In Cognitive Processing, Vol 2 No 1, 2001, pp 177–198

Beyond Shallow Models of Emotion

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

Abstract

There is a huge diversity of definitions of "emotion" some of which are associated with relatively shallow behavioural or measurable criteria or introspectable experiences, for instance use of facial expression, physiological measures, activity of specific regions of the brain, or the experience of bodily changes or desires, such as wanting to run away, or to hurt someone. There are also deeper theories that link emotional states to a variety of mechanisms within an information processing architecture that are not easily observable or measurable, not least because they are components of virtual machines rather than physical or physiological mechanisms. We can compare this with "shallow" definitions of chemical compounds such as salt, sugar, or water, in terms of their appearance and observed behaviours in various test situations, and their definitions in the context of a theory of the architecture of matter which is mostly concerned with postulated sub-atomic entities and and a web of relationships between them which cannot easily be observed, so that theories about them are not easily confirmed or refuted. This paper outlines an approach to the search for deeper explanatory theories of emotions and many other kinds of mental phenomena, which includes an attempt to define the concepts in terms of the underlying information processing architectures and the classes of states and processes that they can support. A serious problem with this programme is the difficulty of finding good constraints on theories, since in general observable facts are consistent with infinitely many explanatory mechanisms. This "position paper" offers as a partial solution the requirement that proposed architectures be capable of having been produced by biological evolution, in addition to being subject to constraints such as implementability in known biological mechanisms, various resource limits (time, memory, energy, etc.) and being able to account for a wide range of human functionality. Within such an architecture-based theory we can distinguish (at least) primary emotions, secondary emotions, and tertiary emotions, and produce a coherent theory which explains a wide range of phenomena and also partly explains the diversity of theories: most theorists focus on only a subset of types of emotions, like the proverbial blind men trying to say what an elephant is on the basis of feeling only a leg, an ear, a tusk, the trunk, etc.
Keywords: affect, architecture, artificial intelligence, cognitive science, deliberative, emotion, evolution, intelligence, meta-management, mind, reactive, reflective virtual machine.

*This is a revised version of a paper presented at the workshop on *Behaviour planning for life-like avatars*, at the I3 Spring Days Workshop, March 1999, Sitges, Spain. It is not intended to be a thorough and scholarly survey, but a provocative "position paper" which outlines an ambitious approach to the study of mind, building on the various approaches which it criticises as inadequate!

1 Introduction

The study of emotion in cognitive science and AI has recently become very fashionable, with a rapidly growing number of workshops, conferences and publications on the topic, some reporting attempts to produce emotional behaviour in robots or software agents, some concerned with detecting and responding to emotions in human users of computing systems, and some aiming to model and explain human emotions.

This is not a new topic in AI, as shown by Simon's important contribution over 30 years ago (Simon, 1967), and various papers nearly 20 years ago in IJCAI'81 including my first paper on this topic (Sloman and Croucher, 1981), which was much influenced by Simon's notion that emotions, motivations and other affective phenomena were deeply entwined with cognitive processes and the mechanisms for control of internal and external behaviour in intelligent agents.

There are now many useful surveys of issues concerning emotions[1], but it is difficult for newcomers to the field to achieve a balanced overview, not least because (as Oatley and Jenkins point out) there is a very wide variety of definitions of "emotion" offered by researchers with different viewpoints. For AI researchers aiming to produce working systems it is tempting to think of emotions as relatively easily simulated patterns of behaviour. The result is a tendency for researchers to present simplistic AI programs and robots as if they justified epithets like "emotional", "sad", "surprised", etc.

Such programs may be based on an attempt to analyse conditions under which certain emotions are thought to occur and the behaviours typical of such emotions. This leads to the design of an architecture controlling a robot or interactive software system, in which there is a sub-component (possibly labelled "emotion") which tests for those conditions and generates the corresponding behaviours, possibly using state variables with names like "angry", "frightened", "surprised", "pleased", etc. either with boolean values that can be toggled or with a numerical or "qualitative" range of values for each variable. These models are shallow insofar as they have relatively simple relationships between input and output. This is similar to a practice lambasted long ago by McDermott (1981) namely using terms like "goal", "plan", "learn", simply because there are procedures or variables with these names in a program.

2 Shallow models are not all bad

Some researchers (Bates et al., 1991; Reilly, 1996) have quite explicitly acknowledged that they are aiming for *shallow* models whose merits are based on *breadth*, namely possessing a variety of capabilities supported by diverse mechanisms, or mechanisms that can cope with a wide range of cases. Such "broad and shallow" designs may be useful for certain practical purposes such as enlivening computer games or other interactive entertainments or perhaps helping naive users of computing systems by making them appear more "human" than they are. One way to achieve such breadth, while still using a shallow model, is to try to encompass a very wide range of cases, such as those surveyed in (Ortony et al., 1988).

Shallow models are fine if they have a limited purpose which is made clear, e.g. to entertain, or to teach programming, or to model some limited aspect of control of posture or facial expression,

[1] E.g. (Goleman, 1996; LeDoux, 1996; Oatley and Jenkins, 1996; Ortony et al., 1988; Picard, 1997; Elliot, 1998; Hatano et al., 2000)

etc. I have a very shallow model[2] in which simulated mobile robots can be in states described as glum, surprised, neutral or happy, but this is nothing more than an elementary teaching tool. Students play with and extend it in order to learn agent programming techniques. In the near future, there will probably be a growing use of very shallow models of emotion in computer entertainments. There is nothing wrong with that, if they are successful at entertaining. However that does not necessarily make them plausible models of human or animal emotions. They may not even be useful steps in the direction of such models.

Shallow models can sometimes play a role in the search for deeper models. Building inadequate models, and exploring their capabilities and limitations is often an essential part of the process of learning how to design more complex and more satisfactory models, as explained in (Beaudoin and Sloman, 1993; Sloman, 1993b).

3 Inconsistent definitions and usages

If we want to understand and model what are normally referred to as "emotions" in humans and other animals then we need to start from a deeper analysis of the concepts we are aiming to instantiate. This task is made difficult by the fact that we do not all agree in our usage of the word "emotion". For example, some will call *surprise* an emotion whereas others (Ortony et al., 1988) will say that it is just a cognitive state in which an expectation has been violated, as often happens in a complex and dynamic world, and can even occur when doing mathematics. Of course, surprise, like any other state, can trigger states that most people would call emotions.

There are also disagreements over whether pains and pleasures are emotions, some regarding it as obvious that they are, whereas others find it equally obvious that one can have the pain of a pin-prick or the pleasure of eating an ice cream without feeling at all emotional about it. E.g. one can be totally unconcerned about the pin-prick, while acknowledging that it hurt. Of course, very intense pain is a different matter.

Another example: some people believe that emotions, by definition, cannot exist without being experienced, whereas others (including some novelists and playwrights) regard it as obvious that someone can be angry or infatuated (and therefore in an emotional state) without being aware of their state, even if friends notice it. On further investigation this dispute can sometimes turn on whether an emotion's being experienced is taken to imply that the emotion is recognized and labelled as such, or only to imply that it involves being aware of some mental states and processes related to the emotion. At one extreme a theorist will say that you cannot enjoy something unless you recognize and categorise your state as enjoyment. An intermediate position would claim that there must be some experience that you recognize and categorise which is part of the enjoyment, even if the total state is not recognized. At another extreme it is claimed (Ryle, 1949) that intense enjoyment can occur where all one's attention is focused on *external* phenomena, e.g. enjoying a game of football where one is thinking only of the other players, where the ball is, who needs to be marked, etc., without being aware of anything internal to oneself. Another such example is enjoying an opera or play with attention fully engaged by what is happening in the theatre, without being aware of any *additional* processes going on in one's own mind. When it is objected that there must be some additional experienced state for enjoyment to occur it is not clear whether this is a conceptual disagreement or an empirical one. (What evidence could help to settle it?)

[2] See http://www.cs.bham.ac.uk/research/poplog/sim/teach/sim_feelings

A different dimension of disagreement concerns the attribution of emotional states and possibly other mental states, to other animals. Does a fish feel pain when caught on a hook? When a fly detects and escapes just in time from the hand slamming down on it, does it have a state of fear, or relief at its narrow escape? Do pains occur if you pull its legs off? Debates over animal rights frequently revolve around disagreements over what mental states are possible for animals. It is also possible to argue over mental states of a human foetus or neonate. Does the physical response to a prod show that a human foetus, or a snail, finds it unpleasant?

Not only are there differences in theories and usages between individuals, it is even possible for individuals to be inconsistent in their own usage, for instance some people will state that *love* is a type of emotion, then later admit that they (a) they are not in an emotional state and (b) that they love their family, their country, the game of football, etc. It is possible that when such people offer love as an emotion they are thinking of episodes of passion or fervour, whereas when they say they love their family, etc. they are referring to an *attitude* which is primarily a collection of *dispositions* which are dormant most of the time but can be triggered, under certain conditions, to produce emotional episodes, involving various mental and physical processes. Similar inconsistencies can arise over the classification of moods as emotions: someone may regard being in an optimistic mood as an emotional state, yet claim not to be feeling emotional when in a state which they also characterise as optimistic.

Inconsistencies between and within the explicit theories and the non-reflective linguistic usage of people who talk about emotions are an indication that we are dealing with a deep set of confusions about how our ordinary concepts work. Perhaps those concepts are simply inadequate for the purpose of characterising the enormously rich variety of mental states that can occur in humans and other animals.

From this viewpoint it is very rash to assume that the aim of building machines that have emotions or which model them is a well-defined aim.

4 Possible strategies

What can we do about this? There are many alternatives, including the following strategies.

- Give up talk of emotions (and other mental states) in our science (as some behaviourists tried to do).

- Invent a precise definition of "emotion", for instance in terms of a set of condition-response patterns, and use it regardless of how it relates to ordinary usage or the definitions offered by others – the simplest strategy for would-be emotion modellers.

- Treat the concepts as inherently fuzzy or probabilistic and attempt to investigate the associated probabilities by doing research to find out probabilities of various labels being used in various contexts, or the probabilities of various behaviours or expressions being used when people claim to be in an emotional state.

- Attempt to produce a deep theory of the information processing architectures underlying all the different phenomena, and then define new architecture-based concepts that precisely identify subsets of those phenomena. This could include states and processes involving the agent's relations with the perceived physical or social environment.

In the Birmingham Cognition and Affect project we have adopted the architecture-based approach[3], described in papers in the project directory at www.cs.bham.ac.uk/research/cogaff/.

5 Opinions about an elephant

This has led us to hypothesise an explanatory architecture, sketched below, and to identify various types of states and processes that can occur in such an architecture. We can then investigate the properties of those states and processes which seem to correspond to the sorts of phenomena that are of concern both in ordinary conversations about emotions and also in various scientific and philosophical research endeavours. We can then formulate definitions of a wide variety of states and processes supported by the architecture which can be grouped in various ways. For instance, as explained below, we have found it illuminating to distinguish *primary*, *secondary* and *tertiary* emotions, which arise out of different architectural layers that may or may not be present in different animal and robot architectures. Further subdivisions can then be made within these three categories. We can also investigate precisely defined architecture-based concepts that approximate to other loose concepts of ordinary language, such as mood, attitude, intention, desire, etc.

From this viewpoint the contradictory opinions expressed by people studying emotions are rather like the opinions of the proverbial ten blind men each trying to say what an elephant is on the basis of feeling only a small part of it. Instead of arguing over which description is right we can try to characterise the whole elephant, thereby explaining the contradictions between rival definitions and partial theories. This is partly like the approach adopted in (Ortony et al., 1988), namely characterising a space of possible states independently of debates about words and phrases accurately correspond to which states.

The multi-layer architecture described below accommodates several different varieties of states which could be called emotions: very primitive primary emotions rooted in very old biological mechanisms such as startle mechanisms shared with many other animals, and also more sophisticated semantically rich secondary and tertiary emotions that are probably unique to humans (until we build human-like robots), such as being apprehensive about the outcome of a risky plan, being infatuated with someone, or feeling humiliated because some silly mistake you made was pointed out by a famous person in a large public lecture.

The taxonomy of Ortony et al. focuses on a particular set of cognitive and motivational states (including what some people would describe as attitudes rather than emotions) and can be accommodated within the classes of secondary and tertiary emotions described below, though they are less concerned with the specification of a complete architecture.

6 How to achieve greater depth

A desirable but rarely achieved type of depth in an explanatory theory is having a model which accounts for a wide range of phenomena. One of the reasons for shallowness in psychological theories is consideration of too small a variety of cases.

[3]Previously referred to as the *design-based* approach in (Sloman, 1992), where design-based theories are contrasted with phenomena-based theories, which merely look for relationships between observable phenomena and semantics-based theories which use linguistic investigations to discover what we mean by various expressions describing mental phenomena.

If instead of thinking only about normal adult humans (or only about rats as some experimentalists used to do) we consider also infants, people with brain damage or disease, and also other animals including insects, birds, bonobos, etc., we find evidence for myriad information processing architectures each supporting and explaining a specific variety of mental capabilities. Yet more possible architectures, each supporting a collection of possible states and processes can be found in robots, software systems and machines of the future!

Concepts describing mental states and processes in one animal or machine may be inappropriate when describing another, if the latter lacks the required architecture, even if its *behaviour* appears to justify the attribution. For instance, a purely reactive animal reacting to a threatening situation may be thought to be in a state of fear. But a genetically determined automatic escape reaction is different in many ways from an externally similar escape reaction produced by a system that understands the implications of the threat and on that basis decides to escape.

Likewise, concepts relevant to normal adult humans may be inappropriate for new-born infants, victims of Alzheimer's disease, or an entertaining robot which can be made to *look* happy, annoyed, surprised, etc.

Although human adults seem to be innately programmed to attribute all sorts of mental states to infants, it is likely that new-born infants are incapable of having some of them. Most people would agree that a newborn infant is incapable of wondering whether it will ever have grandchildren. Why? Likewise a newborn infant may be incapable of feeling humiliated by people laughing at its facial expression, if it lacks the architecture required for humiliation. It may even be incapable of feeling pain in the same way as an adult, despite displaying compelling external symptoms.

It often goes unnoticed that much of what poets and novelists say about us, and what we say about our friends and ourselves when gossipping or discussing our interests, loves, hopes, fears and ambitions, implicitly presupposes that humans are essentially information processing systems. E.g. when poets distinguish *fickle liking* which is easily diminished by new information and *deep love* which is not, they implicitly presuppose that new information can have powerful effects on information-based control states.

By considering possible descriptive and explanatory concepts generated by a *virtual machine information processing architecture* we obtain a broader and deeper explanatory theory than is normally found in philosophy, psychology or social science, or most computer modelling. Of course, such a theory should satisfy empirical constraints including evolvability, implementability in neural mechanisms, resource limits, etc.

7 Exploring neighbourhoods in design space

Looking at the variety of states and processes supportable by a class of architectures has been likened above to seeing the whole elephant. Unfortunately, there is more than one "elephant" to study, since architectures vary between organisms (and machines) and even within an individual it may develop over time, e.g. between infancy and adulthood.

A full understanding of the various phenomena that might be called emotions, therefore requires comparative analysis of possibilities and trajectories in design space and niche space,[4]

[4] A niche, in biology or engineering, is an abstract set of requirements for an organism or machine, against which instances of a class of designs can be compared. In simple cases we can use a "fitness function", giving a numerical result. In general the relation between a design and a niche is best thought of as a complex qualitative description.

as outlined in (Sloman, 1994; Sloman, 1998b; Sloman, 2000b). 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. We also have a deeper understanding of the architecture if we can see what sorts of pressures and trade-offs led to its evolution, and how it might develop or evolve in future.

This involves going beyond the majority of AI projects or psychological investigations insofar as it requires us both to consider designs for *complete* agents in addition to designs for component mechanisms and also to do *comparative* analysis of different sorts of designs.

A comprehensive theory of emotions and other mental states requires a survey of types of information processing architectures covering humans of various types, other animals, future robots and software agents. For each type of architecture we can precisely define the sorts of states and processes it supports, and if we decide to label some of those as emotional states it becomes a factual question whether particular organisms or machines are in such a state or not: the answer depends on whether the individual (a) has an information processing architecture that is capable of supporting such states, and (b) whether the components of the architecture are in the appropriate functional states to produce the precisely-defined sort of emotion that is in question. (Some of the concepts may be defined in relation to features of the environment. E.g. wanting to go up the Eiffel Tower is a state that depends on the existence of the Eiffel Tower.) Having produced such precise definitions of various kinds of architecture-based mental states we can formulate and, perhaps begin to answer, new, more precise, questions about which agents are capable of having which sorts of emotions, experiences, thoughts, and so on. There is then no risk of being bogged down in endless terminological disputes or philosophical arguments at cross-purposes, as often happens how.

Of course, the fact that a question is a factual one with correct and incorrect answers does not imply that it is easy to determine the answer, as the history of physics shows very clearly. Sometimes the question has to remain unanswered until new technology is available to probe the system in greater depth and precision than previously. Sometimes the theory has to be extended with links to other theories before observation or measurement can provide relevant evidence. This point is well discussed in standard literature on the history and philosophy of science, e.g. in Popper (1934) and chapter 2 of Sloman (1978).

8 Constraints on theorising

If we wish to go beyond the study of sorts of information processing architectures that are theoretically possible, and attempt to describe the architecture of a particular individual or the architectures typical of members of a certain biological species we find that it is extremely difficult to infer the architecture of a machine that we have not designed ourselves if we do not have access to design specifications used in its production. This is analogous to the task of decompiling large "legacy software" systems.

No amount of observation of the external behaviour of any animal or machine can determine the underlying architecture, since in principle any lifelong set of behaviours can be produced by infinitely many different information processing architectures, including totally unstructured, unintelligible, "flat", multi-component architectures, as suggested in Figure 1. Decompiling information gleaned from invasive or non-invasive observation of internal physical structures is just as

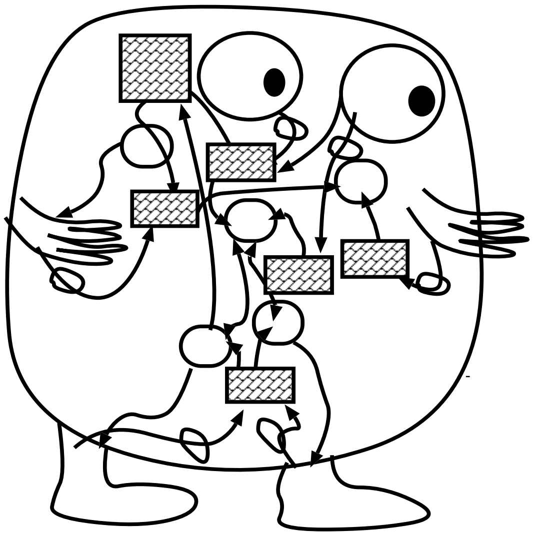

Figure 1: *An unstructured mess?*
Any observed behaviour might be produced by an unintelligibly tangled and non-modular architecture. (Rectangles represent information stores and buffers, ovals represent processing units, and arrows represent flow of information, including control signals.)

hard, e.g. if we don't even know at what physical level most of the architecture is implemented. Do neurons or molecules do most of the information processing?

A common way of avoiding these problems is to formulate theories that address a very narrow range of phenomena so as to yield conditional predictions that can be tested: if we do X to people in conditions C, they will respond by doing Y, etc. The problem is that limiting one's theorising to such easily testable hypotheses prevents formulation of truly deep explanatory theories, such as those which have been of most profound importance in physics.[5]

The study of mind is far more complex than the study of physics as there are so many possible information processing architectures supporting different collections of concepts and different types of laws of behaviour. In general it is not possible to formulate interesting testable hypotheses about how a particular sort of mind works without assuming (explicitly or implicitly) the type of information processing architecture that it uses. But deciding which architecture to propose is very difficult and is not in general constrained by experimental observations, though they certainly provide clues and tests.

We can, however, constrain our theories by combining a number of considerations which I have discussed a greater length in (Sloman, 1998b; Sloman, 2000a), such as: (1) trade-offs that can influence evolutionary developments, (2) what is known about our evolutionary history, (3) what is known about human and animal brains and the effects of brain damage, (4) what we have learnt in AI about the scope and limitations of various information processing architectures,

[5]This topic was discussed at greater length in chapter 2 of (Sloman, 1978), which distinguished the study of the *form* of the universe from the study of its *contents*, including regularities and correlations. Deep science requires the former. Shallow science assumes a form, often implicitly, and then investigates a subset of the contents compatible with that form. A deep theory might state that there exist sub-atomic particles that have various masses and electric charges that can be combined in different ways. A shallower theory might relate the deflection of a stream of electrons to the strength of a magnetic field through which they pass.

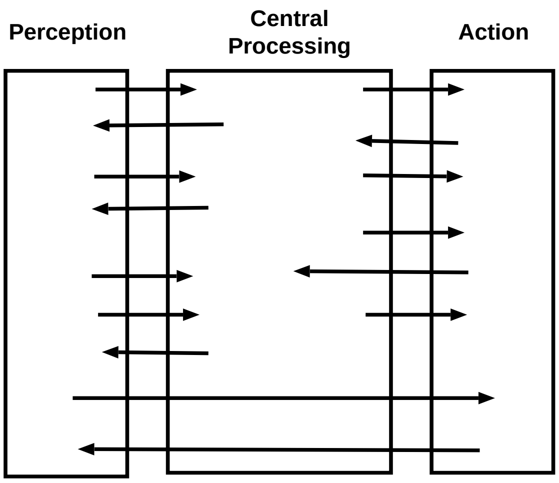

Figure 2: *A "vertical" division into three towers*
Organisms and robots require perceptual mechanisms and action mechanisms of varying degrees of sophistication, along with some persistent internal state which may be modified over various time-scales. This leads to Nilsson's "triple tower" model (Nilsson, 1998). Arrows represent flow of information of various kinds including control signals. The boundaries between the "towers" need not be very sharp, especially where there is rich two-way information and control flow. Later we show that each pillar can be divided horizontally.

mechanisms and representations, (5) introspective evidence, such as my knowledge that before buying tickets I considered and evaluated alternative ways of travelling to the conference where this paper was presented. These constraints are prior to the sorts of requirements more commonly found in philosophy of science texts, such as the requirement of testability, or the requirement to fit statistical data better than alternative theories that have been proposed.

Although our theories will still remain conjectural for some time to come, because of the complexity of human minds and brains, we can at least hope to show that some conjectures are better than others, if we take a broad enough view of what needs to be explained. The next few sections outline a two stage approach. The first stage characterises a general architecture-schema called CogAff which specifies in broad outline a variety of types of functional roles for mechanisms that may occur within organisms or robots of various kinds. In the second stage we present an instance H-Cogaff of this schema which we propose as a first draft model of the information processing architecture typical of human minds.

The CogAff schema defines a framework of possible designs for information processing architectures for organisms or machines. It is useful for thinking about biological organisms, but is not intended to cover *all* possibilities, as it says nothing about many of the architectures designed by engineers, and it does not include distributed multi-agent systems, though it could specify the individuals in such a system.

9 CogAff: an architecture schema

Nilsson (Nilsson, 1998) proposed that intelligent systems can be analysed in terms of the "triple tower" model depicted in Figure 2, which approximately separates perceptual mechanisms, central processing mechanisms and action mechanisms. He calls the central tower the “model tower”, though this label may be too restrictive for the range of functions sketched below. The triple tower model is mainly a result of functional analysis combined with observation of existing organisms.

Another breakdown of information processing functionality comes from both functional and evolutionary considerations. This is the triple layer model sketched in Figure 3, and discussed at greater length in previous papers (e.g. (Sloman, 1997; Sloman, 1999b; Sloman, 1998a; Sloman, 2000a; Sloman and Logan, 1999; Sloman and Logan, 2000)). These three levels are different from the three discussed by Nilsson in chapter 25 of (Nilsson, 1998), though there is some overlap.

If the three layers and the three towers are superimposed as in Figure 4 we arrive at a grid of types of architectural components, where perceptual mechanisms have several layers with different kinds of sophistication required to meet the needs of the different central layers. Likewise the action mechanisms may have different levels of sophistication supporting different sorts of functionality arising out of different levels of central processing. In the figure we have also depicted an “alarm” mechanism, which could also be thought of as merely a part of the central reactive layer, receiving inputs from all over the architecture and sending control signals to many parts of the system, in order to achieve rapid redirection of internal and external processing.

The CogAff scheme thus depicted specifies a variety of components which need not all be present in a particular machine, and which may be related in different ways, giving different specific architectures. For instance an insect or simple robot might have an architecture including only the reactive layers as in Figure 5 whereas some other animals might have both deliberative and reactive mechanisms as in Figure 6.

Moreover, very different designs follow from different functional relations between components. For example, we refer to an Omega architecture as one in which the information flow is essentially a pipeline with information coming in at bottom left, going up the central column to some high level decision making system and then flowing down the centre and out through the bottom right, roughly with the shape of a Greek Ω. For an example see (Albus, 1981). This has some similarities with the Contention Scheduling model in (Cooper and Shallice, 2000).

The subsumption architecture proposed by Brooks (Brooks, 1986; Brooks, 1991), can be seen as a variant in which there is only a reactive layer, containing several parallel pipelines, with information flowing from left to right within each pipeline, but with factual information going up from lower levels to higher levels and control information going down from higher levels to lower levels. A hybrid architecture such as Figure 6 might include a reactive subsumption layer and a deliberative layer.

One of the key features that gives the CogAff schema its generality is the possibility that the different components, instead of forming parts of simple pipelines, can all be concurrently active and concurrently sending information of various kinds to arbitrarily many other components, allowing a wide variety of feedback mechanisms and triggering mechanisms. For instance a high level goal generated within the deliberative or meta-management layer could send information to perceptual mechanisms in order to direct them physically and alter their processing (Sloman, 1989). Likewise different sorts of central processing at different levels of abstraction might send signals with different levels of abstraction to action mechanisms. Since many such things could happen concurrently we infer a need for arbitration mechanisms. One such is the attention filter with dynamically varying filter threshold in Figure 6. Often architectures proposed with diagrams that look superficially similar turn out to be very different when the details are specified, including details such as possible directions of information flow and degrees of concurrency.

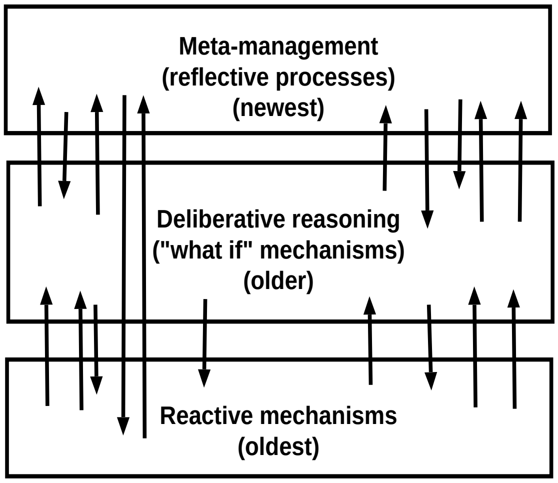

Figure 3: *A "horizontal" division into three layers*

It is now commonplace in AI to distinguish reactive mechanisms in which states detected by sensors (whether external or internal) immediately trigger responses (whether external or internal) from deliberative mechanisms in which alternative possibilities for action can be considered, categorised, evaluated, and selected or rejected. More generally a deliberative mechanism may be capable of "what if" reasoning about the past or future or even how the present might have been. The depth, precision and validity of such reasoning can vary. A meta-management layer adds the ability to monitor, evaluate, and to some extent control processes occurring within the system in something like the way the whole system observes and acts on the environment. The two bottom layers differ in that the second evolved much later and requires a far more sophisticated long term memory and symbolic reasoning capabilities using a short term re-usable memory. The third layer may have evolved later and requires explicit use of concepts referring to states of an information processing architecture. The earliest organisms, like most existing organisms, were totally reactive. Deliberative and meta-management layers evolved later. Adult humans appear to have all three types of processing, which is probably rare among other animals. The three layers operate concurrently, and do not form a simple dominance hierarchy. As previously, arrows represent flow of information and control, and boundaries need not be sharp in all implementations.

10 Sketch of a theory of humans: H-Cogaff

Within the general framework of the CogAff schema we have developed a particular instance which we now call H-Cogaff, depicted in Figure 7, and discussed in more detail in earlier papers e.g. (Sloman, 2000a). Our conjecture is that the information processing architecture of a normal adult human is something like H-Cogaff (augmented with sub-mechanisms not shown in the figure). This conjecture is based on evidence of many kinds from several disciplines, and the sorts of constraints on evolvability, implementability and functionality mentioned above. According to this theory:

(a) Evolution, like engineers, found that (partly) modular designs are essential for defeating combinatorics in the search for solutions to complex problems (with only 4,000,000,000 years and one biosphere on an earth-sized planet available).

(b) Human information processing makes use of (at least) three different concurrently active

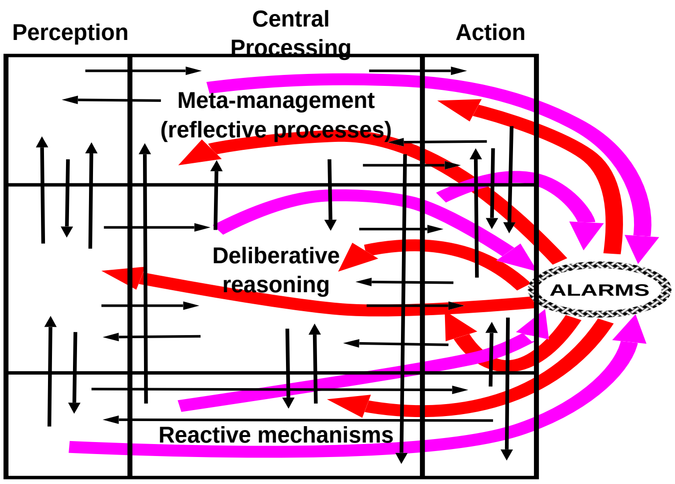

Figure 4: *The CogAff Schema: pillars, layers and alarms*
If we consider a system in which both the division between perceptual, central and motor systems can be made, and also the division between reactive, deliberative and meta-management layers, and if we assume that the perceptual and motor systems include components related to the needs of all three central layers, then we have a three by three grid of architectural components with different sorts of functionality. If some of the internal processing is slow relative to the speeds at which things happen in the environment, then it may be useful to have inputs from many parts of the system to a fast pattern driven reactive "alarm" mechanism that can redirect the whole system. Solid arrows are as before. The shaded arrows represent information flowing to and from the alarm mechanism. The alarm mechanism being purely reactive and pattern driven will typically be stupid and capable of mistakes, but may be trainable.

architectural layers, a reactive layer, a deliberative layer, and a meta-management layer which evolved at different times, which we share with other animals to varying degrees, along with various additional supporting modules such as motive generators, "global alarm" mechanisms and long term associative storage mechanisms. The different layers and supporting mechanisms may have evolved from purely reactive mechanisms by means of the typical evolutionary trick of making another copy of an existing mechanism and then gradually transforming the functions of the new copy. This almost certainly happened several times in the evolution of brains.

(c) Reactive systems may be very complex, and powerful, especially if internal reactions can be chained together and can cause modification of internal states which trigger or modulate other reactions. I do not claim that deliberative or meta-management mechanisms provide behavioural capabilities that could not *in principle* be provided by purely reactive mechanisms. Rather I have argued elsewhere that achieving the same functionality by purely reactive means would have required a far longer period of evolution with more varied circumstances, and a far larger brain to store all the previously evolved reactive behaviours. The time and brain size required for a purely reactive human-like system are probably too large to fit into the physical universe. Some people who argue in favour of purely reactive systems do not consider the trade-offs involved in these resource issues. Merely showing that in principle reactive systems suffice proves nothing about what can work in practice.

(d) Reactive, deliberative and reflective layers support different classes of emotions found in humans and other animals, including the primary and secondary emotions discussed by Damasio and Picard (Damasio, 1994; Picard, 1997), and the tertiary emotions I have discussed in commenting on their work (Sloman, 1998a; Sloman, 1999a).

1. the reactive layer, including a global alarm mechanism, accounts for *primary* emotions (e.g. being startled, frozen with terror, sexually aroused);

2. the deliberative layer supports *secondary* emotions like apprehension and relief which require "what if" reasoning abilities (these are semantically rich emotions);

3. the meta-management (reflective) layer supports not only control of thought and attention but also loss of such control, as found in typically human *tertiary* emotions such as infatuation, humiliation, thrilled anticipation of a future event. (This layer is also crucial to absorption of a culture and various kinds of mathematical, philosophical and scientific thinking.)

All the layers are subject to interference from the others and from one or more fast but stupid partly trainable "global alarm" mechanisms (e.g. spinal reflexes of various sorts, the brain stem, the limbic system including the amygdala, etc.)

(e) A more fine-grained analysis of types of processes that we tend to call "emotions" in humans would show that the above three-fold classification into primary, secondary and tertiary emotions is somewhat superficial. For instance, there are different ways emotions can develop over time, and the three-fold distinction does not say anything about that. A short flash of anger or embarrassment which quickly passes is very different from long term brooding or obsessive jealousy or humiliation which gradually colours more and more of an individual's mental life.

(f) Perceptual and motor systems are also layered: the different layers evolved at different times, act concurrently, and have different relationships to the "central" layers. E.g. deliberative mechanisms make use of high level characterisations of perceived states, e.g. seeing a bridge as "rickety" or an ornament as "fragile". Using some of Gibson's ideas, this can be described as perception of abstract affordances.

(g) Analysing ways in which components of such an architecture might bootstrap themselves, develop, reorganise themselves, acquire and store information, or go wrong, will provide far richer theories of learning and development than ever before.

(h) The three layers account for different cognitive and affective states, as well as different possible effects of brain damage, and other abnormalities. For instance, some aspects of autism seem to involve malfunctioning or non-functioning higher level perceptual mechanisms (as suggested in (Sloman, 1989)).

(i) A multi-layered architecture of the sort proposed could give robots various kinds of human-like mental states and processes, including *qualia* arising out of inward focused attention. As science fiction writers have noted, this might lead some robots to re-discover philosophical confusions about consciousness. Software agents could have similar capabilities. However, detailed differences in physical embodiments and virtual machine architectures could entail many kinds of minor differences in the mental states of which they are capable. This is no different in principle from the fact that mental states possible for adults and children are different, or for males and females, or humans and cats.

Many doubt these claims about robots because they see the limitations of existing computer-based machines and software systems and cannot imagine any ways of overcoming these limitations. They do not realise that we are still in the early stages of learning how to design information

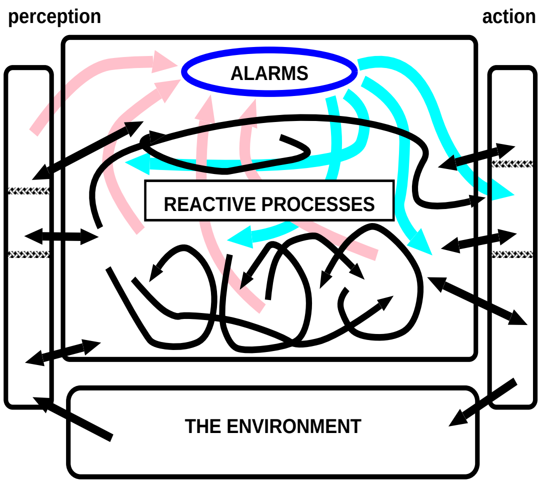

Figure 5: *A reactive system with global alarms.*
Something like this might be an architecture for a fairly sophisticated insect.

processing systems. (Claiming that computers will be ever more powerful is not enough to allay these doubts: we also need deep analysis of the concepts used to express the doubts.)

11 Alternatives in design space

Although the above theory includes a sketch of an architecture for human-like intelligent systems, there is no suggestion that this is the only sort of intelligence. ‘Intelligence’, like ‘emotion’, is a *cluster concept*, referring to a variable cluster of capabilities, and admitting a wide variety of types of instances, with no sharp boundaries. In particular, animals (and perhaps humans) exist with different subsets of the full array of mechanisms described above, and within those mechanisms considerable variation is possible.

For example, many insects appear to be capable of remarkable achievements based entirely in complex collections of purely reactive mechanisms, such as termites constructing their “cathedrals”, with air conditioning, nursery chambers and other extraordinary features.

So I am not denying that there can be organisms (and robots) which are purely reactive, or which combine a reactive mechanism with a separate global alarm system, as in Figure 5.

More sophisticated organisms have both a reactive and a deliberative layer, providing “what if” reasoning capabilities, as illustrated in Figure 6. Such mechanisms provide the ability to construct specifications of hypothetical past or future situations and to reason about them. Many writers, including Craik (Craik, 1943) as long ago as 1943, have pointed out that such abilities may increase biological fitness.

It seems that some other animals besides humans have deliberative mechanisms though they vary enormously in their richness and flexibility. For instance, how effective such capabilities are, will depend on a number of factors including the type and size of re-usable short term working memory, the type of representational mechanisms available, the type and size of the trainable associative memory which can store generalisations about the environment, and so on.

The deliberative layer might have evolved as a result of a mutation which at first led to the copying of a trainable associative memory in a purely reactive system. After that, the new copy

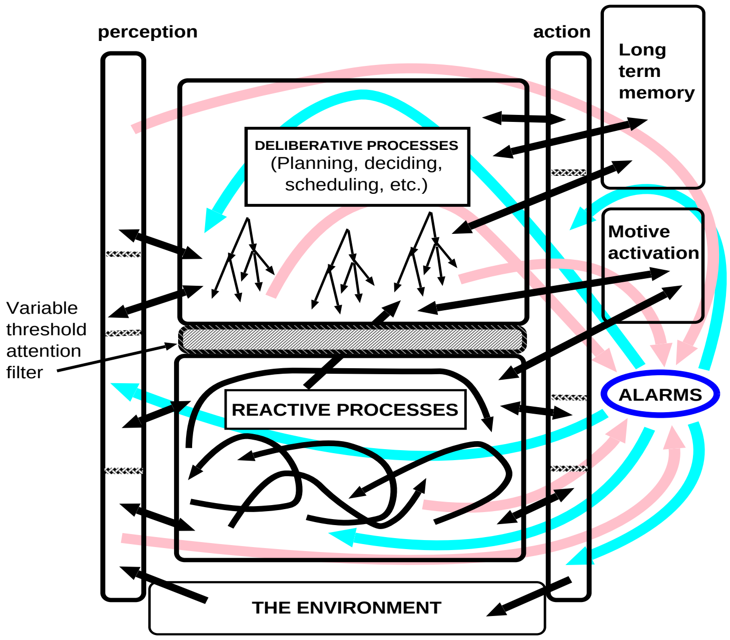

Figure 6: *A hybrid architecture with global alarms.*
In a hybrid reactive and deliberative system, it may be necessary to have an "attention filter" with dynamically varying filter threshold to protect the resource-limited deliberative mechanism from being interrupted too often during urgent and intricate tasks. However an alarm system or intense perceptual inputs may be capable of exceeding the filter threshold.

might have gradually evolved, along with other mechanisms, to provide the ability to answer questions about "what would happen if" instead of "how shall I react now". Making good use of such a "what if" reasoning capability requires being able to store generalisations about the environment at an appropriate level of abstraction to allow extrapolation beyond observed cases. This in turn could generate evolutionary pressure towards perceptual systems which include higher level abstraction mechanisms. All this is, of course, highly speculative, and needs to be tested empirically, though it is consistent both with what is known about evolutionary mechanisms and with the at least partly modular structure of the brain.

More generally, within this framework we can see a need for a generalisation of Gibson's theory of perceptual affordances (Gibson, 1986) (contrasted with Marr's theory of vision in (Sloman, 1989)) to accommodate different perceptual affordances for different components in the more central processing mechanisms. This requires the sharing of sensory resources between concurrently active subsystems, and can generate conflicts, as discussed in (Sloman, 1993a).

Deliberative capabilities bring their own problems, such as how they should be controlled, how different deliberative strategies should be selected or interrupted, how they should be evaluated and modified. For this purpose and others, it seems that an even smaller subset of animals, including humans, have evolved a third architectural layer providing the ability to direct attention *inwardly* and to monitor, evaluate, and in some cases modify what is happening internally. Luc Beaudoin first drew my attention to some aspects of the need for this layer, and called it meta-management. Some of the requirements were analysed in his PhD thesis (Beaudoin, 1994).

Earlier papers (e.g. (Wright et al., 1996)) have discussed some of the ways in which this theory

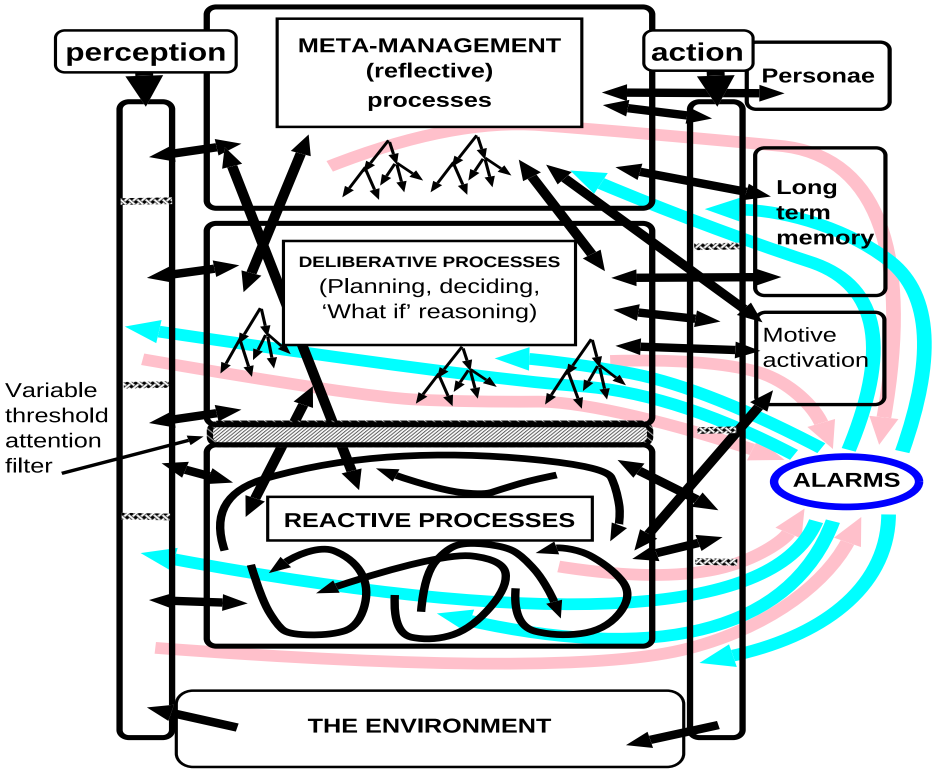

Figure 7: *H-Cogaff – a three layer architecture.*
The meta-management layer provides the ability to attend to, monitor, evaluate, and sometimes change internal processes and strategies used for internal processes. However, all the layers and the alarm system(s) operate concurrently, and none is in total control. A collection of high level culturally determined "personae" may be available, turned on and off by different contexts and causing global features of the behaviour to change, e.g. switching from bullying to servile behaviour. Note that some of the divisions between layers are a matter of taste: some authors e.g. (Davis, 1996) prefer to separate out reflexes from the reactive layer, and some would prefer to separate out some of the high level functionality of the meta-management layer.

accounts for distinctively human emotions such as grief, infatuation, excited anticipation, humiliation, involving partial loss of control of attention. We used to call these emotions "perturbances", but now refer to them as tertiary emotions, to distinguish them from the primary and secondary emotions discussed by Damasio and others.

Since these tertiary emotions (perturbances) involve loss of control of attention, and you cannot lose what you have not got, only an organism which has something like meta-management capabilities can get into such states. This does not mean that all humans have this capability. New born infants, people with degenerative brain disease or brain damage, may lack such capabilities.

12 Are emotions required for intelligence?

It is clear that local reflexes and global alarm mechanisms can be useful in organisms or machines which sometimes require very rapid reactions to occur faster than normal processes of perception, reasoning, deliberation, and planning. Such reactions can produce simple and obvious effects such as freezing, fleeing, producing aggressive sounds or postures, pouncing on prey, sexual responses, and more subtle internal effects such as attention switching and "arousal" which might involve different kinds of information processing. Because these reactions often need to happen very quickly they can be triggered by a relatively stupid, but trainable, pattern recognition system.

Many human emotions seem to involve the operation of such mechanisms. These and other emotions are connected with resource-limits in more "intelligent" subsystems. If those systems could operate faster, and with more complete information, it would not be necessary for more "stupid" mechanisms to override them.

Damasio (in (Damasio, 1994)) pointed out that certain kinds of frontal lobe damage can simultaneously remove the ability to have certain classes of emotions and also undermine the ability to achieve high level control of thought processes required for successful management of one's life. Pending further investigation of details, this gives some support for the claim that there are classes of emotions, referred to as "tertiary emotions" above, which depend on mechanisms that are concerned with high level management of mental processes.

Damasio argued from this that emotions are a *requirement* for intelligence, and since then the argument has been repeated many times: it has become a sort of *meme*. However, the reasoning is fallacious, as I have argued in (Sloman, 1998a; Sloman, 1999a). The brain damage in question might merely have disabled some mechanisms involving control of attention, required *both* for tertiary emotions and for management of thought processes. It doesn't follow that emotions somehow contribute to intelligence: rather they are a side-effect of mechanisms that are required for other reasons, e.g. in order to overcome resource limits as explained above.

Here's an example of similarly fallacious reasoning that nobody would find convincing. Operating systems which support multiple concurrent processes are extremely useful, but they can sometimes get into a state where they are "thrashing", i.e. spending more time swapping and paging than doing useful work. If some damage occurred which prevented more than one process running at a time that would prevent the thrashing, and remove the useful benefits of multiprocessing. It doesn't follow that a thrashing mechanisms is required to produce useful operating systems. In fact, by adding more memory and CPU power, thrashing can be reduced and performance enhanced. Likewise, it is possible for mature humans to learn strategies for avoiding emotions, and this can often improve the quality of their lives and the lives of people they live with or work with.

I am not arguing that all emotions are undesirable or dysfunctional. There are many emotions that have an important biological role (e.g. sexual passion, and aggression in defending a nest), and some emotions that humans value highly, including aesthetic emotions and the joy of discovery. I also accept, as most AI researchers have accepted over many years, that there are many purely intellectual problems which require exploration of search spaces that are too large for complete, systematic, analysis. The use of heuristic pattern-recognition mechanisms is often useful in such cases, to select avenues to explore and to redirect processing. But they can operate without generating any emotions.

13 Conclusion

This paper is a snapshot of an ongoing long term multi-disciplinary research project attempting to understand the nature of the human mind and how we fit into a larger space of possible designs for biological organisms and artificial agents of many kinds.

The ideas have many links with previous work by others. Some aspects of the methodology (defining an architecture-based collection of concepts and then investigating their relations with those in any particular language) have much in common with the strategy in (Ortony et al., 1988). Besides the strong and obvious connections with work of Simon, Gibson, and Nilsson's ideas cited previously, there are also links with work of Dennett, Minsky, Picard, Damasio and many others, not all listed in the bibliography. . However there is no room in this paper for a full survey of similarities and differences between the various theories.

There has also not been space to explore all the implications of the ideas presented here (e.g. showing how they can accommodate the space of possibilities presented in (Ortony et al., 1988)), but one thing is very clear: we are a long way from implementing artificial systems with the full richness and complexity of systems containing all the types of mechanisms defined by the CogAff scheme or the H-Cogaff architecture.

There are many gaps in what current AI systems can do, insofar as they are thought of as steps towards modelling human intelligence, and beyond. Existing AI systems do not yet have whatever it takes to enjoy or dislike doing something. They do not really *want* to do something or *care* about whether it succeeds or fails, even though they may be programmed to give the superficial appearance of wanting and caring, or feeling happy or sad. Animal-like wanting, caring, enjoying, suffering, etc. seem to require types of architectures which have not yet been analysed.

Simulated desires and emotions represented by values for global variables (e.g. degree of "fear") or simple entries in databases linked to condition-action rules may give the appearance of emotion, but fail to address the way semantically rich emotions emerge from interactions within a complex architecture, and fail to distinguish different sorts of emotions arising out of different types of processing mechanisms within an integrated architecture.

Current AI models of other animal abilities are also limited: for example, visual and motor capabilities of current artificial systems are nowhere near those of a squirrel, monkey or nest-building bird. To understand animal comprehension of space and motion we may need to understand the differences between precocial species born or hatched with considerable independence (chickens, deer) and altricial species which start utterly helpless (eagles, cats, apes). Perhaps the bootstrapping of visuo-motor control architectures in the latter yields a far deeper grasp of space and motion than evolution could have pre-programmed via DNA. The precocial species may have much simpler visual capabilities, largely genetically determined.

There are many issues that are still unclear, and a vast number of remaining research topics. In particular it is not clear how much of this is relevant to the design of software agents inhabiting virtual machine environments only, and lacking physical bodies. Many of the human reactive mechanisms and some of their motivators and emotional responses are closely linked to bodily mechanisms and functions. E.g. if you don't have a body you will never accidentally step on an unstable rock, and you will not need an "alarm" mechanism that detects that you are about to lose your balance and triggers corrective action, including causing a surge of adrenalin to be pumped around your body.

Nevertheless events can move fast in a virtual machine world (as many system administrators fighting malicious intruders will confirm) and even pure software agents may need reactive mechanisms. Still, it is likely that the combinations required for software agents may include some architectures never found in agents with physical bodies. Whether the reverse is the case depends on whether all sorts of physical bodies and physical environments can, in principle, be simulated on sufficiently powerful physically implemented computers: an open question.

Artificial agents which do not share our deep grasp of spatial structure and motion will be limited in their ability to communicate with us. However, it is not obvious that in order to share this knowledge such agents *must* have similar bodies and processing architectures. For instance, people who have never wanted to kill someone, may nevertheless understand some of the thought processes of a murderer (a fact on which the success of many novels and plays depends). Similarly someone who has been blind from birth can understand a great deal about visual capabilities of sighted people, for instance, that colours are extended properties of 2-D surfaces, somewhat like tactile textures.

So it remains possible that some software agents which are very unlike us will be able to engage in rich communication with us, though the detailed requirements for this are still not clear.

And of course, in the meantime, teachers and designers of computer games can build many entertaining or didactic, shallow simulations which lack most of the features discussed here. That is fine, as long as they take care how they describe what they have done.

Acknowledgements

I am grateful for useful critical comments by anonymous referees and regret that it was not possible for me to follow up all of their suggestions. I am grateful to Elisabeth André for encouraging submission of this paper to the journal. Many colleagues and students have helped with the development of these ideas, most recently Brian Logan, Steve Allen, Catriona Kennedy and Matthias Scheutz. Work of the Cognition and Affect group is presented at length in papers listed in the bibliography and at the web site
http://www.cs.bham.ac.uk/research/cogaff/
The project is summarised in:
http://www.cs.bham.ac.uk/~axs/cogaff.html
Our software tools including the SimAgent toolkit (Pop-11 based code and documentation) can be found as part of the Free Poplog FTP site:
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
SimAgent is not committed to any particular architecture: rather it supports exploration of a wide range of architectures in single or multiple simulated agents. It also includes teaching materials which can be used to introduce students to some of these ideas through practical experience of designing and modifying simple agents in simulated worlds, including a sheep and sheep-dog scenario, and various kinds of obstacle avoiding and goal seeking agents.

References

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

Bates, J., Loyall, A. B., and Reilly, W. S. (1991). Broad agents. In *AAAI spring symposium on* *integrated intelligent architectures*. American Association for Artificial Intelligence. (Repr. in SIGART BULLETIN, 2(4), Aug. 1991, pp. 38–40).

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. and Sloman, A. (1993). A study of motive processing and attention. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., and Ramsay, A., editors, *Prospects for Artificial Intelligence*, pages 229–238. IOS Press, Amsterdam.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2:14–23. 1.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Cooper, R. and Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4):297–338.

Craik, K. (1943). *The Nature of Explanation*. Cambridge University Press, London, New York.

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

Davis, D. N. (1996). Reactive and motivational agents: Towards a collective minder. In Mueller, J., Wooldridge, M., and Jennings, N., editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*. Springer-Verlag.

Dennett, D. (1996). *Kinds of minds: towards an understanding of consciousness*. Weidenfeld and Nicholson, London.

Elliot, C. (1998). Hunting for the Holy Grail with 'Emotionally Intelligent' Virtual Actors. *SIGART Bulletin*, 9:20–28. 1.

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

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

Hatano, G., Okada, N., and Tanabe, H., editors (2000). *Affective Minds*. Elsevier, Amsterdam.

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

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In Haugeland, J., editor, *Mind Design*. MIT Press, Cambridge, MA.

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

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

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

Ortony, A., Clore, G., and Collins, A. (1988). *The Cognitive Structure of the Emotions*. Cambridge University Press, New York.

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

Popper, K. (1934). *The logic of scientific discovery*. Routledge, London.

Reilly, W. S. N. (1996). *Believable Social and Emotional Agents*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Technical Report CMU-CS-96-138.

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

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

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. (1992). Prolegomena to a theory of communication and affect. In Ortony, A., Slack, J., and Stock, O., editors, *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, pages 229–260. Springer, Heidelberg, Germany.

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

Sloman, A. (1993b). Prospects for AI as the general science of intelligence. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., and Ramsay, A., editors, *Prospects for Artificial Intelligence*, pages 1–10. IOS Press, Amsterdam.

Sloman, A. (1994). Explorations in design space. In Cohn, A., editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester. John Wiley.

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

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

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

Sloman, A. (1999a). Review of *Affective Computing* by R.W. Picard, 1997. *The AI Magazine*, 20(1):127–133.

Sloman, A. (1999b). What sort of architecture is required for a human-like agent? In Wooldridge, M. and Rao, A., editors, *Foundations of Rational Agency*, pages 35–52. Kluwer Academic, Dordrecht.

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

Sloman, A. (2000b). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In et al., M., editor, *Parallel Problem Solving from Nature – PPSN VI*, Lecture Notes in Computer Science, No 1917, pages 3–16, Berlin. Springer-Verlag.

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

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

Sloman, A. and Logan, B. (2000). Evolvable architectures for human-like minds. In Hatano, G., Okada, N., and Tanabe, H., editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam.

Wright, I., Sloman, A., and Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126. Repr. in R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000.