Presented at workshop on Emotions in Humans and Artifacts Vienna, August 1999
Published in *Emotions in Humans and Artifacts,* Eds Robert Trappl, Paolo Petta and Sabine Payr,
MIT Press, Cambridge MA, 2002, pp. 35–114.
NOTE:This version (produced 8 Sep 2006) has some improved diagrams and now makes a distinction between the CogAff schema and the H-Cogaff architecture, which is a special case of the schema. The idea of the schema was only implicit in the original paper but the distinction has been made clear in more recent papers here http://www.cs.bham.ac.uk/research/projects/c

HOW MANY SEPARATELY EVOLVED EMOTIONAL BEASTIES LIVE WITHIN US?

Aaron Sloman
School of Computer Science
The University of Birmingham
http://www.cs.bham.ac.uk/~axs
A.Sloman@cs.bham.ac.uk

Abstract

A problem which bedevils the study of emotions, and the study of consciousness, is that we assume a shared understanding of many everyday concepts, such as ‘emotion’, ‘feeling’, ‘pleasure’, ‘pain’, ‘desire’, ‘awareness’, etc. Unfortunately, these concepts are inherently very complex, ill-defined, and used with different meanings by different people. Moreover this goes unnoticed, so that people think they understand what they are referring to even when their understanding is very unclear. Consequently there is much discussion that is inherently vague, often at cross-purposes, and with apparent disagreements that arise out of people unwittingly talking about different things. We need a framework which explains how there can be all the diverse phenomena that different people refer to when they talk about emotions and other affective states and processes. The conjecture on which this paper is based is that adult humans have a type of information-processing architecture, with components which evolved at different times, including a rich and varied collection of components whose interactions can generate all the sorts of phenomena that different researchers have labelled “emotions”. Within this framework we can provide rational reconstructions of many everyday concepts of mind. We can also allow a variety of different architectures, found in children, brain damaged adults, other animals, robots, software agents, etc., where different architectures support different classes of states and processes, and therefore different mental ontologies. Thus concepts like ‘emotion’, ‘awareness’, etc. will need to be interpreted differently when referring to different architectures. We need to limit the class of architectures under consideration, since for any class of behaviours there are indefinitely many architectures which can produce those behaviours. One important constraint is to consider architectures which might have been produced by biological evolution. This leads to the notion of a human architecture composed of many components which evolved under the influence of the other components as well as environmental needs and pressures. From this viewpoint, a mind is a kind of *ecosystem*[1] of co-evolved sub-organisms acquiring and using different kinds of information and processing it in different ways, sometimes cooperating with one another and sometimes competing. Within this framework we can hope to study not only mechanisms underlying affective states and processes, but also other mechanisms which are often studied in isolation, e.g. vision, action mechanisms, learning mechanisms, ‘alarm’ mechanisms, etc. We can also explain why some models, and corresponding conceptions of emotion, are shallow whereas others are deeper. Shallow models may be of practical use, e.g. in entertainment and interface design. Deeper models are required if we are to understand what we are, how we can go wrong, etc. This paper is a snapshot of a long term project addressing all these issues.

1 What kinds of emotions?

The study of emotions has recently become fashionable within AI and Cognitive Science. Unfortunately all sorts of different things are labelled as ‘emotions’. This is perhaps understandable among young engineers who have not been trained in philosophy or psychology. However even among specialists there many different definitions of ‘emotion’ and related concepts, such as ‘feeling’, ‘affect’, ‘motivation’, ‘mood’, etc.

For instance some define emotions in terms of observable physical behaviours (such as weeping, grimacing, smiling, jumping for joy, etc.). Some define them in terms of measurable physiological changes which need

[1] The published version of this paper used the word ‘ecology’ here.

not be easily discernible externally, though they may be sensed internally (referred to by Picard as 'sentic modulation'). Some define them in terms of the kinds of conscious experiences involved in having them – their phenomenology. Some define them in terms of the brain mechanisms which may be activated.

Even when behavioural manifestations do occur they may be to some extent culturally determined, casting doubt on behavioural criteria for emotions. For instance the sounds people make when exhibiting pain can vary according to culture: 'ouch' in English is replaced by 'eina' in Afrikaans! Some researchers regard emotions as inherently social or cultural in nature, though this may be more true of having a guilty conscience than being terrified during an earthquake.

There is also disagreement over what sorts of evidence can be taken as relevant to the study of emotions. For instance, some will regard the behaviour of skilled actors when asked to show certain emotions as demonstrating connections between emotions and externally observable behaviour. Others will object that that merely reveals what happens when people are asked to act *as if* they had certain emotions, whereas naturally occurring emotions may be quite different. In some cases they may have no external manifestations, since people can often conceal their emotions.

For some researchers, emotions, by definition, are linked to and differentiable in observable behaviour, like weeping, grimacing, jumping for joy, growing tense, etc., whereas others are more interested in semantically rich emotions for which there are no characteristic, non-verbal, behavioural expressions, e.g. 'Being worried that your work is not appreciated by your colleagues' *vs.* 'Being worried that your political party is going to lose the next election', or 'Being delighted that the there is a sunny weather forecast for the day you have planned a picnic' *vs.* 'Being delighted that someone you admire very much is impressed by your research', etc. Most of the empirical, laboratory, research on emotions has studied only simple, shallow emotions, largely ignoring semantic content, whereas most of the important human emotions (the ones that are important in our social lives, and which are the subject matter of gossip, poems, stories, plays, etc.) are deep and semantically rich.

Another common difficulty is that some people use the word 'emotion' so loosely that it covers almost any affective state, including having a desire or motive, whereas in ordinary parlance we do not normally describe someone as being emotional just because they have goals, purposes, or preferences, or because they are enjoying a meal or finding their chair uncomfortable to sit in. If all such affective states were included as emotions, it would follow that people constantly have a large number of different emotions, since we all have multiple enduring goals, ambitions, tastes, preferences, ideals, etc.

Another source of confusion concerns whether having an emotion necessarily involves being conscious of the emotion. According to some this is a defining criterion, yet that does not square with the common observation that people can sometimes be angry, jealous, infatuated, or pleased at being flattered, etc. without being aware of being so, even though it may be obvious to others. Another problem with the criterion is that it may rule out certain animals having emotions if they lack the ability to monitor and characterise their own states or lack the conceptual framework required to classify some states as emotions. Presumably a new-born infant cannot classify its own mental states using our adult categories. Does that mean that it has no emotions? Perhaps it *has* them but does not *feel* them? Perhaps an infant's behavioural manifestations of pain, distress, discomfort, pleasure, etc. are simply part of the biologically important process of generating appropriate nurturing behaviour in parents rather than being expressions of what the infant is aware of? There is no obvious way of resolving disagreements on these issues because of the ambiguities and confusion in the key concepts used.

Yet another confusion concerns whether, in order to have emotions, an organism or machine must contain an emotion-producing module of some kind, or whether some or all emotions are simply states involving interactions between a host of processes which are not intrinsically emotional, as was argued in (Wright et al., 1996). On the first view it makes sense to ask how the emotion mechanism evolved, and what biological function it has, whereas on the second view such questions make no sense. Another possibility is that the ambiguous word 'emotion' sometimes refers to states and processes conforming to the first view, and sometimes to the second, because our usage is inconsistent.

Because of this conceptual mess, anyone can produce a program, label some component of it the 'emotion module' and proudly announce that they have developed a robot or software agent which has emotions. It will be hard to argue against such claims when there is no agreement on what emotions are. This is an extreme form of the phenomenon in AI of attributing mental states and human capabilities to programs on the basis of very shallow analogies, for which McDermott chided the AI community in (McDermott, 1981) many years ago, though he was concerned with the undisciplined use of labels such as 'plan', 'goal', 'infer'.

2 A modest proposal

Is there any way to remove, or at least reduce, the muddle and confusion? Our proposal is to step back from the problem of defining emotions, or other affective states, and then ask two related questions. The first question is what sorts of animals or artifacts we may be talking about, where different sorts are defined by *their information processing architectures*: the varieties of mechanisms and sub-mechanisms they have for acquiring, interpreting, transforming, storing and using information including taking decisions, initiating actions, generating and pursuing goals, and doing various kinds of self-monitoring. It is important that the kind of architecture under discussion is not defined by a *physical* mechanisms and their organisation, but rather by what a software engineer might refer to as a virtual machine architecture, which need have no simple relationship to the underlying physical architecture.

The second question can then be asked about each sort of architecture: what sorts of states and processes can occur in animals or artifacts which have this architecture? We may find that different kinds of systems can produce different sorts of things that might loosely be called emotions.

Having distinguished these different sorts of cases, we can then replace questions like ‘Can computer-based robots have emotions?’, ‘Can software agents have emotions?’ ‘Can insects have emotions?’, ‘Can newborn infants have emotions?’ with questions like ‘What sorts of emotions can various sorts of robots or software agents have?’, ‘Which kinds of insects can have which kinds of emotions?’, ‘Which kinds of emotions can newborn infants have, if any?’

We can then perhaps define a large number of different sorts of emotions: E1, E2, E3, ... E25, ... E73, ..., related to different types of architectures and different states and processes that can occur within those architectures. Moreover, having defined a certain class of emotions, E25, in terms of the architectures which make them possible, we can then ask ‘Which kinds of animals, or computing systems, can have emotions of type E25?’

All this assumes that the characterisation of types of organisms that is potentially the most fruitful for philosophical understanding, for scientific advance and for applications such as therapy and education, is one that refers the organism’s or machine’s information processing architecture. This amounts to adopting what Dennett (1978) referred to as ‘the design stance’. It is different from ‘the physical stance’ which describes the physical components and organisation of the system, and ‘the intentional stance’ which ignores internal mechanisms and merely assumes that it is predictively useful to regard certain behaving systems as acting rationally on the basis of beliefs, desires, intentions, etc.[2]

The idea of explaining mental phenomena in terms of information processing architectures which support them is not new. For instance, Herbert Simon’s highly recommended paper (Simon, 1967) written in the early 1960s, makes a start on the project of analysing requirements for an architecture capable of supporting human emotions and motivations. Since then there have been many proposals regarding architectures for intelligent systems, including, for instance, those described in the books by Albus (1981), Minsky (1987), Russell and Norvig (1995), Dennett (1996), Nilsson (1998), and my own early primitive efforts in my 1978 book.

Different architectures will support different collections of states and processes: different mental ontologies. Using the design stance, we can then define different sorts of emotions, different kinds of awareness, different kinds of learning, different kinds of intentionality, etc. in the context of the architectures which produce them. Then vague and fruitless debates about which animals and which machines can have emotions or consciousness can be replaced by more productive factual investigations: we can ask which kinds of animals and machines can have which kinds of emotions, which kinds of awareness, etc. So if we define twenty five classes of emotions produced by different architectures we may find that normal adult humans can have one subset, newborn infants another subset, chimpanzees yet another subset, particular sorts of robots yet another subset, etc.

For this activity we do not need to assume that the systems we are analysing are wholly or even partly rational, as would be required if we adopted the intentional stance, or if we attempted to build theories conforming to the ‘knowledge level’ defined in (Newell, 1982).

All this may help to counteract the fast-growing practice of publishing papers reporting yet another robot or softbot ‘with emotions’ where ‘emotion’ is defined so as to permit the author’s claim (often in ignorance of McDermott’s warnings). Instead authors will have to specify precisely which class of emotions they are referring to. We shall find that some are far less interesting than others, e.g. because they depend on relatively

[2] It seems to me from Dennett’s recent writings that over the years he has gradually shifted from stressing the intentional stance to adopting the design stance, which explains how a system works in terms of its (usually hierarchical) functional decomposition into interacting components, and which can justify ascribing mental states and processes to organisms and machines on the basis of how they work. I don’t know if he is aware of the shift or would even acknowledge it. His 1996 book for instance discusses designs for different kinds of minds. However, his definition there of the design stance is closer to McCarthy’s definition of ‘the functional stance’ in (McCarthy, 1995), which explains behaviour of a system in terms of what its function is, without reference to how it works.

shallow architectures, though they may be of some use for entertainment or educational purposes, or enhancing a user interface for inexperienced users.[3]

We can also explain, from this standpoint, the proliferation of definitions of emotion and theories explaining how they work. Human minds have extraordinarily complex architectures, with components whose effects are not clearly separable without a deep theory of the architecture. People with a limited view of the system will unwittingly focus on different aspects of the complete system, unaware of what they are leaving out, like the proverbial ten blind men saying what an elephant is on the basis of which part they can feel (tusk, trunk, ear, leg, belly, tail, etc.) Each has a small part of a large and complex truth. The same is true of most theories of emotions, and more generally most theories of mind.

3 Phenomena to be explained

In parallel with producing theories about explanatory architectures it is helpful to collect examples of many types of actual phenomena, so that we have a broad range of cases for the theories to explain. Of course, we may misdescribe some of the phenomena if we don't yet have good theories on which to base our categorisations. So we should try to use descriptions which avoid premature commitment to a particular type of explanatory architecture.

An example of a fairly neutral description that would be applicable to many animals might be *"When a large object is detected rapidly moving towards an organism it will rapidly switch to some kind of evasive action."* We do not need to assume that such generalisations express universal laws: they will have many exceptions, including individuals which are injured or not fully awake, or which freeze instead of jumping to one side. There is also an inherent vagueness in the terms 'large' and 'rapidly'. So, since small or slow moving objects may trigger different kinds of behaviour, e.g. repulsive, or blocking or catching behaviour, the generalisation may have exceptions relating to intermediate cases, which are neither definitely small nor definitely large, or neither definitely rapid nor slow in their motion.

The phenomena to be explained can include both commonplace examples that most people know about from their own experience, observing others, gossiping, reading novels etc, and also phenomena that are discovered by experts from field studies, or in various 'unnatural' laboratory situations, and also information derived from studies of effects of brain damage, or in ethological studies of other animals.

We should try to build a theory which explains all such phenomena, while respecting constraints from neuroscience, psychology, biological evolution, as well as what has been learnt from computer science, software engineering and AI about feasibility, tractability, etc. of various solutions to problems.

Of course, there may be disagreements regarding which phenomena to include as initial constraints. Some researchers may wish to work bottom up from the most fine grained empirical observations of behavioural details, or physiological changes, or brain structures and processes, while others may wish to start from more global characterisations of behaviours and their interactions and the global designs that might explain some of their features. The former involves attempting to design mechanisms that replicate or model the detailed phenomena, worrying later about how to put them together in the design of a complete organism, whereas the latter approach involves attempting to design a complete working system with a wide range of capabilities which at first merely approximate to the original phenomena, and then progressively adding more and more detailed mechanisms to match the finer grained requirements. A mixture of bottom-up and top-down approaches is also possible. Our own work is primarily top-down (aiming simultaneously for progressive deepening and progressively increasing breadth of the architecture) though we attempt to learn from those working in bottom up mode, especially people doing research on brain functions and mechanisms.

4 Pitfalls

A danger in the bottom-up approach is that the mechanisms produced may be capable of performing very specific tasks, but be incapable of contributing adequately to the functioning of a larger architecture. An example might be a visual feature detector which analyses an input image and produces as output another image with the required features (e.g. edges) highlighted. This would be an inadequate mechanism if a complete visual system requires the output of an edge detector to be something quite unlike a modified image.

A danger in the top-down approach (sometimes referred to, following (Bates et al., 1991) as the 'broad and shallow approach') is that it can lead to designs which are incapable of being refined in ways that match the

[3] E.g. I have used a very shallow model for tutorial purposes in my introduction to the SIM_AGENT toolkit: http://www.cs.bham.ac.uk/research/poplog/sim/teach/sim_feelings

fine grained detailed specifications added later. For example, suppose a model of a complete agent includes a specific sub-mechanism described as an 'emotion-generator', perhaps assigning varying numerical values to a collection of emotion descriptors. It could turn out that this sub-mechanism was totally spurious if the actual emotional phenomena found in organisms are not produced by a specific emotion generator but are emergent features of the interactions of other mechanisms with different functions. (An analogy would be supposing that the thrashing of an operating system had to be produced by a 'thrash-generator' mechanism, rather than being an emergent feature of the interactions of other mechanisms under conditions of heavy load.)

A danger common to all approaches is that they can lead to excessive focus on one particular design or architecture. This could be a serious mistake if there is considerable variation in architectures, e.g. between different kinds of animals or even between humans at different stages of development. For instance, it is likely that newborn human infants do not have the same sort of architecture as adults. Apart from the fact that any particular design might correspond to only a small subset of the naturally occurring cases, it is also the case that we cannot fully understand the functional importance of particular designs unless we compare and contrast them with alternative designs. In other words, for a full understanding of a particular architecture we need to explore a neighbourhood in 'design space' (Sloman, 1993b).

More specifically, we should explore architectural variation:

- Across the natural/artificial divide
- Across species,
- Within species,
- Within an individual during normal development
- Caused by brain damage

5 Objectives of the analysis

In what follows I shall sketch in schematic outline a hypothesised architecture which has components which have evolved at different times, with older portions found in many organisms and newer portions in relatively few animals. It is conjectured that the oldest layer is a purely reactive layer. A newer portion is a deliberative layer capable of creating descriptions of possible but not yet actual situations, i.e. doing 'what-if' reasoning. The newest portion is the meta-management (or reflective) layer which is a collection of mechanisms for monitoring, categorising, evaluating, and (to some extent) controlling the rest of the system. The third layer provides a type of self-consciousness that was previously lacking. We do not share the assumption made by Rolls (1998, forthcoming) that the third layer necessarily requires the use of a human-like external language, though we acknowledge that its functionality will be considerably modified by the existence of such a language. It is possible that some other primates have meta-management capabilities without having an external, human-like language. However all information processing mechanisms, in all animals, require some type of formalism for encoding information whether it is factual information or control information. How many such formalisms there are, on what dimensions they vary, and what they are useful for is an important (and difficult) topic of ongoing research e.g. (Glasgow et al., 1995; Karmiloff-Smith, 1996; Peterson, 1996; Sloman, 1996b).

5.1 Architecture-based concepts of emotions

On the basis of the hypothesised architecture it is possible to produce a provisional crude classification of many familiar types of emotional processes into three different classes (a) *primary emotions* which depend only on the oldest reactive subsystems, (b) *secondary emotions* which depend in part on newer deliberative mechanisms and (c) *tertiary emotions* (previously referred to as 'perturbances' in (Wright et al., 1996)), which depend on interactions with the recently evolved reflective or 'meta-management' layer (argued for in (Beaudoin, 1994)). Within the three classes we can distinguish further sub-classes depending on precisely which sub-mechanisms are involved.

We can also distinguish different types of perceptual mechanisms of varying degrees of sophistication and different types of action mechanisms. Then, just as different classes of emotions are supported by different architectures, so also are different classes of perceptions and actions, some much richer and more flexible than others.

Likewise if there are different sorts of mechanisms for acquiring, storing and later on re-using information, we can use the differences to define different classes of learning, and then ask which classes occur in which animals, or which occur at different stages in human development.

All this provides a framework in which concepts of different sorts of mental states and processes are defined in terms of the kinds of information processing architectures they presuppose and the processes which occur in those architectures. By specifying architectures at the virtual machine level familiar to software engineers rather than at the physical level, and defining mental concepts in terms of states and processes within virtual machines, not physical machines, we avoid mistakes of materialist reductionism, though that is not a topic for this paper.

5.2 Types of variation of architectures

There are several different ways in which sub-mechanisms in an architecture may vary from one design to another. At the lowest level there may be differences in the type of physical implementation, e.g. whether the system uses biological mechanisms, such as neurons, or computer-based mechanisms. In the latter case there may be differences between systems implemented directly in hardware and those implemented using software which is interpreted by other software or possibly directly by the hardware. At present we know very little about the implications of using one or other form of bottom-level implementation though much prejudice abounds as to whether using non-biological components does or does not rule out accurate replication or simulation of human mental states and processes. (For many people the desire for a particular answer to be correct gets in the way of investigating which answer is actually correct.)

Whatever the lowest level implementations used, there are also differences between the forms of representation available to the system, e.g. whether it simply has a fixed collection of state registers whose values can vary over time; whether it is capable of using distributed overlaid representations, such as the rules encoded non-locally in weights in certain kinds of neural nets; whether it can use forms of representation with a recursive syntax, so that structures of varying degrees of complexity can be created as needed; whether it can use hybrid modes of representation; whether it can discover the need for new forms of representation and start using them, etc. Moreover for a given general type of representation there may be considerable variations in the type of semantic content which organisms can use those representations to encode. In particular, organisms which need different information about the environment will typically be restricted in what sorts of objects, properties, relations, etc. they can refer to. I.e. there may be differences in the ontologies available to different organisms, or to different components of an organism. For instance, a chimpanzee and a flea close to it can be expected to be capable of acquiring, storing, and using quite different sorts of information about the environment. This is partly because they *need* different sorts of information.

Less obviously, similar comments may be applicable to different sub-components within the same organism, e.g. components which evolved at different times and perform different functions.

For instance, within humans one of the functions of our visual system is to provide information for posture control mechanisms, including information about optical flow patterns in the visual field which can be clues to whether the individual has begun to fall forward or backward. Such information may be directly encoded as control signals causing various muscles to contract or relax to correct the detected motion. The posture control subsystem may be incapable of receiving or using any other kind of information about what is in the environment.

Similarly where complex actions, such as grasping or catching a moving object, require fine-grained control of motion using a visuo-motor feedback loop, the information detected by the visual system for this purpose may be quite different from the information needed for another purpose, e.g. formulating a plan of future action, or recognising a person's facial expression. The different kinds of information may be processed by different sub-systems even though they all get their information from the same optic array. (These issues were discussed in more detail in (Sloman, 1989; Sloman, 1993a; Sloman, 1996a).)

In a later *section 12*, we shall return to a more detailed discussion of dimensions of variation of architectures, and ambiguities in descriptions of architectures by different researchers.

6 Can we expect intelligible structure?

It is possible in principle that any attempt to build a theory of the human information processing architecture is doomed because the actual architecture is a completely unstructured mess, as suggested in *Figure 1*. However, I suggest that in producing a really complex, multi-function design, evolution, like a human designer, will be constrained to use a modular organisation, e.g. so that changes to improve one module will not impact disastrously on functionality of other modules, and also so that some economy of genetic encoding is possible for a design using several copies of approximately the same module. (This argument does not apply to evolution of relatively simple designs, where there is less need for functional decomposition to aid the search for a solution.)

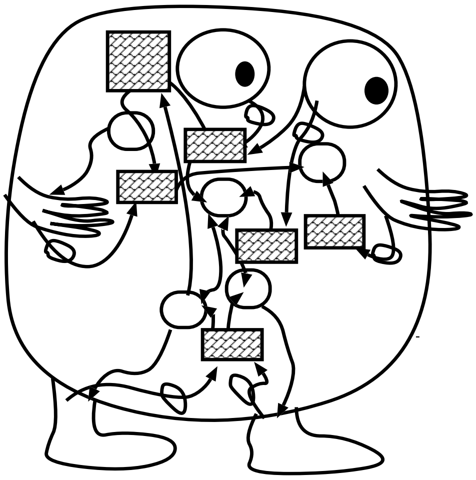

Figure 1: Could the architecture be an unintelligible mess?
The ovals represent processing components and the shaded rectangles information stores or buffers. In principle, there could be very large numbers of such components inter-connected in such a complex and unstructured fashion that humans could never comprehend it.

It is important that in this context ‘modular’ does not mean or imply ‘rigid’ or ‘innate’. As (Simon, 1969) pointed out, the boundaries between modules need not be sharp and clear: a complex system may be ‘nearly decomposable’. Moreover during learning and development new modules may be added, and boundaries can change. Despite these qualifications the claim that the design is modular contrasts with the claim that there is no intelligible structure in the architecture.

An intermediate position can be found in (Fodor, 1983), where it was suggested that the central cognitive component of the human information processing architecture is an unintelligible mess, although it has various ‘encapsulated’ sensory and motor modules connected to it as indicated schematically in *Figure 2*. Fodor’s idea was that the encapsulated components were largely determined innately, and their mode of functioning could not be modified by cognitive processes.

In (Sloman, 1989) I discussed and criticised Fodor’s model along with the theory of vision proposed by David Marr in his very influential book (Marr, 1982), which proposed that a visual system functions largely in data-driven mode, extracting information from the visual array which is then used to construct descriptions of objects in the environment in terms of their geometrical and physical properties, e.g. shape, location, motion, colour, texture, etc. Marr’s work essentially took one of the petals from Fodor’s sunflower and expanded it by dividing it into processing layers through which data-flowed, starting from retinal images passing through various intermediate databases, such as a database of edge features, a binocular disparity map, an optical flow map, a texture feature map, a primal sketch, a 2.5D surface feature database giving orientations of surface fragments, viewer centred object models, and scene-centred object descriptions, ultimately feeding into the cognitive system information about the shapes, motion, colour, and texture, of objects in the environment, along with hierarchical structural descriptions in some cases, and if object-types are recognised, their categories, e.g. ‘dog’, ‘horse’, ‘bucket’.

This *one-way, data-driven, geometry/physics-based* view of a visual system’s architecture can be contrasted both with models (e.g. (Albus, 1981; Sloman, 1989)) which allow perceptual systems to be partly driven in top down mode, influenced by current needs and expectations, and also with the ideas of the psychologist Gibson (1979), according to which instead of being constrained to produce descriptions of the geometrical and physical properties of objects in the environment, a sensory system can produce descriptions which relate to the needs and capabilities of the organism. Gibson called these ‘affordances’. E.g. vision may inform an animal of *positive* affordances such as support, graspability, the possibility of passage, access to food, etc. and *negative* affordances, such as obstruction, risk of injury, etc.

The richness and diversity of the output of a human visual system can be seen both from the descriptions we use of how things look (e.g. ‘He looks angry,’ ‘That spider web looks fragile,’ ‘That bar looks as if it is holding up the shelf’, ‘That knife looks dangerous’, ‘That equation looks solvable’) but also from the varieties of visual ambiguities in pictures illustrated in *Figure 3* and others found in textbooks on vision, such as (Frisby, 1979). This suggests that a visual system can include many subsystems able to detect and report information of very different kinds. In (Sloman, 1989) I proposed that some of those subsystems instead of producing *descriptions*

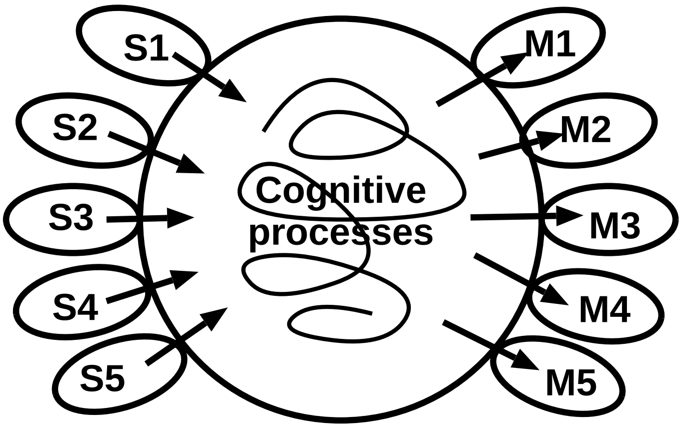

Figure 2: Fodor's modular 'sunflower' architecture
Fodor's 1983 book suggested that humans have a collection of 'encapsulated' sensory modules, shown as S1, S2, etc. feeding information into a complex and messy central cognitive system, which in turn sends signals to a collection of encapsulated motor mechanisms.

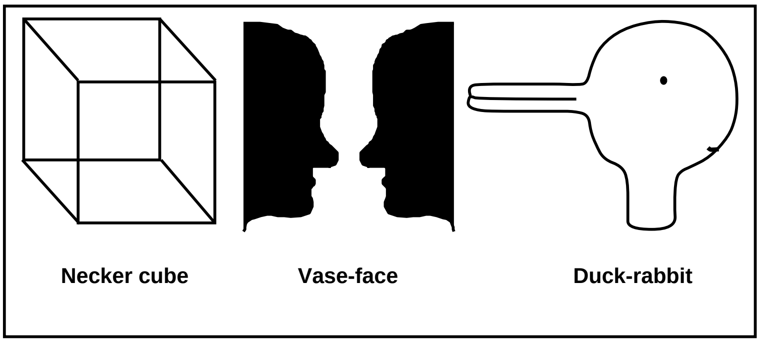

Figure 3: Various kinds of visual ambiguity
The necker cube has two interpretations whose differences are purely geometrical (involving distances and angles). The vase-face figure has interpretations differing both in subtle figure-ground relationships, and also in which objects are recognised. The duck-rabbit figure also has an ambiguity regarding the direction in which the perceived animal is facing, a very abstract property which could be perceptually relevant to predators or prey – i.e. an affordance.

produced *control signals*, e.g. signals to trigger saccades or signals to the posture control system. Some of these different visual sub-systems evolved early in our evolutionary history. Some evolved much later. Some, e.g. modules for sight reading music, are produced by individual learning rather than evolution. Later I'll relate all this to a conjectured architecture for adult humans.

7 A schematic overview of the architecture

The particular sort of architecture we have been developing within the Cognition and Affect project[4] is rather complex and hard to understand. In what follows I shall start from two simplified views of the architecture and then combine them in such a way as to provide a mnemonic overview, in the form of the CogAff Architecture Schema depicted below, which allows for many special cases.[5] However the combined picture is still an oversimplification, as various additional modules are needed to complete the specification. These will be described very briefly, and the whole thing related to various kinds of mental states and processes to be found

[4] Our implementations are still far behind our theoretical analysis, partly because of lack of resources and partly because there are still many gaps in the design.

[5] The distinction between the CogAff architecture schema and the special instance, the architecture now referred to as H-Cogaff, described later, was implicit in the published version of this paper. A few small changes were made in September 2006 to make the distinction clear.

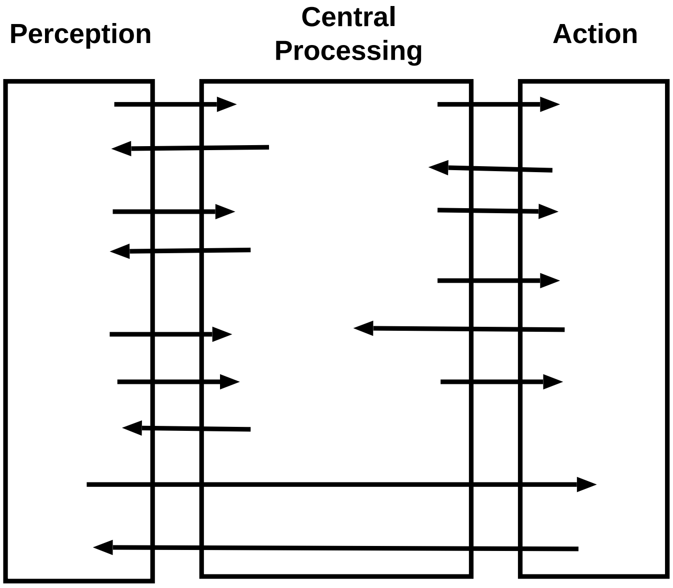

Figure 4: The ‘triple tower’ model of information flow
This is a modified version of the ‘triple tower’ model in (Nilsson, 1998). Arrows represent flow of descriptive or control information. Such models can vary according to how much and what kind of processing is allowed in the perceptual and action modules, and whether information flow is unidirectional or includes internal feedback or ‘hypothesis driven’ or ‘goal driven’ control.

in humans.

We can also view other animals as sharing different subsets of the architecture, though possibly with different design and implementation details since they have to operate with different larger systems. For instance, mechanisms within a bird of prey for perceiving and using optical flow may be different from the optical flow mechanisms in a tiger or a human.

7.1 The triple tower model

The first schematic view of a behaving system is based on the notion of flow of information through an organism (or robot) from sensors to motors, with some intervening processing mechanisms. A simplified version of this model is shown in *Figure 4.* It assumes a division between sensory mechanisms, central processing mechanisms (possibly including changeable long term sates), and action mechanisms.

Models of this type can vary according to how ‘thick’ the perceptual and action pillars are, e.g. how much processing occurs between physical input and creation of central representations, and also whether the perceptual mechanisms are *stratified*, i.e. with different types of sensory or motor processing occurring in parallel, dealing with different levels of abstraction.

In very simple cases (single-celled organisms?) a perceptual mechanism will be little more than a physical transducer, transforming external energy into a form that can be processed internally, and likewise action mechanisms may simply consist of motors which receive signals from the centre, perhaps in the form of control vectors. We have already suggested that in humans visual processing is far more complicated than that, including at least the sorts of layers or analysis and interpretation described by Marr, and possibly also additional layers allowing central control or modulation of the processing, so that the visual system is not an encapsulated package as suggested by Fodor.

Likewise, action mechanisms may include not only direct and simple signals to motors, but also, in sophisticated organisms, a control hierarchy in which high level symbolic instructions can be decomposed into instructions for several subsystems to execute concurrently.

There may be some organisms which are too simple or unstructured to be usefully represented by the triple tower model, e.g. single-celled organisms. This paper does not discuss them, though a more complete investigation of architectures should.

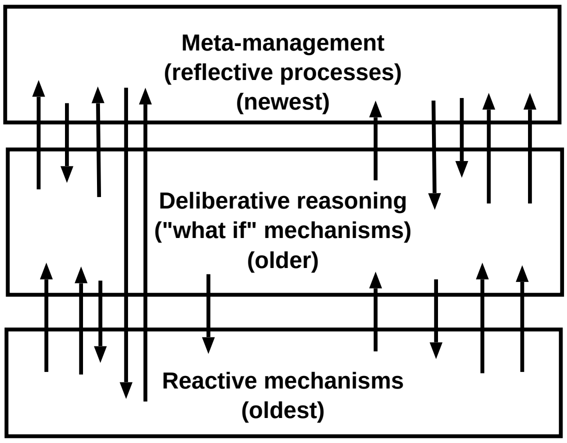

Figure 5: A triple layer model

Multi-layer models can vary according to the mechanisms and functions of the different layers, what sorts of descriptive and control information can flow up and down between the layers, whether the layers operate concurrently and asynchronously, and whether there is some sort of dominance hierarchy, e.g. with high levels dominating lower levels. Various levels can be added, e.g. by splitting off a 'reflex action' level below the reactive layer (as in (Davis, 1996)), or by splitting the top level into different layers. E.g. Albus (1981) proposes a dominant top level labelled 'Will'. Our model assumes a division related to evolutionary age, mechanisms used, and functional role, with all layers operating in parallel, asynchronously, all with sensory input and motor output connections. Differences between the layers are explained later.

7.2 Multi-level systems

Another common architectural partition, depicted in *Figure 5* is concerned with a collection of processing *layers* which lie between sensors and motors and operate at different levels, where the levels may differ in type of processing, or in degree of power over lower levels, or in other respects. The idea is quite old in neuroscience. 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. 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; Nilsson, 1998).

In some theories that propose different architectural layers it is assumed that information comes from sensors via the lowest level which then passes information up the hierarchy until the topmost level takes major decisions about what to do next and sends control signals down the hierarchy to the lowest level which then drives motors, a sort of Ω model of information flow (e.g. Figure 15.2 in (Nilsson, 1998)). By vertically stretching the central circle in Fodor's architecture (*Figure 2*) and adding some internal structure we can see that the Omega architecture and Fodor's model have much in common.

Many such systems are based on engineering designs motivated by some set of practical requirements. Our architecture is motivated largely by analysis of requirements for living organisms, especially humans, including the requirement that the system should have been produced by evolution (i.e. via an *e-trajectory* in the terminology of (Sloman, 1998b)).

Humans are able to carry out many skilled tasks involving perception and action, e.g. walking, driving a car or operating a machine, at the same time as performing other more cognitively demanding tasks such as deliberating about where we should go next, or holding a conversation, or trying to assess and evaluate our state of mind (e.g. 'Am I feeling too tired to go on driving?'). This suggests that there are subsystems which can to some extent behave as complete organisms, acting in parallel with other sub-organisms. In previous papers, e.g. (Beaudoin, 1994; Wright et al., 1996; Sloman, 1998a; Sloman, 2000a) we have explained some of the main features of the reactive, deliberative and meta-management (reflective) layers.

Within this sort of architecture it is not assumed that higher level systems can totally dominate the others, though they may have partial control over them. This distinguishes our system from a 'subsumption' architecture (Brooks, 1991). Another difference is the ability of the second layer to perform 'what-if' reasoning about chains of possible future actions in order to form plans along with the closely related ability to reason about possible past actions in order to form either explanations or analyses of previous successes and failures.

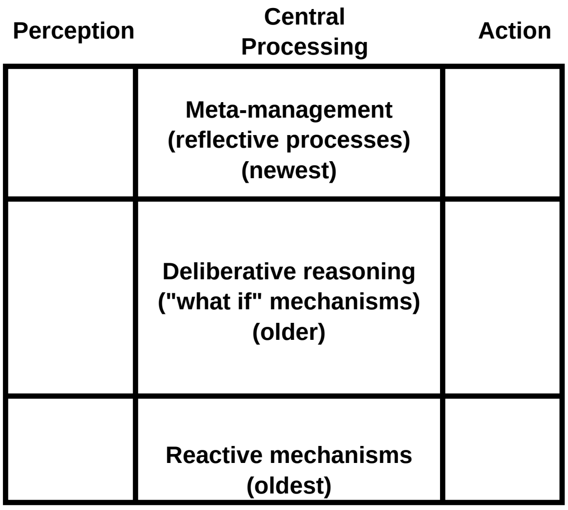
Perception
Central Processing
Action
Meta-management (reflective processes) (newest)
Deliberative reasoning ("what if" mechanisms) (older)
Reactive mechanisms (oldest)

Figure 6: Combining the tower and layer views in the CogAff schema
This gives a schematic overview of some of the functional differentiation within an architecture. Components not shown are described in the text and later figures. The CogAff schema represents many possible architectures, depending on which of the boxes are occupied and what sorts of connections there are between sub-mechanisms.

It is worth emphasising that although only one of the three layers is reactive, everything ultimately has to be implemented in reactive mechanisms. It should not surprise anyone familiar with computing systems that a mechanism of a certain type can be implemented in a mechanism of a very different type. In particular, mechanisms which consider actions before performing them can be implemented in mechanisms that do no such thing.

7.3 Combining the two views

To indicate how the the triple tower and triple layer models can be combined we superimpose the previous diagrams to produce the CogAff schema depicted in *Figure 6* which indicates more clearly that both the sensory and the action sub-systems may have different components operating in parallel at different levels of abstraction and feeding information to or getting information from different levels in the central three-layered system.[6]

The general idea is that as higher level cognitive functions developed they required more sophisticated perceptual input, producing evolutionary pressure for the development of perceptual systems able to operate with more abstract categories, as described above. For example Rolls remarks that view-invariant object recognition (perhaps using object-centred forms of representation) is much less well developed in non-primates, e.g. rodents. The more abstract representations have the advantage of generalising over a wider range of conditions (along with the disadvantage of providing less information about the details of any particular instance, though that might be compensated for by additional mechanisms operating in parallel).

The more abstract percepts are needed for chunking of information in a form that supports the learning of associations that can be useful in deliberative reasoning. If handled by fast dedicated perceptual mechanisms they can support faster high level reactions, compared with having to apply *general purpose* inference mechanisms to results of low level processing. Such perceptual mechanisms can process biologically important information that is often required, e.g. whether another animal is angry, pleased, afraid, or where it is looking (as in *Figure 3*). Some of the more abstract perceptual tasks may be given very specific sub-mechanisms dedicated to those tasks, e.g. face-recognition in humans. (Precisely how all this high level functionality is achieved remains a hard research problem, especially the problem of characterising the many functions of vision and how they are achieved.)

Similarly it may be useful if higher level mechanisms are able to give abstract commands to an action system, leaving it to highly trained or evolved sub-mechanisms in an action control hierarchy to produce the detailed execution, sometimes simply by modulating actions generated by lower levels (walking faster to catch up with someone), and sometimes by replacing them e.g. stopping to ask someone the way. To illustrate the

[6] Note added Sept 2006: As indicated in an earlier footnote, the distinction between the CogAff architecture schema and the special instance, the architecture now referred to as H-Cogaff, described later, was only implicit in the published version of this paper.

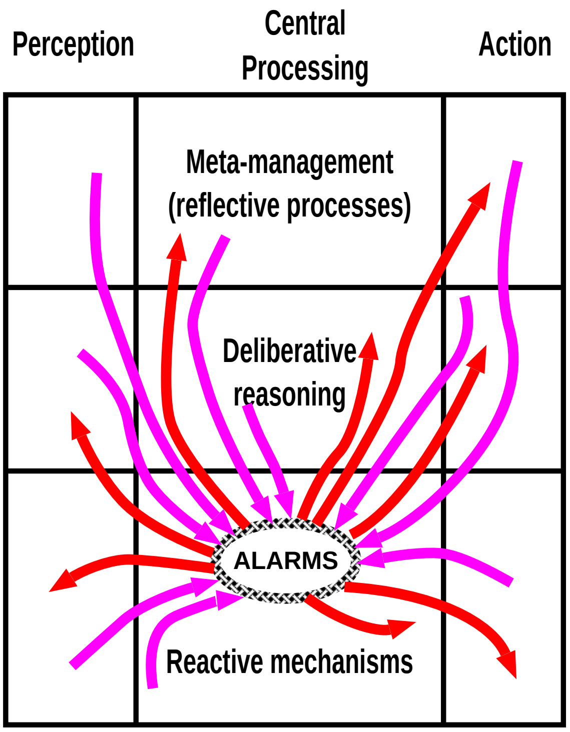

Figure 7: Adding alarm mechanisms

If there are contexts in which producing a behavioural response by 'normal' means takes too long it may be be necessary for a very fast (trainable) pattern-directed 'alarm' system to be able to interrupt and redirect other parts of the system. The figure shows only one global alarm system, though there may be several more specialised alarm systems.

power of specialised action routines compare trying to write your name with your normal writing hand and with the other hand. Although you have full knowledge of the task it is much harder to do it with the unusual hand, unless you first spend a lot of time practising, i.e. training the appropriate action mechanisms. Knowing exactly what you want to write and being able to control your muscles is not enough!

On this theory we can think of many of the mechanisms in the various boxes as if they were sub-organisms which have co-evolved in such a way that the needs and capabilities of each sub-organism applied evolutionary pressure influencing the development of other sub-organisms. (For a similar view see Popper, 1976.) For instance, development of mechanisms for standing and walking with an inherently unstable upright posture on two legs produces pressure for the sensory systems to produce perceptual information that can rapidly detect motion that indicates a need for corrective action. So in humans the visual system can use information from optical flow to send signals to posture control sub-systems (Lee and Lishman, 1975). This can happen in parallel with performing other visual functions, such as recognising objects, people and affordances in the environment and also providing visual feedback for grasping actions, or direction control while walking. This model implies that information coming in via a particular sensory subsystem may be processed concurrently in several different ways, and that different kinds of information will be routed in parallel through various different sub-systems, according to their current needs. This 'labyrinthine' architecture (Sloman, 1989) contrasts with the 'one-way , data-driven, geometry/physics-based' view of a visual system mentioned above. It is also consistent with recent discoveries of multiple visual pathways (e.g. (Goodale and Milner, 1992)) though the ideas presented here suggest that far more functionally distinct sensory pathways exist than have been discovered so far, including linkages between mechanisms for different sensory modalities which enable them to share some tasks, such as mutual disambiguation.

The task of cataloguing the full variety of perceptual functions in humans and other animals has barely begun. This seriously restricts our ability to design human-like or, for instance, squirrel-like, robots. In particular, completing the task will require a full analysis of the many kinds of perceptual affordances that need to be detected to meet all the needs of different sub-systems of the organism. (Some of the more complex affordances involving perception of possibilities and impossibilities are discussed in (Sloman, 1996a).)

8 The need for ‘alarm’ systems

Reactive mechanisms are generally faster than deliberative mechanisms because, as explained in (Sloman, 2000a), deliberative systems are inherently discrete and serial. They have these characteristics both because they need access to a re-usable short term memory structure and because they repeatedly need to access the same associative memory store, e.g. to answer questions like ‘Which actions are possible in situation S?’ and ‘What will happen if action $A3$ is performed in situation S?’

Reactive systems can sometimes also be relatively slow, if they use a lengthy sequence of chained internal reactions between sensory input and motor output, for instance in the teleo-reactive systems described in (Nilsson, 1994).

In some situations, even a delay of fractions of a second may make the difference between catching or losing an opportunity to eat, or to avoid injury or death. For this reason it may be useful, even in a purely reactive organism, to have an ‘alarm’ mechanism that is relatively fast and which, either as a result of genetically determined design or as a result of previous learning from experience, or both, can very rapidly detect a need for some change in behaviour and very rapidly produce that change by sending control signals to other sub-systems. Such an alarm system is schematically depicted in *Figure 7*, where the alarm system receives inputs from all the main parts of the architecture and can send its outputs to all the main parts. If the alarm mechanism uses an appropriate neural net architecture with not too many layers it can function very quickly, although being fast rules out performing complex deductions, exploring alternatives, etc.

This means that such an alarm mechanism will necessarily be relatively stupid (compared with a deliberative system or a more sophisticated reactive system), and it can be expected sometimes to get things wrong. However, if in situations where global interrupts are triggered false positives are generally not as serious as failures to react quickly to dangers and opportunities, the advantages of a fast and stupid alarm system will outweigh the disadvantages.

An alarm system may trigger various kinds of global responses, including freezing, fleeing, attacking, rolling up in a ball, rapidly jumping sideways to avoid a fast approaching large object, or changing direction or speed if already in motion. More subtle reactions might be changes in high level attention control systems (e.g. attending to some visual or auditory stimulus), or changes in physiological or cognitive state to enhance *readiness* for some kind of action or some kind of cognitive process, such as rapid decision-making, or planning. Some of these will be chemical changes either localised or throughout the organism. Other changes will be muscular, or changes in heart rate, etc. These physiological phenomena underly various familiar emotions. However, as we shall see they are not essential to *all* types of emotions.

It is worth noting that although the figure depicts the alarm system as if it were a separate architectural sub-system, its specification is actually consistent with requirements for the reactive layer in the architecture. So alarm mechanisms can also be seen simply as sub-components of the reactive layer in the architecture – which is how they were construed in some of our earlier papers. However they typically differ from other reactive components in receiving inputs from many parts of the organism, and being able to send *high priority* control signals to many parts of the organism. Insofar as this means that a reactive mechanism can sometimes dominate the other layers it differs from a simple subsumption architecture, as mentioned above, though some of the principles are similar.

Empirically, it seems that in nature many organisms have one or more such mechanisms. In humans there are various innate reflex mechanisms which serve this type of purpose, some of them very narrow in their field of influence such as the blinking reflex triggered by rapid motion, which can protect eyes from danger. Athletes, boxers, drivers of fast cars and aeroplanes, musicians who sight-read musical scores, all illustrate the existence in humans of very fast trainable reflexes, some of them involving very sophisticated information processing. Trainable fast reactive systems seem to be in the brain stem, the cerebellum and the limbic system, and no doubt elsewhere in the brain.

Precisely how to define the boundary between alarm systems and other fast reactive systems is not clear, though not all will have the global control capability required for an alarm system as indicated in the diagram. They certainly need not all be innate: in a complex and changing environment, it may be useful for organisms to be able to learn which situations need a fast global response.

Later we shall link some of the processes in reactive systems, especially processes involving alarm mechanisms triggered by sensory inputs or physiological state monitors, and which produce global physiological changes, to a subset of emotions, namely the *primary emotions*. If global alarms are triggered by events in the deliberative and meta-management layers they are *secondary emotions*. The subset of cases in which the effect of an alarm mechanism, or other ‘interrupt’ mechanisms, is a disposition to override control at the meta-management level will be described as *tertiary emotions*. All this provides only a crude, provisional architecture-based distinction among types of emotions, and some further sub-divisions will be indicated later,

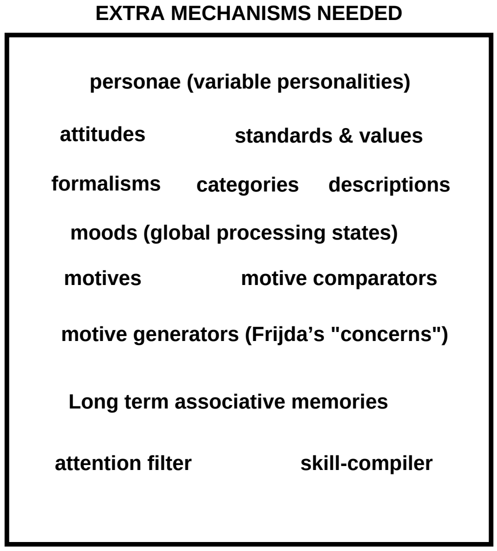

Figure 8: Additional components required
***For the systems described previously to work, several additional sub-systems are required, described briefly in* section 9.**

though this paper cannot present a complete overview.

9 Additional components

The nine-fold architectural decomposition previously outlined does not amount to a specification for a functional system. There are many details missing, listed in *Figure 8*. This section gives a brief overview of some of them.

9.1 Explicit and implicit goal mechanisms

An organism may have implicit or explicit goals. A goal is implicit when detection of a situation directly triggers an appropriate reaction, though generally there is no *unique* implicit goal. The reaction may have been designed or may have evolved to achieve some result, but without using any explicit representation of that result. Where there is an explicit, enduring, representation which can be used to check whether the current situation matches the goal specification, we can say that there is an explicit goal.

If sensors detecting contact with a very hot object trigger a withdrawal reaction, then the implicit goal may be to avoid damage from the hot object, or to move to a state where there is no longer contact with the hot object. (In general, appropriate descriptions for implicit goals are not uniquely determined by the condition-action relationship.)

For a very simple organism swimming in a soup of nutrients, detecting a shortage of energy or some other resource may trigger behaviour such as opening the mouth so that nutrients are swallowed, perhaps along with orientation or movement controlled by sensors which detect density gradients for the nutrients.

In some cases, however, detecting a need does not automatically determine action required to satisfy the need. Which action is appropriate will depend on the context, which may have to be ascertained by examining the environment, searching information stores, etc.

For instance, if a need for food is detected, the organism may have to use knowledge to work out what to do in order to get food. If the knowledge is not available, then that determines another need, namely to acquire knowledge that can then be used for the purpose of working out how to get food. Even if all the required knowledge is available, e.g. that the food is buried in a particular location some distance away, the achievement of the goal of consuming food may require many intermediate steps.

It is possible for a purely reactive system to cope with these cases, provided that it already has a suitable collection of reactive behaviours. Detection of energy shortage may trigger production of an explicit internal state which effectively represents the goal of obtaining food. The new state has that representational function insofar as it helps to trigger additional behaviours which are terminated only when food is found, which in turn triggers that representation to be removed (or de-activated) and the food to be eaten.

All this may involve several intermediate reactive steps, because the existence (or high level of activation) of the goal representation may repeatedly trigger actions determined by the combination of the goal and the changing context. This can continue until the current situation 'matches' the goal, e.g. food is reached. This reactive goal driven behaviour can work because the organism in effect has a pre-stored plan for achieving the goal. The plan is implicit in a large collection of condition-action rules.

However, if there are situations in which the pre-stored plans are not adequate, a planning mechanism may be needed for working out a sequence of actions that can achieve the goal.

9.2 Plan-formation in reactive systems

Organisms with a sufficiently rich and varied evolutionary history may have evolved a large enough collection of such plans to cope with most problems that arise in their environment – e.g. insects and perhaps many other animals. However there are problems if the evolutionary history of the type of organism is not adequate, or if the genetic structures are not large enough to store all the plans required, or if the individual brains are not large enough for them.

A partial solution would be for individuals to be born with a partially specified reactive plan store that has surplus capacity. Then various kinds of learning (e.g. learnt sequences of responses, such as can be used to train performing animals) might encode new plans that can thereafter be used when the need arises, in that sort of environment.

The implicit reactive plan-following mechanisms may be more or less sophisticated depending on the formalism available for expressing goals and various conditions and actions. For example, the goal formalism may simply be a collection of on-off flags (perhaps allowing different activation levels between "off" and fully "on"). Alternatively the goal states may include parameters which can be bound to different values at different times, introducing more flexibility in the selection of appropriate actions. In some cases it may be possible to formulate syntactically complex goals (e.g. 'find food of type X not further away than distance D'). However, syntactic complexity is more likely to be found in deliberative mechanisms.

9.3 Deliberative mechanisms and what-if reasoning

In some environments the previously mentioned partial solution to the need for novel plans is inadequate, e.g. because the training process in reactive systems is too slow or too dangerous. Training a purely reactive system is satisfactory if the environment provides suitable 'safe' contexts for learning. A deliberative layer is required only when evolutionary history and individual training opportunities do not provide a wide enough variety of plans which can be learnt safely. Forms of learning which require trial and error, or reward and punishment, are not safe where the error or punishment entails death or serious injury. A deliberative mechanism enabling analysis of future possibilities can sometimes provide a safer way of discovering useful new plans, as Craik pointed out in 1943, and various others have also noticed since. This allows our 'hypotheses to die in our stead' as the philosopher Karl Popper noted. (See also (Dennett, 1996).)

This requires 'what-if' reasoning mechanisms. Here, instead of a new goal immediately triggering behaviours in the reactive layer, it can trigger planning behaviour in a deliberative layer. This typically involves exploring combinations of future possible actions until one is found that achieves the goal, or at least achieves a state in which it may be easier to achieve the goal. Only when the deliberative mechanism has created and selected an appropriate plan or partial plan (a process which may be more or less complex, depending on the organism and the environment), will it be possible to begin producing appropriate behaviour under the control of the plan.

Research in AI over the last 40 years shows that such planning may be a far more complex process than most people might suspect if they have not tried to design working systems with such capabilities. Things get even more complex if the planning mechanism is interruptable because it operates concurrently with perceptual mechanisms, goal generators (discussed below), and plan execution mechanisms (which may be busy executing a plan to achieve another goal).

Among the important mechanisms required for such deliberative processes to work are the following:

(a) A re-usable short term memory in which candidate plan steps and predictions of plan execution can be constructed and compared, prior to the selection of a plan, which can then be 'frozen' in a longer term, though still not necessarily permanent, memory until the plan has been achieved or abandoned. If there are mechanisms for learning by comparing the results of different plan execution processes, a more sophisticated plan memory may be required.

(b) A long term extendable associative memory which can store re-usable information in order to answer questions of forms like: 'What actions are possible in situations of type S?' 'What type of situation can result from performing action A in situation S? 'What are the good and bad aspects of being in a situation of type S?'.

As suggested in (Sloman, 2000a) this might have evolved from a reactive condition-action system by a typical biological process of copying and modifying a previously evolved mechanism. The same basic content addressable associative engine for determining what action to produce in particular conditions could be copied (as is common in evolution) and the new one modified so that it is used to determine responses to inputs representing *hypothetical* situations. However, in order to be able to produce a *set* of possible actions, rather than an individual action, the mechanism would need to be able to cope with non-unique associations. This would also be required for a predictive mechanism in conditions of uncertainty, where a given event could be followed by alternative consequences. The precise sequence of evolutionary steps required to produce such functionality is a topic for further research.

Although details are not known it seems clear that some sort of trainable associative neural net could implement all the required functionality for such a memory. On this basis, along with further mechanisms of types explored in AI research on planning and problem solving, it would be possible for a planning system to be able to explore branching sets of alternative possible futures. If the associative memory also included information about alternative conditions which could produce a particular consequence, it could also be used for a backward-chaining planner.

If such a deliberative mechanism when presented with a goal is capable of discovering two or more alternative (partial or complete) plans for achieving that goal, then some additional *evaluative* criteria will be required for selecting among the alternative plans.

In some cases this could be done on the basis of how the plan relates to other current goals, e.g. by preferring a plan which most promotes or least hinders other goals or by using standards, values, preferences, or ideals (compare chapter 6 of Sloman(1978)). Notice that not all selections are necessarily based on *selfish* considerations. In addition the organism may have general preferences, e.g. for shorter plans rather than longer ones, or for less dangerous rather than more dangerous routes, or for easier terrain over shorter routes, or for familiar strategies over unfamiliar ones, at least in contexts where outcomes are partially uncertain.

As I have explained in previous papers, the sort of deliberative mechanism which can produce new plans is likely to be discrete, serial and slow, i.e. resource-limited. Although parallel neural mechanisms may be used as an underlying implementation, the main operations are likely to be serial. One reason is the need to re-use a limited short term memory for building temporary structures such as possible plans. Access to the long term associative memory will be inherently serial if it can deal with only one query at a time. It may also be necessary to have a single high level 'control' mechanism for resolving major conflicts, since otherwise multiple inconsistent decisions could be taken (e.g. going east to get food and west to get water).

Finally if the system is to be able to learn associations between events in the deliberative mechanism, the number of co-occurring events needs to be limited, for the complexity of searching for associations between subsets of events is exponential in the number of events. So learning requires parallelism to be constrained. (I first heard this argument from Dana Ballard.)

9.4 Skill compilers

For all those reasons, creating new plans is likely to be a slow process. However, once a plan has been created it may be possible for that plan to be stored and re-used later without having to repeat the planning process. That is relatively easy for a computer to do. In humans it seems that plans that are created and followed a number of times, and are usually found to be successful, can have the appropriate training effect on the reactive system's memory (the cerebellum perhaps?).

I.e. there are mechanisms whereby structures learnt in a slow deliberative processes can be transferred to the reactive system so that in future they are available for rapid retrieval and execution, without using the deliberative mechanism, which is then able to perform some other task at the same time. In other words the architecture includes one or more *skill compilers*.

This leaves open many questions about the precise details by which such stored reactive types of expertise are created and selected when needed in future. For instance is the same formalism used by the slow deliberative mechanism or does the developing of fluency and speed require transformation to a different formalism, as has often been suggested by analogy with the difference between interpreted and compiled programs?

9.5 Varieties of motivational sub-mechanisms

The discussion so far has assumed that changes in body state (e.g. need for food or liquid), or in perceived situations, can produce new goals, which may either trigger reactive behaviours if suitable ones exist, or can trigger deliberative processes to produce one or more plans to achieve the goal.

But that assumes that there is some mechanism for producing the goal in the first place. I.e. the sort of organism we are discussing requires *motive generators*. Several issues regarding the functionality of motive generators were discussed in (Beaudoin, 1994). There are many different kinds of goals, or motives. They can be short term, long term, or even permanent. They can be triggered by physiological changes, by percepts, by deliberative processes (e.g. subgoals of a pre-existing goal), by recollections, by meta-management processes.

Thus, embedded in various parts of the organism there will be a collection of motive generators: $MG1, MG2, MG3$, ... of various types. These can be activated asynchronously in parallel with other processes. E.g. while you are in the middle of building a shelter you may become tired and wish to rest, or hungry, or itchy, or curious about the the noise coming from something out of sight.

As a result of asynchronous generation, motives may be in conflict, so motive comparators are also needed, for resolving conflicts. (Sloman, 1987a). Simple models often use a single numerical motive comparator, based on the assumption that every motive has an associated utility, where utilities are either scalar values or are at least partially ordered. However, in real life things may be more complex, and the organism may have to learn which motives to treat as more important in various contexts, e.g. by discovering the short term and long term consequences of failure to act on them. Such selection skills may be expressed in a collection of different specific motive comparators, $MC1$, $MC2$, ..., some within the reactive layer, some within problem solving or planning mechanisms (e.g. comparing subgoals) and some as part of meta-management, e.g. determining which 'top level' goals to select when there are conflicts.

Some organisms may have fixed, innately determined, motive generators and motive comparators. Others may be capable of acquiring new ones through processes of learning, development, or being influenced by a culture. In other words there may be motive generator generators (MGG) and motive comparator generators(MCG).

We can speculate about even more sophisticated organisms (or robots) which may also have motive generator comparators (MGC), and motive comparator comparators (MCC), which could play a role in creating or learning motive generators or motive comparators, respectively. Perhaps we should consider yet more possibilities: $MGGG$, $MGGC$, $MCGG$, $MCGC$, $MGCG$, $MGCC$, ... etc ?

9.6 Evaluation mechanisms

There are also evaluators. Various mechanisms within an organism may evaluate some aspect of the current state as good, or bad, i.e. something to be preserved or something to be terminated. This can be independent of a comparison with any goal or motive. Aesthetic experiences involve evaluation of the current state as worth while, even if it serves no purpose, e.g. looking at a rainbow or enjoying a tune. Evaluations play a role in many theories of emotions, for instance as 'concerns' in Frijda's theory.[7]

Similarly, possible futures, including actions or action sequences, can be evaluated as desirable or undesirable independently of any comparisons with specific alternatives. This may be part of the process of deciding whether a particular goal should be adopted or rejected. If the only available means are evaluated as bad then the goal may be rejected, or perhaps postponed in the hope that new means may become available. Humans can also use their 'what-if' reasoning capabilities to consider possible past events which did not actually occur, and evaluate them as good or bad. For instance, after some failure has occurred, this could be part of trying to understand what went wrong and what alternatives were possible.

What it *means* to say that a mechanism evaluates a state or event as good or bad will be slightly different in different contexts. This cannot be defined simply in terms of the form of output of some classification mechanism, whether it is binary, symbolic, or a value on a continuous scale. Rather, what makes it an *evaluation* is the functional role of the classification within the larger architecture.

For example, to say that the *current* situation is evaluated as good by an organism, or a sub-mechanism, implies that, in the absence of other factors, the organism or mechanism will be disposed to try to preserve (or intensify) the situation, or resist occurrences which are likely to reduce or terminate it. Evaluation as bad is similar, with appropriate changes. This is different from evaluating as good or bad a *future* action or situation, or one that is merely considered as a *possibility* during planning: in these cases different effects will follow from the evaluation.

[7] More precisely, concerns, for Frijda, are 'dormant demons', i.e. dispositions which can manifest themselves through evaluations when appropriate. They are close to the attitudes discussed below.

Evaluation is also related to learning processes. For instance, learning mechanisms based on positive and negative reinforcement tend to increase the likelihood of actions that are found to produce results evaluated as good and to decrease the likelihood of actions if results are evaluated as bad. (This simple summary does not do justice to the variety of such learning phenomena.) Rolls (1998, and to-appear) has emphasized the importance of such evaluations and, like many other researchers, assumes that the evaluations make use of a 'common currency', so that conflicts between motives can be resolved by comparing evaluations. However, having something like a simple numerical scale for goodness and badness of states or events is often too crude, as consumer reports show when they present the results of comparisons in terms of a collection of descriptions of objects on different dimensions, such as cost, ease of use, reliability, results achieved, etc. (Compare the analysis of 'better' in (Sloman, 1969)). So although simple organisms, and simple sub-mechanisms in more sophisticated organisms, use a one-dimensional ordered scale of value, in general evaluation is a more complex process, and different evaluations need not be commensurable.

This is one reason why we have conjectured that humans use learnt motive comparators: there would be little need for them if all comparisons could be based on relative value or utility. If there is no fixed, general, common currency, then different rules for comparing motives or their outcomes of actions may have to be learnt separately. In some cases they may be learnt through cultural processes (including forms of indoctrination) rather than through individual learning. In other, sometimes tragic, cases conflicts remain unresolved, e.g. where there is a choice between looking after one's sick mother and leaving her to fight for one's country. The existence of a problem does not imply the existence of an answer.

I have deliberately not described evaluations as if they were always produced by the *whole* organism, because different sub-mechanisms, which evolved at different times, and which serve different functions within the larger architecture, may perform their own evaluations concurrently and independently. This allows the possibility of conflicting evaluations. Normally this will not arise, but it can and does occur, for instance when curiosity and fear pull in different directions, or when pain and sexual pleasure have the same cause.

Talk of evaluations does not imply anything about consciousness. However if there is a meta-management layer in the architecture, then since it can monitor some internal states and processes it may detect some of the evaluations produced by other subsystems. In that case the organism is aware of its evaluations.

9.7 Pleasure and pain

Evaluation is closely related to the familiar concepts of 'pleasure' and 'pain', concepts which are extremely difficult to analyse (compare chapter 11 of (Dennett, 1978)). However the concepts seem to include the use of something like what I have called evaluators. A state of pleasure is one in which the current situation or activity is evaluated positively, producing a disposition to try to preserve it. Likewise pain is connected with negative evaluation, producing a disposition to terminate or reduce something. The effects of such evaluations are only *dispositions* to act, since in many contexts the effects will be overridden by other factors: a person's behaviour may irritate you producing the disposition to try to change him, or to avoid his presence. But there may be other factors which override the disposition, so that the negative evaluation prevents the action, even if the disposition to act is quite strong. Self-control can be valuable.

Evaluations can occur at different levels in the system, and in different subsystems, accounting for many different kinds of pleasures and pains. However, we would not be inclined to refer to them as pleasures or pains unless they occur at a fairly high level in a control hierarchy and potentially affect relatively global behaviour rather than simply the reactions of a minor component embedded deep within the system performing some homeostatic function. However, there is still work to be done clarifying the relationship between evaluations and pleasures and pains.

Pleasure and pain are often thought of as emotions (e.g. in Rolls(forthcoming)). However, one can be enjoying a meal, a view, or a conversation without being at all emotional about it. Likewise it is possible to give oneself a mild pain or displeasure, e.g. by pressing a sharp point against one's skin, without being at all emotional about it. Despite all this, some people insist that these are cases of 'low intensity' emotions. This is one of many ways in which the pre-theoretical notion of 'emotion' is seriously indeterminate. The account we offer below restricts the term 'emotion' to cases where evaluations produce a (fairly strong) disposition to trigger some sort of interruption or redirection of ongoing activities, e.g. through activation of a global alarm mechanism.

Similarly not all evaluations involve emotions. For instance, within a perceptual mechanism detection of positive and negative affordances may involve evaluation of some object or situation as good or bad in relation to current goals or general needs. But this need not involve any kind of emotion, though in special cases such a percept may trigger activity in the alarm mechanism causing some disruption or re-direction, or a disposition to produce disruption or redirection, i.e. an emotion.

However such dispositions may vary in strength, and perhaps only those above a (vague) threshold are normally classed as emotions, if the disposition is overridden, as weak ones often are. In this respect, the normal concept of 'emotion' has inherently vague boundaries.

9.8 Moods and attitudes

Moods and *attitudes* are also affective concepts that can be defined architecturally. They are related to, but distinct from, emotions (though sometimes confused with them). They have in common that they tend to be longer lasting than goals, desires or (most) emotions, though in other respects moods and attitudes are very different from each other.

Moods are relatively straightforwardly construed as semantics-free global states modulated by either perceived environment or internal processes, often involving chemical changes in humans. Moods in turn have a general modulating effect on other states and processes.

Note, however, that when someone says 'I am in the mood for dancing' this expresses a specific desire rather than the sort of mood discussed above, which might be a general state of elation, calm satisfaction, depression, irritability, optimism, etc. I don't know if this is simply a quirk of English, or whether the word for 'mood' in other languages has similar divergent usage. In what follows I'll ignore such cases.

Some moods (i.e. global states) may be adaptive or 'rational' reactions to the environment. E.g. in an environment where most attempts at partly risky actions fail, this can induce a cautious or pessimistic mood in which there is a strong disposition to select the less risky of available alternatives (e.g. lying low, or doing nothing), even where there are options that could achieve far better results.

Similarly in a more 'friendly' environment where moderately risky actions are found generally to be successful this can induce an optimistic mood, including a strong tendency to select the action with the most highly valued consequences, even when there is a risk of failure.

Of course, in calling them rational reactions I do not imply that there is any conscious or deliberate adoption of the mood. Rather, it is implied that it might be rational for a designer to build in a mechanism for producing such moods. Similarly evolution might select such mechanisms because they help organisms to tailor their behaviour appropriately to the environment, even if individual organisms have no idea that that is happening. However, moods, like desires, emotions, attitudes, preferences, may sometimes be dysfunctional, e.g. when they are results of addictions or brain damage of some sort.

Further mechanisms for producing and changing moods will not be discussed here, though they are very important. There are probably several of them, e.g. some chemical, some symbolic. It is very likely that what I have called 'moods' (enduring, global, general state modulations without any specific semantic content) need to be further subdivided according to their different architectural bases: another topic for future research.

Attitudes, like moods, are generally long term states, though they are far more complex and varied than moods and they have semantic content. It is possible simultaneously to have attitudes to many individual people, to one's family, one's country, one's job, political parties, particular styles of music, particular life-styles, etc.

Moreover an attitude to one thing may have many components, including a collection of beliefs, expectations, motive generators, preferences and evaluations and will generally be rich in semantic content (unlike moods, which need not be 'about' anything). The fact that one can simultaneously have many attitudes, including loving some people and hating others, implies that the majority of them will typically be *dormant* at any particular time, like Frijda's 'concerns'. As a result we may have far more attitudes than we are aware of. Usually terms like 'love' and 'hate' refer to such attitudes even though they are often thought of as examples of emotions. Your love for members of your family, or the music of Bach, or your hatred of religious bigotry, endures at times when you are thinking about other things and you are not at all emotional.

However a new percept, thought, memory or inference can interact with a dormant attitude and trigger a new mood, emotion, or motive produced by an associated motive generator. For instance, hearing that a beloved relative is seriously ill, can trigger a state of great anxiety, and seeing a newspaper headline announcing electoral success of an individual for whom one has a very strong attitude of disapproval can trigger an emotion of anger and motivations to expose the person. In this sense, each attitude will involve a large collection of unrealised dispositions. Not all of these need be behavioural dispositions. For instance there are also dispositions to produce new mental states (as noted in (Ryle, 1949)). There is still much work to do, clarifying architectural requirements for mechanisms involved in production, 'storage', modification (including decay) and activation of attitudes.

9.9 Attention filters

For reasons discussed in (Simon, 1967) and our previous papers (including (Sloman and Croucher, 1981; Beaudoin, 1994; Wright et al., 1996)) it may be that newly generated motives, alarm signals, or other pieces of information transferred from the reactive layer to the deliberative layer may be disruptive if an already active goal or plan or activity is (1) urgent and (2) important and (3) requires close and continuous attention.

It was argued above that the deliberative system is likely to be resource-limited. Consequently, when new goals are generated, or new important items of perceptual information are presented to the deliberative mechanism, dealing with the new goal or information can cause diversion of resources from urgent, important and demanding tasks. In extreme cases, even thinking about whether to continue thinking about the new information or new goal may disrupt a really intricate process, e.g. listening to complex instructions for performing a very important task, or watching a child that may be about to do something dangerous.

Such disruptions and wasteful frequent redirections of attention could be reduced if the deliberative mechanism had some sort of attention filter with a dynamically variable threshold for determining which new goals, alarm signals or other potential interrupts should be allowed to divert processing in the deliberative layer. The interrupt threshold could be high when current tasks are important, urgent and demanding, and relatively low at other times. Since extremely important and urgent new goals could in principle be generated at any time (e.g. by the news that the building is on fire), it should never be possible for the filter totally to exclude *every* kind of interrupt.

If the mechanisms which generate the potential interrupters also assign to them a quickly computed crude, heuristic measure of urgency and importance, which our previous papers labelled 'insistence', then whether a new motive diverts the deliberative process will depend on whether its insistence is above the current threshold. The suppression of pain produced by injury while a battle is still raging or an important football match is still in progress may be an example of the functioning of such a filter.[8] It is also useful to contrast *insistence* of a motive, defined in relation to power to penetrate of an interrupt filter, with *intensity* defined as ability to remain active once selected for action. Boosting intensity of motives currently being acted on can reduce wasteful cycling between different plans and goals.

Since the process of deciding whether or not to allow something to disturb resource-limited higher level processes must itself not divert those resources, the interrupt filter will have to be relatively fast and stupid, like the global alarm system, and both are for that reason potentially error prone.

A global interrupt filter for resource-limited deliberative mechanisms is a sort of converse of the global alarm mechanism. Whether the filter is actually implemented as a separate mechanism or whether it is merely part of the operation of the alarm system or motive generators is not clear. Perhaps several implementations are used, for filtering different types of interrupts,as Beaudoin proposed.

9.10 Switching personae

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', mechanism 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 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.

[8] There is considerable further discussion of various types of filters in Beaudoin's PhD thesis.

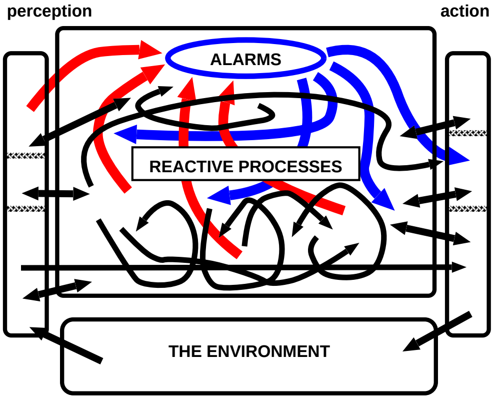

Figure 9: A reactive system with alarm mechanism
How to design an insect with emotions?

10 Organisms with subsets of the architecture

Not all parts of the grid are necessarily present in all animals. The type of architecture described so far and conjectured as an explanation of much human functioning is relatively sophisticated. Organisms which evolved earlier, including most types of organisms still extant, do not have all this complexity. It seems that various subsets of the architecture are found in nature, which is consistent with the conjecture that the complete system may have evolved from simpler architectures by adding new layers, with increasing internal functional differentiation. Studying such architectures, i.e. exploring the neighbourhood of humans in 'design space', may give us a deeper understanding of the trade-offs in our design.

10.1 Purely reactive organisms

As far as I know insects, spiders and many other evolutionarily old types of organisms are purely reactive, though they may include some very complex genetically determined reactive plans. In the case of web-construction by spiders the plans control behaviour of individual organisms. In social insects, such as termites and bees, there are very complex achievements which result from parallel activation of reactive plans in different individuals, e.g. the construction of termite 'cathedrals'. It is as if evolution was presented with an enormous variety of different problems (different evolutionary niches) and consequently found an enormous variety of different solutions.

So far, the discussion of reactive mechanisms has been very sketchy. The main feature is negative: reactive mechanisms do *not* include 'what-if' reasoning mechanisms that can construct representations of possible futures, or, more generally, possible situations past, present or future, can categorise them, evaluate them, etc. However, as explained previously, reactive mechanisms can include implicit or explicit goals.

The following are the sorts of features that characterise a reactive agent, or the reactive sub-system of a hybrid agent:

- Many processes may be analog (continuous), while others use discrete condition/action mechanisms
- Mechanisms and space are dedicated to specific tasks, so that parallelism can achieve great speed, and different processes can proceed without mutual interference.
- All internal information-bearing states use dedicated memory locations, registers, or neural components: there are no general purpose re-usable information stores as are required for deliberative mechanisms.
- Both conditions and actions may be purely internal, in sophisticated reactive systems.
- Reinforcement learning is possible, though this usually requires alterations to weights in a pre-existing architecture.
- Stored plans can be implicit in learnt chains of responses.
- There is no construction of new plans or structural descriptions in advance of execution, though implicit plan learning or plan modification can occur, e.g. by learning new condition-action chains.
- There is no explicit creation of alternative new structures (e.g. descriptions of action sequences) followed by evaluation and selection.
- Conflicts between reactions triggered simultaneously may be handled by vector addition, winner-takes-all nets or conflict resolution rules triggered by the existence of conflicts – implicit reactive motive comparators.
- Most behaviours are genetically determined though there may also be some learning, e.g. modification of

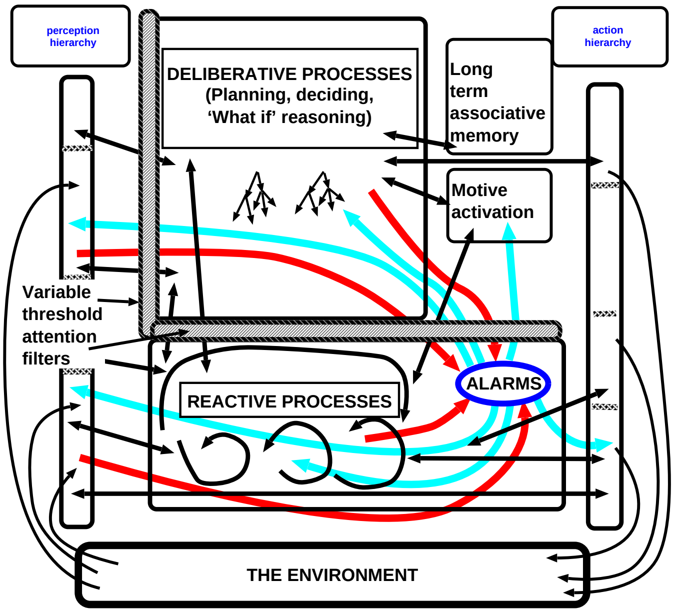
perception hierarchy
action hierarchy
DELIBERATIVE PROCESSES (Planning, deciding, 'What if' reasoning)
Long term associative memory
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Figure 10: A hybrid reactive and deliberative agent, with alarms

This type of hybrid architecture includes some of the mechanisms discussed in* section 9*, including motive generators, a content addressable long term memory, a variable threshold attention filter, and layered perception and action systems.

tunable control loops, change of weights by reinforcement learning.

- Disaster may follow if the environment requires new plan structures. (In some species this may be partly compensated for by having large numbers of expendable agents.)

There are different classes of reactive architectures. Some use several processing layers: e.g. hierarchical control loops, or subsumption architectures. Some include actions that manipulate internal state, including states representing temporary goals, as pointed out above and in (Nilsson, 1994). If reactive mechanisms (like many human forms of expertise) require rapid detection of fairly abstract features of the environment or rapid invocation of complex structured actions, then the perceptual and action 'towers' may be layered, with hierarchical concurrent processing.

If the reactive system includes a global alarm mechanism (as depicted in *Figure 9*), this allows additional features, including rapid redirection of the whole system in the light of sudden dangers or sudden opportunities. This may include behaviours such as: freezing, fighting, attacking, feeding (pouncing), fleeing, mating, and more specific trained or innate automatic behavioural responses. Such reactions seem to be closely related to what Damasio and Picard call "Primary Emotions".

Another type of response (another sort of primary emotion) which may be rapidly produced by an alarm system is general arousal and alertness (e.g. attending carefully to certain sights, sounds or smells, or increased general vigilance). If this is a purely internal change it is distinct from the type of emotion which involves changes in external behaviour or externally detectable physiological changes.

These alarm-driven reactions producing global changes can also occur in 'hybrid' organisms which have deliberative as well as reactive components.

10.2 Hybrid reactive and deliberative organisms

In addition to reactive mechanisms, many animals closer to humans in design space seem to have a deliberative capability at least in a simple form allowing brief 'look-ahead', e.g. some mammals and perhaps some birds, though it is hard to be sure exactly what is going on in such cases. In 1927 Kohler reported studies of creative problem solving in chimpanzees which seemed to demonstrate an ability to 'think ahead' and grasp the consequences of actions not yet performed – an ability which varied between individual apes.

A hybrid reactive and deliberative architecture is depicted schematically in *Figure 10*, including reactive and deliberative layers, along with a global alarm system, and some of the mechanisms mentioned previously in *section 9*.

Some of the features of a deliberative mechanism in a hybrid architecture have been discussed above. To recapitulate:

- Explicit goals or motives, possibly with considerable syntactic richness supported by a re-usable short term memory, are manipulated, and can drive the creation of plans.
- Descriptions of new possible sequences of actions are constructed and evaluated, without the actions being performed.
- A re-usable general purpose short term memory mechanism is required, which can be used for a variety of different tasks at different times, storing different information, unlike reactive systems where all state information is in dedicated components.
- Plans found useful repeatedly can be transferred (by a skill compiler) to the reactive layer where they can be invoked and executed quickly.
- Sensory mechanisms operate concurrently at different levels of abstraction, and high level (abstract) perceptual layers produce 'chunked' information that is useful for stored generalisations required for 'what if' reasoning.
- Action mechanisms may also be layered, with the higher levels accepting more abstract descriptions of actions, which are automatically decomposed into lower level actions.
- As explained earlier, the deliberative layer's actions will be largely serial and comparatively slow.
- A fast-changing environment (including bodily changes triggering events in the reactive layer) can cause too many interrupts and new goals for a speed-limited deliberative layer to be able to process them all.
- Filtering such interrupts and new goals using dynamically varying interrupt thresholds (Sloman, 1987a; Beaudoin, 1994; Wright, 1977) may reduce the problem, as explained above. The filter has to act quickly and operate without using deliberative resources or sophisticated reasoning capabilities. Consequently it is likely to make errors, though it may be trainable so that its heuristics fit the environment.

A hybrid reactive/deliberative system is likely to require a global alarm mechanism for the reasons given in *section 8*. In addition to the effects an alarm mechanism has on reactive systems, in a hybrid system it might have new features. The alarm mechanism might, for instance, be triggered by the occurrence of certain *hypothetical* future or past possibilities represented in the 'what-if' reasoning mechanism, instead of being triggered only by *actual* occurrences when they are detected by sensors and internal state monitors. In addition, in a hybrid system the alarm mechanism might be able to interrupt and redirect or modulate the deliberative layer as well as triggering changed external behaviours and modulating the reactive processes.

These extensions to the functionality of the alarm mechanism could produce new sorts of states, such as becoming apprehensive about anticipated danger and later being relieved at discovering that the anticipated unpleasant events have not occurred. The alarm system may also produce a host of more specialised learnt effects on the deliberative system, e.g. switching modes of thinking in specific ways that depend on current goals and the current environment. For instance, detection of rapidly approaching danger might trigger a new less careful and detailed mode of deliberation in order to increase the chance of finding suitable plan of evasion rapidly, even if it is not necessarily the best plan.

Where deliberative mechanisms can trigger global events through the alarm system, this seems to correspond to cases where cognitive processes trigger 'secondary emotions' as described in (Damasio, 1994). We can distinguish two types of secondary emotion produced in this way.

(a) *Purely central* secondary emotions where a certain pattern of activity in the deliberative mechanism causes the alarm mechanism to trigger certain rapid changes in the type or content of deliberative processing, but without producing any new external behaviour or externally detectable physiological changes

(b) *Partly peripheral* secondary emotions, where the new global signals change not only central but also 'peripheral', externally detectable states of the body, such as sweating, muscular tension, changes in blood flow, etc. (Described as 'sentic modulation' by (Picard, 1997)).

There may also be certain classes of secondary emotions which produce only peripheral changes without any modification of internal deliberative processes. Lie-detectors may be dependent on such mechanisms.

Notice that within the context of an architecture we can define a collection of concepts distinguishing various types of processes independently of whether those processes are or are not actually found in real organisms. In other words the architecturally defined concepts allow us to formulate empirical research questions which can then be answered by investigating organisms with the architectures. Yet more such concepts can be supported if a meta-management layer is added to the architecture.

10.3 Meta-management with alarms

We now return to another look at the complete system postulated previously, and depicted schematically in *Figure 7* and *Figure 8*. We can present this sort of architecture in a slightly enriched form, but still schematically, as in *Figure 11*. Meta-management processes allow the following processes which are not possible in the

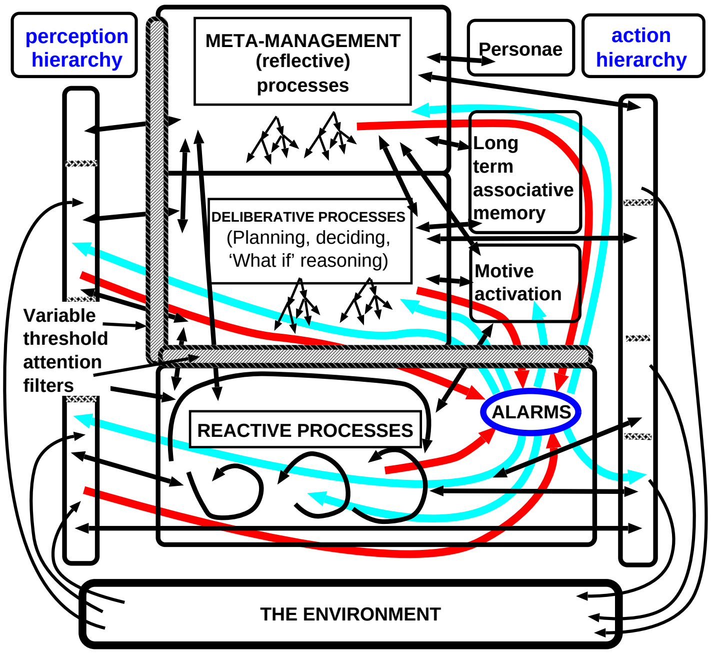
perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, 'What if' reasoning)
Motive activation
Variable threshold attention filters
REACTIVE PROCESSES
ALARMS
THE ENVIRONMENT

Figure 11: A three layered system with alarms: the H-Cogaff architecture.
The H-Cogaff architecture is a special case of the CogAff schema, proposed as a first draft specification of a human architecture. A meta-management mechanism receives information from many parts of the system. On this basis it can classify, evaluate and to some extent control various internal states and processes, including deliberative processes. However it may be distracted *or* interrupted *by information coming from the reactive mechanisms or perceptual system, or by global control signals from alarm mechanisms. The variable threshold interrupt filter mentioned previously can be used to protect both the deliberative and the meta-management processes.* NOTE: *this is an improved version of the diagram, not the one in the original paper.*

previous systems:
(1) concurrent self monitoring (of many internal processes, possibly including some intermediate databases in a layered perceptual system)
(2) categorisation and evaluation of current states and processes, including those in the deliberative layer
(3) self modification (self-control) – though this may be partial, since concurrently active reactive or alarm mechanisms may be able to disrupt meta-management processes, e.g. because new items exceed the current threshold in the attention filter mechanism mentioned previously.

Why should this layer evolve? There are several functions that it could perform. Deliberative mechanisms with evolutionarily determined strategies may be too rigid. Internal monitoring and evaluation mechanisms may help the organism improve its planning and reasoning methods by doing the following:
• detect situations when attention should be re-directed, e.g. when working on one problem produces information that acts as a reminder of some other important task,
• improve the allocation of scarce deliberative resources e.g. detecting "busy" states and varying the interrupt threshold in a more context sensitive fashion, compared with an automatic and inflexible adjunct to a resource-limited deliberative system,
• record events, problems, decisions taken by the deliberative mechanism, e.g. to feed into learning mechanisms,
• detect management patterns, such as that certain deliberative strategies work well only in certain conditions,
• allow exploration of new internal problem solving strategies, categorisations, self-description concepts, evaluation procedures, generalisations about consequences of internal processes,
• allow diagnosis of injuries, illness and other problems by describing internal symptoms to others with more experience,
• evaluate high level strategies, relative to high level long term generic objectives, or standards,
• using aspects of intermediate visual representations to communicate more effectively with others, e.g. by using viewpoint-centred appearances to help direct someone else's attention (e.g. 'look at where the hillside meets the left edge of the roof'), or using drawings to communicate how things look.

(Having access to contents of intermediate perceptual buffers and other internal states may cause some philosophically inclined robots to discover sensory qualia, and perhaps start wondering whether humans have them too!)

By doing all these things meta-management can promote various kinds of learning and development; reduce the frequency of failure in tasks; prevent one adopted goal interfering with other goals; prevent endless, time-wasting efforts on problems that turn out not to be solvable; notice that a particular planning or problem solving strategy is slow and resource-consuming, then possibly replace it with a faster or more elegant one; detect possibilities for structure sharing among actions, and finally allow more subtle cultural influences on behaviour

However, none of these functions is likely to be performed perfectly, for various reasons. For instance, self monitoring cannot give complete and error-free access to all the details of internal states and processing: it is just a form of perception, and like all forms of perception will involve abstracting, interpreting, and possibly introducing errors.

Even when self-observations are accurate, self-evaluations may be ill-judged and unproductive, e.g. because some incorrect generalisations from previous experiences cause wrong decisions to be made about what does and does not work, or because religious indoctrination causes some people to categorise normal healthy thoughts and motives as ‘sinful’.

Even when self-observations are accurate, and self-evaluations are not flawed, decisions about what to do may not be carried, out either because some of the processes are not within the control of meta-management (e.g. you cannot stop yourself blushing) or because processes that are partly under its control can also be disrupted by other processes, e.g. intrusive noise or other salient percepts or features of the situation which trigger the alarm situation to intervene when it would be better not to. The latter are characteristic of tertiary emotions (perturbances), where there is partial loss of control of attention. This is possible only in the presence of meta-management which allows some control of attention.

Contrary to the claims of some theorists (perhaps Damasio and Picard?) there could be emotions at a *purely* cognitive level – an alarm mechanism triggered by events in the deliberative mechanism interrupting and diverting processing in deliberative and meta-management systems without going through the primary emotion system.

Some people are more prone than others to react with bodily symptoms, including externally detectable symptoms, when secondary or tertiary emotions occur. We could therefore distinguish partly peripheral and purely central secondary and tertiary emotions, producing four sub-categories in addition to primary emotions.

11 Architectural layers and types of emotions

Our everyday attributions of emotions, moods, attitudes, desires, and other affective states implicitly presuppose that people are information processors. This is partly because they generally have *semantic* content, including reference to the objects of the emotions. For instance, you cannot be angry without being angry about something, and to long for something you need to know of its existence, its remoteness, and the possibility of being together again. That information plays a central role in the production and character of the emotion.

Besides these *semantic* information states, anger, longing and other emotions also involve complex *control* states. One who has deep longing for X does not merely occasionally think it would be wonderful to be with X. In deep longing thoughts are often *uncontrollably* drawn to X.

Physiological processes (outside the brain), such as changes in posture, facial expression, blood pressure, muscular tensions, or amount of sweat produced, may be involved in some cases, e.g. in primary emotions and in a subset of secondary and tertiary emotions, but not necessarily all emotions. The importance of such physiological reactions is often over-stressed by experimental psychologists under the influence of the James-Lange theory of emotions. Contrast the views of (Oatley and Jenkins, 1996), and what poets say about emotions. The fact that some theorists regard certain physiological phenomena as *defining* emotions or somehow *central* to all emotions, illustrates the comparison in *section 2* with a collection of blind men each trying to understand what an elephant is on the basis of the part he can feel. On the basis of our theory we would predict such misunderstandings.

On this theory, many sorts of control phenomena, including not only production of physical changes, but also redirection of perception, thought processes, attention, motivation and actions, are possible if the information processing architecture is rich enough. In particular, the various architectural layers discussed above, along with the alarm system and mechanisms for generating new “high-insistence” motives, can explain (at least in outline) what have been called primary, secondary, and tertiary emotions.

Primary emotions involve events triggered in reactive mechanisms which cause something like an alarm system to rapidly redirect reactive behaviours and usually also perceptual and motor subsystems. This can include such things as being startled, disgusted by horrible sights and smells, being terrified by a large fast-approaching object, and perhaps various kinds of sexual and aesthetic arousal, though they should probably be given separate categories. In simple organisms such primary emotions may provoke fighting, fleeing, freezing, etc. In more sophisticated cases they could produce *readiness* for various kinds of actions rather than those actions themselves, along with increased general, or directed, alertness.

Secondary emotions are triggered by events in the deliberative mechanism, during 'what if' reasoning processes, or by perceiving something relevant to an important deliberative activity or state, e.g. a long term goal. Such secondary emotions could occur during planning, during reflections on past events, or idle thinking or reminiscences. The results could be various kinds of anxiety, relief, fear, pleasure at unexpected success, etc. An example would be noticing during planning that an action being contemplated could go horribly wrong. This might produce a mixture of apprehension and extreme attentiveness to the task. The effects of such emotions may or may not include those typical of primary emotions.

Within different architectural frameworks we can explore different kinds of secondary emotions, their various causes and various effects. For instance, in an architecture which supports long term dormant attitudes, a new high level percept (e.g. unexpectedly seeing an animal mistreated) can interact with a dormant attitude (e.g. love of animals) to generate a host of reactions including generating new goals and mobilising resources to achieve those goals.

Secondary emotions will often include effects characteristic of tertiary emotions, namely disruption or redirection of meta-management processes, where the architecture includes meta-management.

Tertiary emotions depend on the third type of architectural layer (meta-management, reflection), which provides capabilities for self-monitoring, self-evaluation and (partial) self-control, including control of attention and thought processes. In that case there is also the possibility of some loss of control, which we previously called "perturbance", and now refer to as "tertiary emotions". This can include such states as feeling overwhelmed with shame, feeling humiliated, being infatuated or besotted, various aspects of grief, anger, excited anticipation, pride, and many more. These are typically *human* emotions, and are generally the stuff of plays, novels and gossip. Many, though not all of them involve social interactions, or the possibility of various kinds of social interactions. It is possible that some primates have simplified versions of tertiary emotions.

Most socially important human emotions require having sophisticated concepts and knowledge and rich control mechanisms embedded in sophisticated architectures. Some of these emotions, e.g. patriotic fervour, dismay at the success of a despised politician, resenting being passed over for promotion, delight at solving a famous hard mathematical problem, are too semantically rich and complex to have 'natural' behavioural expressions, though they can be expressed in language. (Many of the speeches written by Shakespeare and other great playwrights are marvellous examples of the use of language to express emotions.)

The situation is somewhat more complex than the previous remarks may appear to indicate. First of all, actual emotional states may be a mixture of all three kinds and other things, and the labels we use do not necessarily identify a simple category. In particular, the things we call "love", "hate", "jealousy", "pride", "ambition", "embarrassment", "grief", "infatuation" can involve effects found in all three categories, since in ordinary usage these words do not refer unambiguously to any of the particular categories that can be defined precisely in terms of the underlying architecture.

Further, as shown in the previous section, the sort of architecture we have been discussing could potentially explain more varieties of emotions which differ according to which parts of the of the system are affected. Moreover, since we are not discussing *static* states but *developing* processes, with very varied aetiology, more subtle distinctions could be made according to the dynamics of different sorts of processes, how they are generated, which kinds of feedback loops they trigger, how long they last, how their intensity rises and falls, which long term effects they have, etc.

More generally, within the sort of framework we have been presenting we can begin to define an *architecture-based* ontology of mind (Sloman, to-appear). Different sorts of architectures, found in different animals, and in humans at different stages of development, or with various kinds of brain damage, will support different mental ontologies.

In particular, different animals and different sorts of humans will not all be capable of the same classes of emotions or emotion-like states. For instance, as suggested above, if newborn human infants do not have adult information processing architectures, then they will not be capable of having burning ambition, religious guilt, or national pride.

From this viewpoint it is a mistake to claim that all sorts of emotions have physiological effects outside the brain in the manner suggested by William James. Some will and some will not, e.g. the purely central secondary and tertiary emotions, when the architecture supports them. For which individuals or species this is possible will be an empirical question, not a question of definition.

Likewise it is a mistake to claim, as many do, that having an emotion necessarily involves being *aware* of having that emotion. Leaving aside the ambiguity or unclarity of the concept of 'being aware' there is first of all the fact that people can often be in emotional states that are evident to others but not to them, e.g. being angry or jealous or infatuated: a point often used in plays and novels. Secondly, from an architectural viewpoint, if being aware of X involves meta-management capabilities and having self-monitoring capabilities directed at X, then organisms (or very young infants) that lack meta-management may be incapable of being aware of internal states such as primary emotions, even if they have them. By defining various sorts of awareness in terms of the architectures and mechanisms that support them we can replace one ambiguous, unanswerable, question with a collection of more precise empirical questions.

We can also use this to explain the existence of qualia, if they involve kinds of self-awareness made possible by the action of the third layer.

The theories presented here are highly speculative. But that does not make them false! If they are close to the truth then that has many implication for our understanding of many topics in cognitive science, e.g. regarding kinds of individual development and learning that are possible, the varieties of perceptual processes that can occur, and the different sorts of affordances that are processed in parallel, varieties of brain damage that might occur and their possible effects the kinds of emotions and other affective states and processes that can occur.

Finally, it should be acknowledged that our architecture-based theory of types of emotions and other affective states is not necessarily in conflict with theories which emphasise other aspects of emotions, such as their causes, semantic contents, and their effects. For instance, the theory in (Ortony et al., 1988) concentrates on aspects of emotions and attitudes that are mostly orthogonal to the architectural issues, and attempts to account for our pre-theoretic emotional labels. A complete theory would need to encompass both viewpoints, though we see no reason to expect that pre-theoretical taxonomies enshrined in colloquial language will survive unscathed.

12 What sorts of architectures are possible?

We know so little about possible information processing mechanisms and architectures (especially the extraordinarily powerful visual mechanisms implemented in animal brains) that it is premature to hope for a complete survey of types of architectures and their capabilities. It could turn out, as some have claimed, that any information-processing architecture produced by millions of years of evolution is bound to be far too messy and unstructured for us to understand as engineers, scientists or philosophers (*Figure 1*).

Alternatively, it may turn out that evolution, like human designers, must use principles of modularity and re-usability in order to achieve a robust and effective collection of architectures, such as we find in many kinds of animals. *Figure 6* and *Figure 7* and our earlier discussion present more structured and modular architectures, combining a three-fold division between perception, central processing, and action, and three levels of processing, with and without a global 'alarm' mechanism. However, such diagrams can be misleading partly because they convey very different designs to different researchers. A frequent confusion is between diagrams indicating state-transitions (flow-charts) and diagrams indicating persisting, interacting components of an architecture. In the former an arrow represents a possible change of state. In the latter it represents flow of information between components. My diagrams are of the latter kind.

To help us understand what to look for in naturally occurring architectures, it may be useful to attempt a preliminary overview of some features of architectures that have already been proposed or implemented. We can then begin to understand the trade-offs between various options and that should help us to understand the evolutionary pressures that shaped our minds.

Researchers on architectures often propose a collection of layers. The idea of hierarchic control systems is very old both in connection with analog feedback control and more recently in AI systems. There are many proposals for architectures with three or more layers, including not only ours but also those described by Albus and Nilsson mentioned previously, the subsumption architecture of Brooks (1991), the ideas in Johnson-Laird's discussion (1993) of consciousness as depending on a high level 'operating system', the multi-level architecture proposed for story understanding in (Okada and Endo, 1992), Minsky's notion of A, B and C brains in section 6.4 of (Minsky, 1987) and many others.

On closer inspection, the layering in multi-level architectures means different things to different researchers, and in particular different researchers refer to a so-called 'three-layer architecture' but propose very different distinctions between the layers.

There seem to be several orthogonal distinctions at work, which, at present, I can classify only in a very crude fashion. The following should be read as a first categorisation based on (Sloman, 2000b), which is likely to be revised in the near future.

12.1 Concurrently active vs pipelined layers

In Albus (1981) and some of what Nilsson (1998) writes, the layers have a sequential processing function: sensory information comes in (e.g. on the 'left') via sensors to the bottom layer, gets abstracted and interpreted as it goes up through higher layers, then near the top some decision is taken on the basis of the last stage of abstraction or interpretation, and then control information flows down through the layers and out to the motors (on the other side). I call this an "Omega" architecture because the pattern of information flow is shaped like an Ω.

Many AI models have this style. This can include hybrid architectures, e.g. where the lower levels are competing neural nets whose activity is triggered by incoming sensory information and the higher levels are symbolic processes which help to select one of the competing sub-nets.

An alternative is an architecture where the different layers are all concurrently active, with various kinds of control and other information constantly flowing within and between them in both directions, as in figure *Figure 6* and the 'Cogaff' architecture in *Figure 11*.

12.2 Dominance hierarchies vs functional differentiation

A second distinction concerns whether higher levels *dominate* lower levels or merely attempt to control them, not always successfully and sometimes with the direction of control reversed. In the subsumption model (Brooks 1991) higher levels not only deal with more abstract state specifications, goals and strategies, but also completely dominate lower levels. I.e. they can turn lower level behaviour off, speed it up, slow it down, modulate it in other ways, etc. This conforms to the standard idea of hierarchical control in engineering.

By contrast, in a non-subsumptive layered architecture (such as the Cogaff architecture) the 'higher' levels manipulate more sophisticated and abstract information, but do not necessarily dominate the lower levels, although they may sometimes attempt to do so. Higher levels may be able partially to control the lower levels but sometimes they lose control, either via alarm mechanisms or because other influences divert attention. For instance, attention can be diverted by sensory input with high salience (loud noises, bright flashes) and by newly generated motives with high 'insistence' (e.g. hunger, sitting on a hard chair, etc.).

In the Cogaff model the *majority* of lower level reactive mechanisms cannot be directly controlled by the deliberative and metamanagement layers, especially those concerned with controlling bodily functions. Some training may be possible, however, a possibility allowed in the next dimension of variation.

12.3 Direct control vs trainability

In some layered systems it is assumed that higher levels can directly control lower levels. A separate form of control which is not 'immediate' is re-training. It is clear that in humans higher levels can sometimes retrain lower levels even when they cannot directly control them.

For instance, repeated performance of certain sequences of actions carefully controlled by the deliberative layer can cause a reactive layer to develop new chained condition-action behaviour sequences, which can later run without higher level supervision. Fluent readers, skilled athletes, musical sight-readers, all make use of this. (The nature of the interface between central mechanisms and action control mechanisms, discussed in *section 12.8* below, is relevant here.)

12.4 Different kinds of processing vs different control functions

On some models, different layers all use the same kinds of processing mechanisms (e.g. reactive behaviours) but perform different functions, e.g. because they operate at different levels of abstraction. In other models there are different kinds of processing as well as different functional roles.

For instance, our figures showing layered architectures include a lowest level that is purely reactive, whereas the second and third levels can do deliberative, 'what if', reasoning, using mechanisms able to represent possible future actions and consequences of actions, categorise them, evaluate them, and make selections. This is not how reactive systems behave.

Traditional AI planning systems can do this, and similar mechanisms may be relevant to explaining past events, doing mathematical reasoning, or doing general reasoning about counterfactual conditionals.

However, it is likely (indeed inevitable) that the deliberative mechanisms which go beyond reactive mechanisms in explicitly representing alternative actions prior to selection are themselves *implemented* in reactive mechanisms, e.g. reactive mechanisms that operate on structures in a temporary workspace.

Reactive mechanisms may be implemented in various kinds of lower level mechanisms, including chemical, neural and symbolic information-processing engines, and it is possible that the reliance on these is different at different levels in the architecture. Some kinds of high level global control may use chemical mechanisms which would be too slow and unstructured for intricate problem solving.

Some have argued that human capabilities require quantum mechanisms though I have never seen a convincing account of how they could explain any detailed mental phenomena.

12.5 Where are springs of action?

A fifth distinction concerns whether new 'intrinsic' motives (which are not sub-goals generated in a planning process) all come from a single layer or whether they can originate in any layer. In one variant of the Omega model, information flows up the layers and triggers motivational mechanisms only at the top. In other models, processes anywhere in the system may include motive generators.

For instance, in the Cogaff architecture, physiological monitors in the reactive layer can generate new goals, e.g. to acquire food, to adjust posture, to get warmer. Some of the motives thus generated may be handled entirely by reactive goal-directed behaviours, while others have to be transferred to the deliberative layer for evaluation, adoption or rejection, and possibly planning if there is no previously stored strategy for achieving such goals in this sort of context. A full survey of theories of motive generation is beyond the scope of this paper, but it is perhaps worth noting that not all motives benefit the individual when satisfied. In particular, it would be useful to attempt to explain various kinds of addictions (to drugs, eating, sex, power over others, gambling, computer games, etc.) in terms of sources of motivation in an architecture.

12.6 Handling competing motives

Not all motives will be mutually consistent, so there has to be some way of dealing with conflicts. Architectures differ regarding the locus of such conflict resolution and the mechanisms deployed.

For instance, in some forms of contention-scheduling models, schemata form coalitions and oppositions on the basis of fixed excitatory and inhibitory links in a network, and then some kind of numerical summation leads to selection, which is always done at the same level in the hierarchy. In other models the detection of conflicts might use symbolic reasoning, and the resolution might be done at different levels for different sorts of conflicts.

For instance the decision whether to stay home and help granny or go to the marvellous concert might be handled in one part of the system, the decision whether to continue uttering the current unfinished sentence or to stop and take a breath in another part, and the decision to use placatory or abusive vocabulary when addressing some who has angered you might be handled by yet another part of the system. In the last example, two parts might compete for control: a reactive part generating an impulse to be abusive and a deliberative or meta-management mechanism deciding that only placatory words will do any good in the long run.

12.7 Perceptual to central connections

Architectures with perceptual components differ in the relationships they propose between modes of processing in perceptual modules and more central layers. E.g. is the perceptual processing itself layered, producing different levels of perceptual information to feed into different central layers, or is there a fixed entry level into the central mechanisms, after which the information may or may not be passed up a hierarchy, as in the Omega model, and in Fodor's model depicted in *Figure 2*?

The Omega model could be described as using a 'peephole' model of perception: sensory information comes in via a limited orifice and then has to be processed and interpreted centrally. The Fodor model (in the version proposed by Marr (1982)) could be described as the 'telescope' model, in which information arrives through a narrow orifice after several layers of specialised processing. Both can be contrasted with the 'multi-window' model of perception in the Cogaff model presented above.

In 'peephole' or 'telescope' perceptual systems, the sensory mechanisms (simple transducers or more complex sensory analysers) produce information about the environment and direct it all to some component of the central architecture. That may trigger processes which affect other parts.

In *Figure 6* and subsequent figures, it is suggested that the perceptual processes are themselves layered, handling different levels of abstraction concurrently, with a mixture of top-down and bottom up processing, and with different routes into different parts of the central system. For instance, deliberative mechanisms may need perceptual information chunked at a fairly high level of abstraction, whereas fine control of movement may require precise and continuously varying input into the reactive system. Differential effects of different kinds of brain damage seem to support the multi-window multi-pathway model, which can also be defended on engineering grounds. (The multi-window model was defended at greater length in (Sloman, 1989), though not with that label.)

12.8 Central to motor connections

An analogous distinction concerns the relationship between central and motor processing. Just as there is what I called 'multi-window' perception, 'telescope' perception, and 'peephole' perception, so too with action. At one extreme there is only a 'narrow' channel linking the motor system only with the lowest level central mechanism, as in the Omega model: there are motors and they all get signals directly from one part of the central mechanism (analogous to 'peephole' perception). At another extreme there can be a layered, hierarchical motor control system where control information of different sorts comes in directly at different levels, from different layers in the central system. In between, is the 'reverse telescope' model, where only fairly high level abstract instructions (e.g. something like "pick up the hammer") are handed to the motor system which then produces one or more transformations down to detailed motor instructions or muscle control signals. The Fodor model in *Figure 2* seems to cover both peephole and reverse telescope models of action.

Humans seem to have motor systems with complex hierarchical skills, and probably also many other animals. An example supporting the multi-window model is our ability to perform some well rehearsed skill based on reactive behaviours while modulating it under the control of high level preferences and goals, for instance while playing a musical instrument, or making a speech, or acting on the stage in such a way as to give the impression of being furtive, or confident, or caring, etc.

In some proposed architectures (e.g. (Albus, 1981)) this hierarchical organisation of action is acknowledged, but instead of the action hierarchy being a separate 'tower' with its own layers communicating with several central processing layers, it is folded in to the central control hierarchy, as if the reverse telescope is part of the central cognitive mechanism. This might give the hierarchical action mechanism more access to powerful central reasoning mechanisms and information stores, but would reduce the opportunities for central processes (e.g. planning, problem solving) to continue uninterrupted while complex actions are being carried out.

The different models could be construed as describing similar systems viewed differently. However, I believe there are significant engineering design differences, with complex trade-offs that have not yet been investigated fully. In particular, designing the central system and the action systems as both having hierarchic organisation and both able to operate concurrently has advantage for an organism or robot that needs to be able to think a long way ahead while carrying out complex actions. Moreover, allowing different parts of the action hierarchy to receive instructions directly from different parts of the central system allows control information to be channelled directly to where it is needed without all having to go through a selection bottleneck. Conflicts can be detected within the action 'tower' and may either resolved there, if they are routine conflicts, or may trigger some sort of interrupt in the central mechanisms.

I conjecture that thinking about such design trade-offs may help us understand how the whole system evolved in humans and other animals as well as helping us come up with better engineering designs. Similar comments are applicable to the trade-offs involved in different architectures for perception.

12.9 Emergence vs 'boxes'

One of the notable features of recent AI literature is the proliferation of architecture diagrams in which there is a special box labelled 'emotions'. Contrast our figures, where there is no specific component whose function is to produce emotions. Instead we explain several varieties of emotions as emergent properties of interactions between components which are there for other reasons, such as alarm mechanisms and mechanisms for diverting attention (which often happens without any emotion being generated).

This is often compared with the emergence of 'thrashing' in a multi-processing architecture. The thrashing is a result of interactions between mechanisms for paging, swapping and allocating resources fairly when there is a heavy load. There is no special 'thrashing' module. As with emotions, thrashing may or may not be detected by the system within which it occurs: this depends on the availability of sophisticated self-monitoring processes.[9]

[9] Emergence in this sense does not involve any mysterious processes. There are two sorts of emergence that are usefully distinguished. In one sense, "agglomerative emergence", emergent states, processes and properties are merely large-scale features of a system, but can be

Disagreements over whether to label components of an architecture using the word "emotion" may be partly terminological: e.g. some theorists write as if all motives are emotions. Then a component that can generate motives may be described as an 'emotion generator' by one person and as a 'motive generator' by another. Separating them accords better with ordinary usage, since it is possible to have motives and desires without being at all emotional, e.g. when hungry, although *intense* desires can have the properties characteristic of disruptive emotional states.

This is just one of many areas where we need far greater conceptual clarity, which may come in part from further study of varieties of architectures, their properties, and the states and processes they support.

There are many cases where it is not clear whether some capability needs to be a component of the architecture, or an emergent feature of interactions between components. The attention filters in *Figure 10* and *Figure 11* are examples. Instead of using a special filtering mechanism, a design can produce the effects of filtering through interactions between competing components. The first alternative may be easier to implement and control. The second may be more flexible and general. There are many such design trade-offs still to be analysed.

12.10 Dependence on language

Some models postulate a close link between high level internal processes and an external language. For instance, it is often suggested (Rolls 1998) that mechanisms analogous to meta-management could not exist without a public language used by social organisms, and in some of Dennett's writings consciousness is explained as a kind of 'talking to oneself'.

A contrary view is that internal mechanisms and formalisms for deliberation and high level self-evaluation and control were necessary pre-cursors to the development of human language as we know it.

The truth is probably somewhere in between, with an interplay between the development of internal facilitating information processing mechanisms and social processes which then influence and enhance those mechanisms, for instance by allowing a culture to affect the development in individuals of categories for internal processes of self-evaluation. (Freud's 'super-ego'). However, it appears from the capabilities of many animals without what we call language, that very rich and complex information processing mechanisms evolved long before external human-like languages, and probably still underpin them.

Since the acquisition, storage, manipulation, retrieval and use of information in such animals has important high level features in common with uses of external languages (including the need for structural variability, extendability, and manipulability of the internal information medium) we can usefully extend the word 'language' to refer to forms of internal representation and say that the use of language to think with (see with, desire with, intend with, plan with, etc.) is biologically prior to its use in external communication. This will almost certainly give us a better understanding of the phenomena normally referred to as language because of the way they depend on older inner languages. When we understand better how they work, we shall be in a better position to understand how social linguistic phenomena influence individual mental phenomena.

12.11 Purely internal vs partly external implementation

A more subtle distinction concerns how far the implementation of an organism or intelligent artefact depends entirely on the internal mechanisms and how far the implementation is shared with the environment. The development in the 70's of 'compliant wrists' for robots, which made it far easier, for example, to program the ability to push a cylinder into a tightly fitting hole, illustrated the advantage in some cases of off-loading information processing into mechanical interactions. Trail-blazing and the design of ergonomically effective tools and furniture are other examples.

From a philosophical viewpoint a more interesting case is the ability to refer to a spatially located individual unambiguously. As explained long ago by Strawson (1959), whatever is *within* an individual cannot *suffice* to determine that some internal representation or thought refers to the Eiffel tower, as opposed to an exactly similar object on a 'twin earth'. Instead the referential capability depends in part on the agent's causal and

defined in terms of the small scale components that make them up, and their causal interactions can be derived logically or mathematically from the laws of behaviour of the low level features. A more interesting type of emergence involves processes in virtual machines whose ontologies are not *definable* in terms of the ontology of the underlying implementation machine, from which it follows that the laws of behaviour of the entities in that ontology cannot be logically or mathematically derived from the laws of behaviour of the implementation machine. An example would be a chess playing virtual machine in a computer, where concepts like "king", "capture", "win" cannot be defined in terms of the concepts of physics and (less obviously) the rules of chess and the playing strategies cannot be derived from the laws of physics and descriptions of the underlying physical machine. In these cases there is circular causation, between the physical layer and the virtual machine layer, even though the physical layer may be causally complete. A more detailed, but still incomplete, discussion of this point can be found in http://www.cs.bham.ac.uk/ axs/misc/supervenience spatial relationships to the thing referred to. So attempting to implement *all* aspects of mental functioning entirely within a brain or robot is futile: there is always a subtle residue that depends on external relations, and an adequate theory of mind has to take account of that.

In referring to parts of oneself, or parts of one's own virtual machine the problem of unique reference is solved partly by internal causal relationships as explained in (Sloman, 1985; Sloman, 1987b).

Allowing that mental states with semantic content are not implemented solely in internal mechanisms, distinguishes the theory developed here from common varieties of functionalism, which reduce mental phenomena to purely internal functional or computational phenomena. Our position is also distinct from methodological solipsism and other varieties of solipsism, since we allow semantic content to include reference to objects in the physical environment: a necessary condition for organisms or robots to function successfully.

12.12 Self-bootstrapped ontologies

I have been arguing that by analysing each type of architecture we can understand what sorts of processes can occur in it, and on that basis we can define an appropriate set of concepts for describing its 'mental' states, i.e. an appropriate mental ontology.

However, some learning mechanisms can develop their own ways of clustering phenomena, according to what they have been exposed to and various other things, such as rewards and punishments. If a system with the kind of meta-management layer depicted in the Cogaff architecture uses that ability on itself, it may develop a collection of concepts for categorising its own internal states and processes that nobody else can understand because nobody else has been through that particular history of learning processes. The role those concepts play in subsequent internal processing in such an architecture may exacerbate the uniqueness, complexity and idiosyncratic character of its internal processing.

For systems with that degree of sophistication and reflective capability, scientific understanding of what is going on within it may forever be limited to very coarse-grained categorisations and generalisations. This could be as true of robots as of humans, or bats (Nagel, 1981).

13 Human-like architectures

I have tried to bring out some of the design options that need to be faced when trying to explain the architecture of a human mind. When we understand what that architecture is, we can use it to define collections of concepts that will be useful for describing human mental states and processes, though we can expect to do that only to a certain degree of approximation for the reasons given in *section 12.12.* However that degree of approximation may suffice to provide useful clarifications of many of our familiar concepts of mind, such as 'belief', 'desire', 'intention', 'mood', 'emotion', 'awareness' and many others.

In particular, so many types of change are possible in such complex systems that we can expect to find our ordinary concepts of 'learning' and 'development' drowning in a sea of more precise architecture-based concepts.

We may also be in a better position to understand how, after a certain stage of evolution, the architecture supported new types of interaction and the development of a culture, for instance if the meta-management layer, which monitors, categorises, evaluates and to some extent controls or redirects other parts of the system, absorbs many of its categories and its strategies from the culture. It seems that in humans the meta-management layer is not a fixed system: not only does it develop from very limited capabilities in infancy, but even in a normal adult it is as if there are different personalities "in charge" at different times and in different contexts (e.g. at home with the family, driving a car, in the office, at the pub with mates, being interviewed for a job, etc.).

This suggests new ways of studying how a society or culture exerts subtle and powerful influences on individuals through the meta-management processes. The existence of the third layer does not presuppose the existence of an external human language (e.g. chimpanzees may have some reflective capabilities), though, as argued above, it does presuppose the availability of some internal information bearing medium or formalism, as do the reactive and deliberative layers.

When an external language develops, *one* of its functions may be to provide the categories and values to be used by individuals in judging their own mental processes (e.g. as selfish, or sinful, or clever, etc.) This would be a powerful form of social control, far more powerful than mechanisms for behavioural imitation, for instance. It might have evolved precisely because it allows what has been learnt by a culture to be transmitted to later generations far more rapidly than if a genome had to be modified. However, even without this social role the third layer would be useful to individuals, and that might have been a requirement for its original emergence in evolution.

We can also hope to clarify more technical concepts. The common reference to "executive function" by psychologists and brain scientists seems to conflate aspects of the deliberative layer and aspects of the meta-management layer. That they are different is shown by the existence of AI systems with sophisticated planning and problem solving and plan-execution capabilities without meta-management (reflective) capabilities. A symptom would be a planner that doesn't notice an obvious type of redundancy in the plan it produces, or subtle looping behaviour.

One consequence of having the third layer is the ability to attend to and reflect on one's own mental states, which could cause intelligent robots to discover qualia, and wonder whether humans have them.

All this should provide much food for thought for AI researchers working on multi agent systems, as well as philosophers, brain scientists, social scientists and biologists studying evolution.

14 Conclusion

This paper has presented a view which can be summed up by saying that the architecture of a human like mind is a complex system which in some ways is like what Minsky called a society of mind and in some ways like an ecosystem[10] of mind insofar as various components evolved within niches defined by the remainder of the system, just as happens in an ecosystem composed of different species of organisms. The mind is a collection of different evolved sub-species of sub-organisms.

This view and its detailed elaboration (barely begun in this paper) have important implications both for the science of mind (including animal minds and robot minds) and also for various engineering activities involving the production of systems to interact with minds, or to model minds for entertainment or other purposes.

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. The work is very difficult, and will need to pursued by different groups adopting different methodologies, some mainly empirical, some mainly theoretical and philosophical, some attempting to design working systems using a variety of approaches: bottom-up, top-down and middle-out, but with an open mind rather than dogmatic commitments to particular architectures or mechanisms. All these researchers will need to communicate their problems, their results and their failures to one another, since otherwise they will not fully understand the constraints and opportunities that are relevant to their own approach.

The particular sort of approach adopted here, emphasising architecture-based ontologies for mind, can bring some order into the morass of studies of affect, e.g. helping us understand why there are myriad rival definitions of "emotion" and helping us move to a more useful synoptic viewpoint. This is partly analogous to the way in which our concepts of kinds of physical stuff (e.g. as represented in the periodic table of elements) and kinds of physical and chemical processes were enriched through an understanding of the previously hidden architecture of matter.

For our task the challenge is much greater: whereas there is one physical world, with a single architecture (albeit with multiple levels) there are many architectures for organisms and possible intelligent machines, and correspondingly varied ontologies for mind. These are not, as is often supposed, differences of *degree* within a continuum of possibilities. There are many large and small *discontinuities* in 'design space', and we need to understand their implications, including the implications for evolutionary mechanisms and for individual learning and development.

A particular application will be development of a conceptual framework for discussing which kinds of emotions and other mental phenomena can arise in software agents that lack the reactive mechanisms required for controlling a physical body, replacing discussions which depend on arbitrary preferences for one or other definition of "emotion".

This work can lead to a better approach to comparative psychology, developmental psychology (the architecture develops after birth), and enhance the study of effects of brain damage and disease. Only when you have a good understanding of the normal functioning of a complex system can you hope to understand ways in which it can go wrong.

Although the scientific and philosophical implications of these ideas are profound, they are also relevant to engineers. The topics are relevant to designers of complex human-like systems for practical purposes, including new advanced forms of computer-based interactive entertainment or development of human-like software or hardware assistants, advisors, teachers, etc.

In particular the issues need to be understood by anyone who wishes
(a) to build systems that effectively model human mental processes, including affective processes,
(b) to design systems which engage fruitfully with human beings, since that requires an understanding of how

[10] The published version of this paper used the word 'ecology' here.

humans work (including an appreciation of the enormous individual variability among humans)
(c) to design good teaching systems for helping people learn mathematics, languages, science or other topics, since without understanding how minds work we cannot design systems (including classroom practices) that effectively enhance those modes
(d) to produce teaching/training packages for would-be counsellors, psychotherapists, psychologists, since those packages need to be based on good theories of how people function normally and how that normal functioning can be disrupted, damaged, etc.
(e) to produce convincing synthetic characters in computer entertainments, since shallow purely behavioural models may suffice for a while, but eventually they will be found to be dull, repetitive, rigid, and the task of extending the behavioural repertoires by adding more and more behaviours either by explicit programming or through imitative learning processes will turn out to be too tedious and too restrictive: deeper models can be expected to have more power and flexibility.

Acknowledgements

The ideas presented here were developed in collaboration with Steve Allen, Luc Beaudoin, Brian Logan, Catriona Kennedy, Ian Millington, Riccardo Poli, Tim Read, Edmund Shing, Ian Wright, and others in the Cognition and Affect Project, at The University of Birmingham.
I have also benefitted from interactions with, or reading the publications of, many colleagues elsewhere including Margaret Boden, Dan Dennett, Stan Franklin, Pat Hayes, John McCarthy, Marvin Minsky, Roz Picard, the members of the workshop on Emotions in Humans and Artifacts in Vienna, August 1999, and those who attended the 'Designing a Mind' workshop at the AISB Convention in April 2000.
Papers and theses by students and colleagues can be found at
http://www.cs.bham.ac.uk/research/cogaff/
Our software tools, including the SIM_AGENT toolkit, are available free of charge at
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
I have of course also learnt much from many others, including the authors listed in the bibliography. I am particularly grateful to Robert Trappl and Paolo Petta for making the workshop possible, and for their patience while I tried to get this paper into a presentable shape. I am aware that there are still many gaps and much that needs to be clarified.

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

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.

Dennett, D. C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, MA.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge Mass.

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

Frisby, J. P. (1979). *Seeing: Illusion, Brain and Mind*. Oxford University Press, Oxford.

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

Glasgow, J., Narayanan, H., and Chandrasekaran, B., editors (1995). *Diagrammatic Reasoning: Computational and Cognitive Perspectives*. MIT Press, Cambridge, Massachusetts.

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

Goodale, M. and 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.).

Karmiloff-Smith, A. (1996). Internal representations and external notations: a developmental perspective, in Peterson (1996), pages 141–151.

Lee, D. and Lishman, J. (1975). Visual proprioceptive control of stance. *Journal of Human Movement Studies*, 1:87–95.

Margulis, L. (1998). *The Symbiotic Planet: A new look at Evolution*. Weidenfeld & Nicolson, London.

Marr, D. (1982). *Vision*. Freeman.

McCarthy, J. (1995). What has artificial intelligence in common with philosophy? In *Proc. 14th International Joint Conference no Artificial Intelligence*. (Accessible via http://www-formal.stanford.edu/jmc/aiphil/aiphil.html).

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.

Nagel, T. (1981). What is it like to be a bat. In Hofstadter, D. and D.C.Dennett, editors, *The mind's I: Fantasies and Reflections on Self and Soul*, pages 391–403. Penguin Books.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, 18(1):87–127.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Nilsson, N. (1994). Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*, 1:139–158.

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

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

Okada, N. and Endo, T. (1992). Story generation based on dynamics of the mind. *Computational Intelligence*, 8:123–160. 1.

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

Peterson, D., editor (1996). *Forms of representation: an interdisciplinary theme for cognitive science*. Intellect Books, Exeter, U.K.

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

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

Rolls, E. (????). Precis of *The Brain and Emotion*. *The Behavioral and Brain Sciences*. (To appear).

Rolls, E. (1998). *The Brain and Emotion*. Oxford University Press, Oxford.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence, A Modern Approach*. Prentice Hall.

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

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

Simon, H. A. (1969). *The Sciences of the Artificial*. MIT Press, Cambridge, Mass. (Second edition 1981).

Sloman, A. (1969). How to derive "better" from "is". *American Phil. Quarterly*, 6:43–52. Online at http://www.cs.bham.ac.uk/research/cogaff/sloman.better.html.

Sloman, A. (1978). *The Computer Revolution in Philosophy*. Harvester Press (and Humanities Press), Hassocks, Sussex. Online at http://www.cs.bham.ac.uk/research/cogaff/crp.

Sloman, A. (1985). What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles.

Sloman, A. (1987a). Motives mechanisms and emotions. *Cognition and Emotion*, 1(3):217–234. Reprinted in M.A. Boden (ed), *The Philosophy of Artificial Intelligence*, 'Oxford Readings in Philosophy' Series, Oxford University Press, 231–247, 1990.

Sloman, A. (1987b). Reference without causal links. In du Boulay, J., D.Hogg, and L.Steels, editors, *Advances in Artificial Intelligence - II*, pages 369–381. North Holland, Dordrecht.

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. (Available at http://www.cs.bham.ac.uk/research/cogaff/).

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, Cambridge, UK.

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. (1996a). Actual possibilities. In Aiello, L. and Shapiro, S., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638, Boston, MA. Morgan Kaufmann Publishers.

Sloman, A. (1996b). Towards a general theory of representations. In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science*, pages 118–140. Intellect Books, Exeter, U.K.

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). Models of models of mind. In Lee, M., editor, *Proceedings of Symposium on How to Design a Functioning Mind, AISB'00*, pages 1–9, Birmingham. AISB.

Sloman, A. (2002). Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)*, pages 403–427, Dordrecht. Kluwer. (Synthese Library Vol. 316).

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.

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

Wright, I. (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., 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.