architecture should accommodate infants and adults, normal and pathological cases, and also animals. An analysis of states and processes that each architecture supports provides a new framework for systematically generating concepts of various kinds of mental phenomena. This framework can be used to refine and extend familiar concepts of mind, providing a new, richer, more precise theory-based collection of concepts. Within this unifying framework we hope to explain the diversity of definitions and theories and move towards deeper explanatory theories and more powerful and realistic artificial models, for use in many applications, including education and entertainment.

2 Approaches to the study of mind

There are many approaches to the study of mind. Experimental approaches involve searching for patterns in data from laboratories, questionnaires, etc. Philosophers try to analyse the concepts we use in thinking about minds, or try to work out general requirements for a mind. Minds can be regarded as biological phenomena and attempts made to trace their evolution. Some physicists claim that mentality is implicated in the basic mechanisms of quantum mechanics and try to derive therefrom explanations of more familiar mental phenomena. Social scientists view minds as products of culture constituted largely by their social context. AI researchers and some cognitive scientists and brain scientists attempt to build computational models.

There are several types of *computational* approaches to mind, which produce theories and models of varying depth. At one extreme, researchers are concerned entirely with practical goals, such as trying to produce robots or software agents which are 'believable' and produce

*Now at the School of Computer Science & IT, University of Nottingham: bsl@cs.nott.ac.uk

appropriate reactions (such as sympathy) in humans, or which are adequate for some practical purpose, such as controlling a factory. Often, shallow simulations achieve such practical goals without explaining very much. When the main research objective is to produce accurate explanations and models of naturally occurring intelligent systems more complex theories are required. These may involve different levels of abstraction. For example, some theorists strive to produce realistic explanations by modelling known neural structures and processes. Others aim for a more abstract level of modelling.

Our approach is to explore *virtual machine* information processing architectures which might explain human mental phenomena. We operate at an intermediate level, expecting that later work will move 'downwards' from the architecture to connect it with more realistic implementations closer to biological details and 'upwards' by showing how the architecture can explain a wide range of phenomena arising out of mechanisms and processes therein, including social phenomena.

There is a *huge* space of potentially relevant architectures although researchers have so far studied only a tiny subset, so any theories produced in the near future must be provisional, and subject to revision as we learn more.

3 Constraining the architecture

There are indefinitely many virtual machine architectures that could in principle produce any observed behaviours of an organism, even if the organism is observed over its lifetime. How can the search for *explanatory* architectures be controlled? There is no guaranteed method of success, here or in any other area of science. At best we can use general guidelines. For instance, the following kinds of information can help to constrain or guide the search for an explanatory theory:

(1) *Information about the physical design of the system.* Studying brain mechanisms provides such information. Brain structures and mechanisms constrain the types of virtual machine (VM) they can support, though only in subtle and indirect ways. For instance the number of possible states of the brain limits the number of distinct states in the VM, though the mapping between physical and VM levels is complicated by such things as use of sparse arrays, or information implicit in axioms or rules from which consequences can be derived as needed, so that information 'in' the system need not be explicitly mapped onto particular physical components.

(2) *Information about the design history.* If some sub-system within an organism evolved, then the mechanisms of biological evolution will constrain the design, e.g. because precursors of the organism must be complete viable organisms, unlike partially built artificial systems. Thinking about trajectories in evolutionary design space may help us find explanatory mechanisms, for instance based on the notion that evolution frequently modifies a design by producing an extra copy of some component then modifying that component to perform new functions. By contrast, for certain artificial systems, designs are possible that could not have evolved in nature because the process of creation requires production of complex sub-mechanisms which would not be viable as independent agents. Similarly if the VM is partly bootstrapped via a learning or development process then things we know about capabilities

of various forms of learning or development can constrain our theories.
(3) *Observed behaviour.* Information about what the system does, whether in a 'natural' environment or in various laboratory or field-test contexts, help to provide information about the machine's capabilities (subject to many qualifications about how misleading such information can be). For example, humans can deceive us about their abilities, or they can be tired, temporarily forgetful, distracted or misled by ambiguous instructions or questions, and so on. Moreover some behaviours may be constrained by the current culture, not by the intrinsic capabilities of the VM.
(4) *Introspection*. When trying to build theories about your own virtual machine, you also have access to introspective information other people do not have: e.g. I know that a few minutes ago I was wondering whether I should reply to a message or mark a student's essay, and nobody else could have told by observing me that that was going on inside me. So I know things about my VM which do not come from observation of behaviour. (Whether measurements of brain activity could ever provide such information is an open question: it will be even harder than 'decompiling' machine code on computers.)
(5) *Broken parts*. A powerful source of information is to tamper with bits of the physical machinery and see how that affects the capabilities of the system (which is often not at all obvious, and poses its own problems). This can provide evidence that the virtual machine has an unexpected modular structure, since different kinds of damage leave different previously unnoticed capabilities intact. Damage caused by strokes or injuries, and genetic brain defects can all contribute such information.
(6) *Common knowledge*. Explicit or implicit common sense knowledge includes such things as knowledge that people can sometimes be jealous or angry without being aware that they are, and that certain emotional states disrupt thinking and attention, that deep emotions are sometimes externally visible and sometimes hidden from others, that many emotions and motives have rich semantic content (e.g. being angry that someone has betrayed your friendship). Even if many widely held beliefs about minds are erroneous, it does not follow that all are false.
(7) *Information from other disciplines*. Combining knowledge from many disciplines also helps. For instance: philosophy provides techniques for revealing surprising aspects of familiar concepts; psychology and brain science provide many facts that need to be explained; ethology reveals the diversity of animal minds; evolutionary biology helps us understand possible routes from simple to complex brains and the functions of various aspects of mind; computer science and software engineering teach us about important general characteristics of information processing mechanisms and architectures; mathematics provides precise analysis of some of the properties of those mechanisms; and AI has provided us with much experience about specific varieties of virtual machines and what they are and are not good for, and the tradeoffs between design options. Computer engineering and AI also provide tools for testing ideas in working models. We learn both from the process of implementation and from the limitations of the systems we build. (Like the disappointing performance of most current robots!)
(8) *Unifying explanations*. Aiming for a unified explanation of many phenomena helps to constrain theories. Too often theorists study only normal humans. Besides normal adult human minds, we should consider infants, people with brain damage or disease, insects,

chimpanzees and other animals.

The following additional criteria are useful in selecting between rival theories.
(9) *Analogies*. When trying to understand a particularly complex system, observable analogies with systems about which we already have more information including those we have designed ourselves, may give clues, though analogies should always be used with great caution.
(10) *What we have learnt in AI and software engineering.* We have learnt a lot about the sorts of designs which are and are not capable of producing various kinds of functionality, and about the trade-offs involved in choosing between options. This can help us rule out unworkable alternatives and avoid premature commitment to particular designs. However, we still have much to learn under this heading.
(11) *Select among competing theories*. A standard method in science is to compare all the available explanatory theories and try to decide which one is best (in terms of explanatory depth, predictive precision, predictive coverage, consistency with known evidence, coherence with other good theories, etc. etc.). As Popper always stressed (e.g. in his 1976 book), such selections are always *provisional* conjectures, and exploration of alternatives to 'accepted' theories should always be allowed. Nothing is ever final in science.

4 Architectural decomposition

Using these constraints and sources of information as inspiration, we have been exploring virtual machine information processing architectures which might explain human mental phenomena.

One approach which we have found fruitful can be explained (approximately) by superimposing two commonly used architectural decompositions, a 'vertical' and a 'horizontal' decomposition illustrated in Figure 1 (a) and (b) respectively. The first corresponds to a view of the flow of information through a system and the second corresponds to a view of an organism as having levels of control, or, alternatively, as having layers that evolved at different stages.

4.1 *Three towers*

The 'three tower' model (Nilson, 1998) shown with vertical divisions in Figure 1(a), is often implicit. There are different versions of this model, depending on the sophistication of the perceptual, central and motor subsystems. For example simple versions treat the perceptual mechanisms simply as physical transducers. More sophisticated models include complex perceptual processes such as segmentation, recognition, interpretation, and direction of attention, with various intermediate information stores used to record partial results. Another kind of variation concerns the kinds and degrees of control of perception by the central tower. Some architectures support explanatory concepts referring to these intermediate information stores or to central control of perception, whereas others do not (Sloman, 1989).

Likewise, the action component might, at one extreme, be regarded as a collection of transducers sending signals to motors, or in more sophisticated cases as a complex hierar-

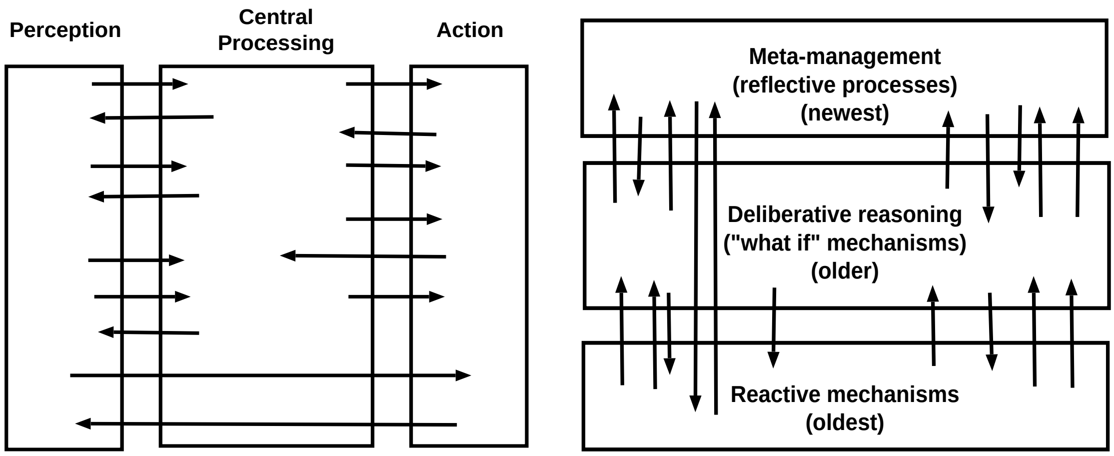

Figure 1: (a) (b)

The diagrams show two ways of partitioning the architecture. In (a) information flow through the system via sensory, central and motor mechanisms is emphasised, where each 'tower' may comprise simple transducers or may have extremely sophisticated multi-functional layered mechanisms. Diagram (b) emphasises differences in degree and type of sophistication of processing layers in the architecture which evolved at different times. Processing in different sub-systems may be concurrent and asynchronous.

chical control mechanism which translates high level instructions into detailed patterns of action by motors or muscles. Examples can be found in (Johnson-Laird 1993) and (Albus 1981, Ch 7), and many studies of action. Action systems may also vary in the amount of feedback they include either within themselves (e.g. proprioceptive feedback) or the extent to which they work in close coordination with perceptual mechanisms, for instance in hand-eye coordination. Again, the more complex and sophisticated action mechanisms will support a wider range of descriptive concepts referring to different types of control of actions.

Thus we can have 'thin' or 'fat' perception or action towers with various kinds of internal layering and various kinds of information flow in different directions.

Additional types of variation depend on whether the system is physically embodied or consists entirely of software, where sensors and motors are virtual machines observing or acting on a software environment.

Yet another complication concerns the ontological level of the model. A three tower model could refer to just the physical architecture. Alternatively it could refer to the more abstract virtual machine functionality involved in processing information of various kinds in various ways, even if the underlying physical (or physiological) implementation does not have clear boundaries. It seems to be the case that in humans much thinking (e.g. some mathematical reasoning, and visualising a rearrangement of furniture) makes use of parts of the brain related to vision (representing and processing spatial structure). It may be that the same physical component implements both part of the perceptual tower and part of the central tower.

4.2 *Three layers*

The 'three layer' model, depicted crudely in Figure 1(b) attempts to account for the existence of a variety of more or less sophisticated forms of information processing and control which can operate concurrently. The version discussed here,[1] postulates three concurrently active layers which evolved at different times and are found in different biological species. The three layers account for different sorts of processes, found in different kinds of animals and will be shown below to provide a framework for distinguishing three different concepts of 'emotion'.

The first layer contains *reactive* mechanisms which automatically take action as soon as appropriate conditions are satisfied. The second *deliberative* layer provides 'what if' reasoning capabilities, required for planning, predicting and explaining. Relatively few organisms have this, and again the forms can vary widely. The *meta-management* layer provides the ability to monitor, evaluate, and partly control, internal processes and strategies.

Roughly, within the reactive layer, when conditions are satisfied actions are performed immediately: they may be external or internal actions. A reactive system may include both analog components, in which states vary continuously, and digital components, e.g. implementing condition-action rules, or various kinds of neural nets, often with a high degree of parallelism.

By contrast, the deliberative layer, instead of always acting immediately in response to conditions, can contemplate possible actions, compare them, evaluate them and select among them. At least in humans, chains of possible actions can be considered in advance, though there are individual differences in such capabilities. The human deliberative system can also consider hypothetical past or future situations not reachable by chains of actions from the current situation, and can reason about their implications. As explained elsewhere (e.g. (Sloman 2000)) physically implementable mechanisms required for such sophistication, including a long term associative memory and especially a re-usable short term memory, will cause the deliberative system to be discrete and serial, and to proceed in much slower steps than a reactive system can.

A meta-management system can act, in a reactive or deliberative fashion, on some of the internal processes involved in the reactive or deliberative (or meta-management) system. This includes monitoring, evaluating and redirecting such internal processes, and possibly reflecting on them after the event in order to analyse what went wrong or how success was achieved. Like the deliberative layer, it will be resource-limited.

We suspect that researchers and therapists who refer to 'executive function' in humans are often unaware that they are discussing mechanisms which combine deliberative and meta-management capabilities. That such capabilities are functionally different is shown by the fact that there are many AI systems that have deliberative capabilities, insofar as they can make plans, execute them, revise them when execution goes wrong, etc, but lack meta-management capabilities. So they may not have the ability to notice that their planning processes are wasteful or that it might be better to abandon the current goal in the light of

[1]Partly inspired by Simon (1967) , and elaborated in our previous papers (Sloman & Croucher 1981, Beaudoin 1994, Sloman 1994, Sloman & Poli 1996, Sloman 1997, Sloman 1998, Sloman 2000, Sloman & Logan 1999)

some new information.

Various forms of reactive mechanisms are found in all organisms and some of them must have developed very early in biological evolution. Deliberative mechanisms evolved later and are found in fewer organisms, though we do not know exactly which ones have them. Can a bumble bee or even a rat wonder what would have happened if...? It seems from Kohler's work that at least chimpanzees can think ahead. Meta-management mechanisms evolved last of all and it is not clear how many organisms have this, apart from humans (though even they may not have it at birth). Perhaps chimpanzees and other animals have less sophisticated versions of meta-management.

How the layers evolved must be largely a matter of speculation. We conjecture that one of the important features of biological evolution making this possible is the process of producing two copies of an old structure, after which one of them develops a new function. In (Maynard Smith & Szathmaŕy 1999) this is referred to as 'duplication and divergence.' E.g. mechanisms which at first stored useful reactive condition-action patterns might later be copied and modified to form a long term associative memory that can be used in 'what-if' reasoning.

Of course, all the different layers must ultimately be implemented in purely reactive mechanisms, otherwise nothing would ever happen. This common implementation feature is consistent with great functional diversity within the layers, just as a common computer architecture can support very different operating systems and software packages.

5 Related theories

The idea of a layered architecture is quite old in neuroscience, including versions similar to the architecture we propose. E.g. Albus (1981, page 184) presents the notion of a layered 'triune' brain with a reptilian lowest level and two more recently evolved (old and new mammalian) levels above that, including hierarchical perceptual and action systems (chapter 7). Freud's distinction between *id*, *ego* and *super-ego* seems to be a related idea. AI researchers have been exploring a number of variants, of varying sophistication and plausibility, and varying kinds of control relations between layers. The 'subsumption hierarchy' in Brooks (1991) is one of many examples. Compare Minsky (1987) and Nilsson (1998). Johnson-Laird's discussion (1993) of consciousness as depending on a high level 'operating system' is related to our third layer. A multi-level architecture is proposed for story understanding in (Okada & Endo 1992).

In some theories presenting layered architectures, it is assumed that as sensory information comes in, increasingly abstract interpretations or summaries are passed up through various layers until at the highest level it may trigger processes which cause instructions to act to trickle down through the layers. By contrast, in our hypothesised architecture all the layers get information (of different degrees of abstraction) in parallel and process it in parallel and may produce action signals (of different degrees of abstraction) in parallel. An example might be walking with a friend and simultaneously discussing philosophy, while digesting food, controlling posture, admiring the view, etc. Most of the processes are unconscious, of course.

6 Combining models, to form a grid

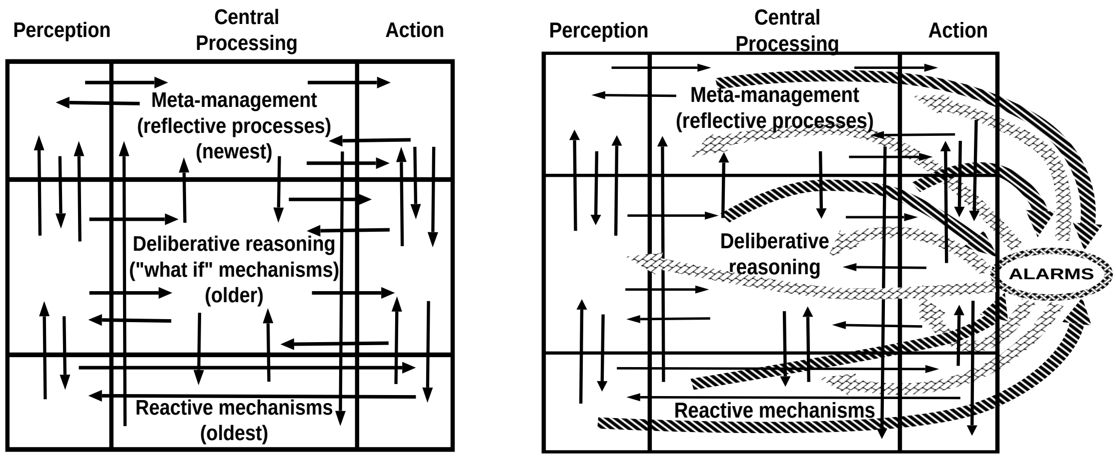

Figure 2: (a) (b)

The first figure serves as a mnemonic indicating simultaneously the triple tower and triple layer views, where the various components in the boxes will have functions defined by their relationships with other parts of the system. In (b) a global alarm system is indicated, receiving inputs from all the main components of the system and capable of sending control signals to all the components. Since such alarm systems need to operate quickly when there are impending dangers or short-lived opportunities, they cannot make use of elaborate inferencing mechanisms, and must be pattern based. Global alarm mechanisms are likely therefore to make mistakes at times, though they may be trainable.

When the horizontal and vertical subdivisions are superimposed we obtain the schema outlined in Figure: 2(a). In Figure: 2(b) we make explicit the role of *global alarm mechanisms*, which receive information from all components of the system and are able to send interrupts and redirection signals to all parts of the system. The idea of this sort of global alarm mechanism was partly inspired by consideration of engineering requirements, partly by the discussion of interrupts in Simon (1967) and partly by studies of the brain, especially the role of the limbic system, e.g. see Albus (1981) and LeDoux (1996).

If processing of information in any of the layers is likely to take too long in relation to the urgency of some need provoked by the environment, for instance if there is a large object coming rapidly towards you, or a fast-moving edible object flying past you when you are very hungry, then it may be necessary for ‘normal’ processes to be interrupted and redirected. This could be achieved by the addition within the reactive mechanisms of one or more modules receiving information from sensors or other parts of the system and using fast and general pattern-recognition techniques to decide to interrupt everything and redirect the whole system towards an appropriate response, e.g. running away, freezing, pouncing, attending to a particular object in the environment, and so on.

Whether this requires a special mechanism or can simply be part of the normal functioning of a reactive system, depends on the relative speeds of various kinds of processing. There is no need to interrupt and redirect a system towards a particular action if it was about to do that anyway, and just as quickly.

A number of additional mechanisms, listed briefly in Figure 3(a), enable the various layers to function, and their shortcomings (e.g. limited processing capacity of the deliberative layer) to be compensated for. The very cluttered Figure 3(b) impressionistically portrays the result of putting various pieces together, including the alarm mechanism.

7 Motive generators

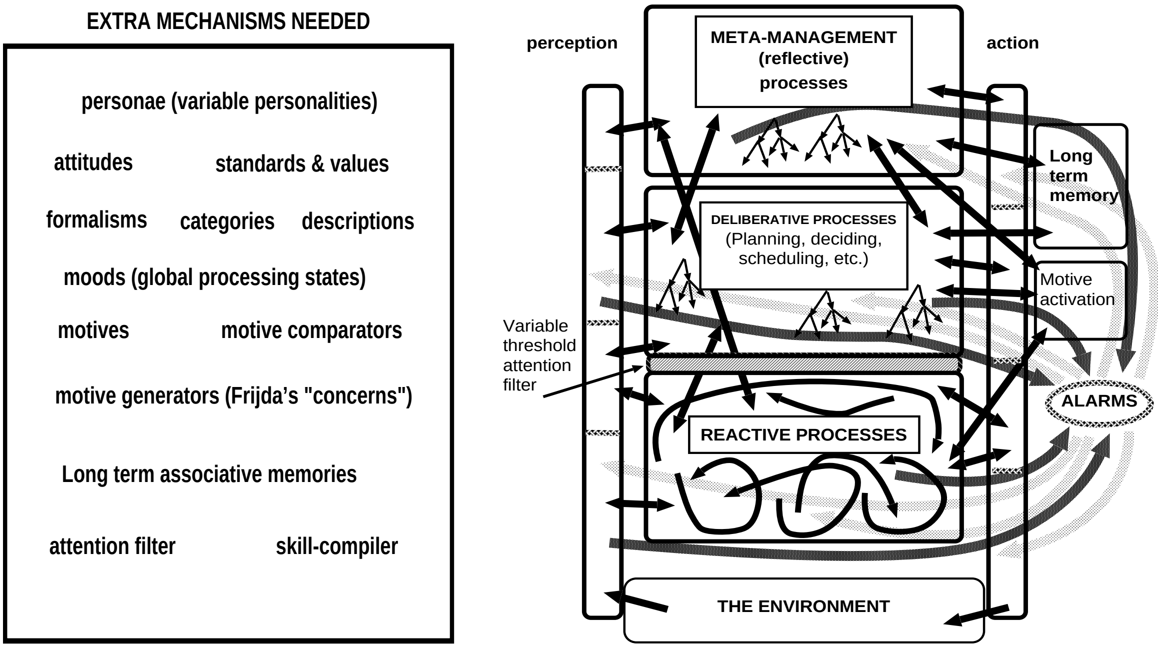

Figure 3: (a) (b)

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) the 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.

Both in a sophisticated reactive system and in a deliberative system with planning capabilities there is often a need for motives which represent a state or goal to be achieved or avoided. In simple organisms there may be a fixed set of drives which merely change their level of activation depending on the current state of the system. In more sophisticated systems not all motives are permanently present, so there is a need for *motive generators* to create new goals possibly by instantiating some general goal category (*eat something*) with a particular case (*eat that foal*). These generators are similar to the dispositional 'concerns' in Frijda's theory. Beaudoin's and Wright's PhD theses discuss various types of generators or 'generactivators' and related implementation issues.

The ability to manipulate goals, and to plan actions, or tailor actions to current contexts, requires various kinds of representational mechanisms. In particular, reasoning about which actions are possible now or in hypothetical future situations, or about the consequences of those actions, requires one or more powerful long term associative memories, whose form will be related to the form of representation used for goals, situations and actions. These will not necessarily map directly onto sensory input (e.g. the perceivable *affordances* discussed by Gibson (1979), such as 'something edible' or 'a support' may have enormously varied sensory manifestations) and that may induce a need both for more abstract central representational capabilities and also for more sophisticated processing of perceptual input in order to provide information at the appropriate level of abstraction for the deliberative mechanism. While this happens the visual system might simultaneously be sending more 'primitive' sensory information to other parts of the system, e.g. for posture control.

It is important that none of the three layers has total control: they are all concurrently active and can influence one another. The degree and type of influence will vary from time to time. In particular, all three layers can be disrupted by the global alarm mechanisms.

8 Dynamic filters and moods

Since the different layers, the sensory systems and the alarms, operate concurrently, it is possible for new information that requires attention to reach a deliberative or meta-management sub-system while it is busy on some task. Because of the resource limits mentioned previously, it may be unable to evaluate the new information while continuing with the current task. But it would be unsafe to ignore all new information until the current task is complete. So new information needs to be able to interrupt deliberative processing.

Under stressful conditions, deliberative mechanisms with limited processing or temporary storage capacity can become overloaded by frequent interrupts. We have argued elsewhere (e.g. (Beaudoin & Sloman 1993)) that variable-threshold attention filters can reduce this problem. Setting the threshold at a high level when the current task is urgent, important and intricate, can produce a global state of 'concentration' on that task. (Malfunctioning of this mechanism may produce a type of attention disorder (Beaudoin 1994).)

Variations in the external context and the individual's needs and resources will require more coarse-grained global control mechanisms. This may account for some *moods*. For example, when the environment can be classified as 'friendly' because most goals are relatively easily achieved, a confident optimistic mode of behaviour may be fruitful. When things often go wrong, and predators abound, a more cautious, even pessimistic demeanour may be much safer. More subtle and complex changes of mood may be triggered by recognition of socially significant contexts. There are some global state changes produced by pathologies, e.g. depression, manic states. Much more research is needed to help us understand the architectural basis of a wide variety of types of global state changes, including, for instance, the nature of sleep.

9 Architecture-based concepts

The model outlined above (like many similar models) allows us to generate systems of cognitive and affective concepts which are grounded in the virtual machine information processing architecture of agents. Such architecture-based concepts may prove superior to those we currently use to formulate research questions and our theories.

We often think we know exactly what consciousness, experience, emotions, etc. are, because we experience them directly. This is mistaken. We may experience *simultaneity* 'directly' sometimes, but that does not guarantee a clear grasp of the concept, as Einstein showed. One way to deepen our understanding of these concepts, and, where necessary, repair their deficiencies, is to seek an explanatory architecture and then use it as a framework for systematically generating concepts, just as the theory of the sub-atomic architecture of matter generated concepts of kinds of elements, kinds of chemical compounds and processes, etc. The relation between the periodic table of elements and modern ideas about the architecture of matter illustrate how an underlying architecture can give new clarity and coherence to a family of concepts. The new concepts do not *replace* our old ones, but extend and refine them, for instance adding concepts of isotopes to old ideas of chemical elements, and adding new ideas about valency to old ideas about chemical processes.

10 Architecture-based emotion concepts

In our previous work we have attempted to show how the three layers might support states and processes which correspond to some of our pre-scientific concepts of 'emotion'. Such processes also explain a common distinction between 'primary' and 'secondary' emotions (e.g. found in Damasio, 1994; Picard, 1997) and suggest a need for an additional category of 'tertiary' emotions.

The reactive layer accounts for *primary* emotions (e.g. being startled). The deliberative layer explains *secondary* emotions with greater semantic content and less dependence on events in the perceived environment (e.g. apprehension concerning what might happen when a risky plan is executed and relief concerning what did not happen). Various processes impinging on the meta-management producing partial loss of control of attention and thought processes account for *tertiary* emotions, states in which people may find it hard to redirect their attention or hard to maintain a focus of attention, or where various kinds of thoughts ('Did she really like me?' 'How can I have my revenge?' 'Why won't he change his mind?', etc.) constantly intrude despite decisions to ignore them and concentrate on important and urgent tasks. These tertiary emotions such as humiliation, jealousy, and thrilled anticipation, are probably unique to humans, though perhaps simplified versions can be found in some other animals.

It is well known that definitions of 'emotion' vary widely (Oatley & Jenkins 1996). We expect that further work on varieties of architecture-based concepts will reveal a still wider range of architecture-based concepts of emotion, along with new, more precise, architecture-based concepts corresponding to old ideas about mood, motivation, attitude, personality, perception, learning, and so on. This sort of model provides a unifying framework which helps us explain the diversity of definitions, caused by different researchers (unwittingly) focusing on different parts of the same architecture.

11 Multiple personalities

In humans it seems that the meta-management layer does not have a rigidly fixed mode of operation. Rather it is as if different personalities, using different evaluations, preferences and control strategies, can inhabit/control the meta-management system at different times. E.g. the same person may have different personalities when at home, when driving on a motorway and when dealing with subordinates at the office. Switching control to a different personality involves turning on a large collection of skills, styles of thought and action, types of evaluations, decision-making strategies, reactive dispositions, associations, and possibly many other things.

For such a thing to be possible, it seems that the architecture will require something like a store of 'personalities', mechanisms for acquiring new ones (e.g. via various social processes), mechanisms for storing new personalities and modifying or extending old ones, and mechanisms which can be triggered by external context to 'switch control' between personalities.

If such a system can go wrong, that could be part of the explanation of some kinds of multiple personality disorders.

It is probably also related to mechanisms of social control. E.g. if a social system or culture can influence the meta-management processes that determine how an individual represents, categorises, evaluates and controls his own deliberative processes, this might provide a mechanism whereby the individual learns things as a result of the experience of others, or mechanisms whereby individuals are controlled and made to conform to socially approved patterns of thought and behaviour. An example would be a form of religious indoctrination which makes people disapprove of certain motives, thoughts or attitudes, leading to redirection of deliberation in more 'socially acceptable' directions.

12 An ecology of mind

We have indicated how during evolution the changing needs of the central processing mechanisms might lead to developments of higher level layers in the perceptual and motor mechanisms. For instance, development of a deliberative layer leads to a requirement for more sophisticated and abstract input from the sensory systems ('chunking' at higher levels of abstraction, to provide knowledge relevant to general planning capabilities). It can also produce pressure for evolution of higher level control mechanisms within the action subsystems, including the ability to perform social actions, such as greeting, performing rituals, or cooperating on complex skilled tasks requiring good coordination. If hierarchical control of action can be devolved to a sophisticated action system this releases central resources for other tasks.

All this suggests that we can think about the boxes on the grid, and the additional components, as forming an ecology in which sub-organisms co-evolve so that developments in some of them generate needs and opportunities for the others. Such co-evolution involves a family of parallel trajectories through both 'design space' and 'niche space' (Sloman 1998). Although the components clearly are not separate organisms, they do co-exist, performing different tasks, making use of different information, sometimes co-operating and sometimes competing with one another. Just as different organisms in the same part of the forest may analyse their sensory inputs in different ways and encode information about the environment using different ontologies and possibly different forms of representation, so also may different sub-components of a single complex organism. For instance incoming visual information, as mentioned previously, may be processed to produce a variety of different descriptions about affordances used in parallel by reactive mechanisms, deliberative mechanisms and meta-management levels, almost as if they had their own eyes. Similarly, different internal state monitoring processes may use different ontologies in recording events, generating goals, etc.

In some ways all this is reminiscent of Minsky's ideas on a 'society of mind' (Minsky 1987) though perhaps the phrase 'ecology of mind' is more apt if we think of the various components as having co-evolved to meet different pressures and opportunities provided by the other components.

This is of course only a metaphor, and some of the differences from more common forms of co-evolution may help us to understand the strengths and weaknesses of the metaphor. Very often co-evolution of whole organisms involves competition. But often it involves cooperation, such as evolution of the shape of a flower and evolution of the shape and behaviour of insects or birds that obtain nectar from the flower. Co-evolution within an organism is more likely to be of the cooperative form, though there could also be competition, e.g. competition for resources, such as information, blood supplies, etc. (See also Ch 8 of Maynard Smith and Szathmáry (1999)).

The main difference is that normal biological co-evolution involves organisms that can replicate independently, whereas parts of a single organism cannot. Nevertheless just as a mutation that changes a type of flower may produce opportunities for change in a bee, so a mutation that alters the capabilities of a perceptual sub-mechanism might produce new opportunities for useful changes in a more central component.

It is not possible for reproductive fitness of one component to increase or decrease independently of fitness of another, since they reproduce together when the organism reproduces. However, different parts or features of an individual may be thought of as having different degrees of 'fitness' according to how fast they spread through a population. This is clearly related to the notion that different genes within a genome may have different reproductive fitness.

Although the notion of an 'ecology' must not be taken too literally, nevertheless, trying to understand processes of incremental changes in different parts of the architecture may help us understand how the whole system evolved, and how that system works. Our discussion is closely related to Popper's proposal (1976, p. 173) to distinguish external and internal selection pressures. For instance he suggests that sometimes in biological evolution new *preferences* evolve, e.g. if having those preferences aids survival and reproduction. This in turn can produce a new 'niche' in which there is pressure for certain skills to evolve. Thus organisms will be favoured by natural selection if they develop skills which serve those preferences. This in turn can produce a pressures which favour certain anatomical changes if they support those skills. Those changes may then support the evolution of new preferences, e.g. if they serve the needs of the new anatomical mechanisms.

13 Some conjectures

It is conjectured that the three layers can be used to explain different sorts of consciousness, ranging from simple sentience to full reflective self-awareness and possession of 'qualia' (Sloman 2000), though there is no space to elaborate on this here.

Likewise it is conjectured that this sort of architecture could give a robot many human-like mental processes – including falling into philosophical confusions about consciousness.

Of course, some robots, like many animals, young children, or even brain damaged adult humans, will have only parts of the system present, and their cognitive and other states will be correspondingly limited.

Similar comments can be made about software agents.

14 Conclusion

As science, much of this is conjectural – many details still have to be filled in and consequences developed (both of which can come partly from building working models, partly from multi-disciplinary empirical investigations).

An architecture-based ontology can bring some order into the morass of studies of affect. We have begun to illustrate this by showing how different concepts of emotion relate to processes arising in different parts of a complex architecture, though there is still much work to be done. This partly helps to explain why there are diverse definitions of emotion in the literature: different researchers unwittingly focus on different subsets of the phenomena we have referred to as primary, secondary and tertiary emotions. It should already be clear from our discussion of the proposed architecture that this is a crude and inadequate classification: additional important subdivisions between types of emotions and other affective states can be based on the differences in mechanisms involved in generating them and the different ways they develop, subside, are suppressed, trigger further emotions, etc. (Wright, Sloman & Beaudoin 1996).

Another feature of the architecture, pointed out in (Sloman 1989), is that it predicts that perceptual information follows many different routes through the brain, supporting and triggering diverse processes within the mental ecology. This may account for phenomena found by (Goodale & Milner 1992) and others involving different visual pathways. However our architectural proposals suggest that far more functionally distinct sensory pathways exist than have been discovered so far. We should also not be surprised to find that sometimes connections go wrong producing phenomena such as synaesthesia in which different sensory modalities become entangled.

It is very unlikely that newborn humans are born with such an architecture fully formed, though simpler organisms may have their architectures largely determined innately. We need more research on how architectures are bootstrapped both in altricial species (where individuals are born or hatched in a relatively helpless and undeveloped state, like humans, hunting mammals, and birds of prey) and precocial species (where individuals are born or hatched far more able to look after themselves, e.g. sheep, deer, grazing mammals, chickens, and many aquatic animals). Such research might lead to deep insights in comparative psychology, developmental psychology (e.g. if much of the architecture develops after birth in humans). This should also provide an improved conceptual framework for studies of effects of brain damage and disease, by enabling us to classify far more precisely than before the many ways in which things can go wrong. It will also point to a much richer classification of types of development and learning within individuals: the more complex the architecture the more varieties of possible change and development there are, at least in principle.

By comparing and contrasting architectures required for embodied animals and those that suffice for software agents we can produce an improved conceptual framework for classifying types of emotions that can arise in software agents, for instance those that lack the reactive mechanisms required for controlling a physical body.

There are implications for engineering as well as science. Designers of complex systems need to understand the issues discussed here:
(a) if they want to model human affective processes,
(b) if they wish to design systems which engage fruitfully with human affective processes, e.g. really convincing synthetic characters in computer entertainments,
(c) if they wish to produce teaching/training packages for would-be counsellors, psychotherapists, psychologists.

There is already recognition of the importance of modelling affective processes in synthetic agents among organisations and researchers involved in the entertainment and games industry. Our introduction pointed out that some of this work can use very shallow models. However, as the requirements for realism become more demanding it could turn out that simulations in computer games and entertainments will have the side effect of leading to very deep advances in psychology and philosophy.

15 Acknowledgements & Notes

Some of this work was done as part of a project funded by the Leverhulme trust. The ideas presented here were developed in collaboration with colleagues and students in the Cognition and Affect Project, at the University of Birmingham. Papers and theses by students and colleagues can be found at:

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

Our tools (including the SIM_AGENT toolkit) can be found here:

http://www.cs.bham.ac.uk/research/ poplog/freepoplog.html
http://www.cs.bham.ac.uk/research/ poplog/newkit/

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.

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

Frijda, N. H. (1986), *The emotions*, Cambridge University Press, Cambridge.

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

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

Johnson-Laird, P. (1993), *The Computer and the Mind: An Introduction to Cognitive Science*, Fontana Press, London. (Second edn.).

Kohler, W. (1927), *The Mentality Of Apes*, Routledge & Kegan Paul, London. 2nd edition.

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

Maynard Smith, J. & Szathmaŕy, 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.

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

Oatley, K. & Jenkins, J. (1996), *Understanding Emotions*, Blackwell, Oxford.

Okada, N. & Endo, T. (1992), 'Story generation based on dynamics of the mind', *Computational Intelligence* **8**, 123–160. 1.

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

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

Simon, H. A. (1967), 'Motivational and emotional controls of cognition'. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

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. (1994), 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. (1997), What sort of control system is able to have a personality, *in* R. Trappl & P. Petta, eds, 'Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents', Springer (Lecture Notes in AI), Berlin, pp. 166–208.

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

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. & 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.

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.