Varieties of Affect and the CogAff Architecture Schema

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

Abstract

In the last decade and a half, the amount of work on affect in general and emotion in particular has grown, in empirical psychology, cognitive science and AI, both for scientific purposes and for the purpose of designing synthetic characters, e.g. in games and entertainments. Such work understandably starts from concepts of ordinary language (e.g. "emotion", "feeling", "mood", etc.). However, these concepts can be deceptive: the words appear to have clear meanings but are used in very imprecise and systematically ambiguous ways. This is often because of explicit or implicit pre-scientific theories about mental states and process. More sophisticated theories can provide a basis for deeper and more precise concepts, as has happened in physics and chemistry. In the Cognition and Affect project we have been attempting to explore the benefits of developing architecture-based concepts, i.e. starting with specifications of architectures for complete agents and then finding out what sorts of states and processes are supported by those architectures. So, instead of presupposing one theory of the architecture and explicitly or implicitly basing concepts on that, we define a space of architectures generated by the CogAff architecture schema, where each supports different collections of concepts. In that space we focus on one architecture H-Cogaff, a particularly rich instance of the CogAff architecture schema, conjectured as a theory of normal adult human information processing. The architecture-based concepts that it supports provide a framework for defining with greater precision than previously a host of mental concepts, including affective concepts. We then find that these map more or less loosely onto various pre-theoretical concepts, such as "emotion", etc. We indicate some of the variety of emotion concepts generated by the H-Cogaff architecture A different architecture, supporting a different range of mental concepts might be appropriate for exploring affective states of other animals, for instance insects, reptiles, or other mammals, and young children.

1 Introduction

The study of emotions is not a new topic, even in AI, as shown by Simon's important contribution over 30 years ago (Simon, 1967), and various papers nearly 20 years ago in IJCAI'81 including my attempt (with Monica Croucher, (1981)) to show why intelligent autonomous robots designed to cope with a rich, dynamically varying, partly unknown, environment, will have the capacity to have certain sorts of emotions, as a side-effect of other design decisions. However, in the last decade and a half, the study of affect in general and emotion in particular has become fashionable in scientific psychology, cognitive science, AI and philosophy. For instance, a leading journal on philosophy of science recently included an article on a computational theory of mood Sizer (2000).

There are at least three different motivations for the interest in computer models of emotions:
(i) an interest in emotions (in humans and other animals) as something to be modelled and explained,
(ii) a desire to give machines which have to interact with humans an understanding of emotions as a requirement for some aspects of that task (Sloman, 1992), and
(iii) a desire to produce new kinds of computer-based entertainments where synthetic agents, e.g. software agents or "toy" robots, produce convincing emotional behaviour.

The requirements for objective (iii), entertainment, are not necessarily the same as for objective (i): since "believable" behaviour in constrained contexts could be the product of widely different models, including at one extreme very large, hand-coded lookup tables specifying what to do when. To some extent this may also work for the second objective, provided that the interaction context is very limited, but in the long run a deep and accurate model of the first type may be required for effectively achieving goals of type (ii). This paper[1] is primarily concerned with objective (i), in particular understanding and modelling human emotions (along with other mental states and processes, since emotions cannot be understood in isolation). Much of the discussion is also relevant to objectives (ii) and (iii) in ways that will not be explained here.

[1]Presented at Symposium on Emotion, Cognition, and Affective Computing at the AISB'01 Convention, 21st - 24th March 2001

2 Architecture-based concepts

Modelling and explaining emotions and other mental phenomena in humans and other animals requires us to use concepts referring to those phenomena. The history of the philosophy of mind, and some of the methodological, terminological and scientific disagreements found in psychology and neuroscience, all point to serious problems in defining these concepts. In the Cognition and Affect project we have been attempting to explore the benefits of developing architecture-based concepts, i.e. starting with specifications of (virtual machine) architectures for complete agents and then finding out what sorts of states and processes are supported by those architectures.

We can illustrate this approach with a non-mental concept, using the familiar concept "thrashing" in an operating system. In a multi-processing operating system with a time-sharing scheduler and a virtual memory mechanism it is a common observation that as the number of large processes increases the more time is spent on swapping and paging as opposed to doing useful work. We can then define a state of "thrashing" as one in which more than half the time is spent swapping and paging. "Deadlock" is another familiar architecture-based concept.

Architecture-based concepts are defined in terms of causal interactions between states and processes within mechanisms in a virtual machine architecture, and in that sense they involve a functional perspective. This is different from the familiar philosophical variety of functionalism that defines mental states in terms of relationships between inputs and outputs of the *whole system* without any mention of the internal architecture. Notice also that our notion of functionalism does not require the concepts so defined to refer to mechanisms or states or processes that have a useful function. As the "thrashing" example shows, mechanisms that do have useful functions can interact so as to produce emergent states that do not. This is very likely to be true of at least some human mental phenomena, which is why therapists are often required!

We can attempt to clarify our pre-scientific concepts of mind using architecture-based concepts that refine and extend them. We first define a space of architectures generated by the CogAff architecture schema, described below, where each architecture supports different sets of possible states and processes. For each architecture, partitions of the set can define concepts of states and processes supported by the architecture. In some architectures we may find analogues of many familiar concepts, e.g. learning, motives, intentions, beliefs, moods, self-awareness. In other architectures only an impoverished set of such concepts will be supported.

In the space of architectures defined by the CogAff schema, we focus on one architecture H-Cogaff, described below, a particularly rich instance of the CogAff architecture schema, conjectured as a schematic theory of human information processing. It is schematic insofar as many details remain to be filled in. Instances of H-Cogaff support architecture-based concepts that provide a framework for defining, with greater precision than ever before, a host of mental concepts, including affective concepts. We then find that these new precise concepts map more or less loosely onto various pre-theoretical concepts, such as "emotion", etc. (Something like this happened to other pre-theoretical concepts as architecture-based concepts of kinds of stuff developed in physics and chemistry during the last two centuries.)

We indicate below some of the variety of emotion concepts generated by the H-Cogaff architecture. Different architectures (also consistent with the general CogAff schema) might be appropriate for exploring affective states of insects, or reptiles, or other mammals, or newborn infants.

In a more general investigation we can study properties of different architectures both analytically and by producing simulations. For instance, in this symposium Scheutz and Logan (2001) describe simulation experiments comparing some very simple varieties of architectures subsumed by CogAff in a variety of environments. This sort of investigation is relevant to finding out under which conditions evolutionary transitions from one architecture to another might occur, which is one of the objectives of the Cognition and Affect Project, described in http://www.cs.bham.ac.uk/~axs/cogaff.html

3 Do we know what we are talking about?

Specifying what we are talking about generates difficult conceptual problems. Whichever of the three motivations listed above drives the modelling of emotions and other mental phenomena, the work understandably starts from concepts of ordinary language (e.g. "emotion", "mood", "feeling", "pleasure", etc.). These concepts can be deceptive to those not trained in philosophical analysis. The concepts are so familiar that they appear to have very clear, commonly understood, meanings, whereas detailed analysis shows that the opposite is true: the familiar labels often refer to concepts that are riddled with confusion and ambiguity, and when people attempt to define them they come up with widely different definitions.

For instance in the psychological literature there are a multitude of definitions of "emotion", some stressing brain processes, some stressing peripheral physiological processes, some stressing patterns of behaviour, some stressing eliciting conditions, some stressing the functional roles, some stressing introspective qualities. This diversity was already evident long ago in the collection edited by Magda Arnold (1968).

The definitions also differ in scope: for instance some writers treat all motives or desires (e.g. hunger, curiosity) as emotions while others do not. Some regard surprise as an emotion, whereas others (e.g. Ortony et al. (1988)) regard it as basically a cognitive state in which a belief or expectation has been found to be violated, which may or may not produce an emotional reaction. Ortony et al., like many others, claim that being experienced is a necessary condition for an emotion (p. 176), whereas it is not uncommon for novels or plays to include characters who are totally unaware that they are infatuated, or jealous, even though other individuals notice the state. In this case, the novelists and playwrights have the deeper insight into the nature of emotions!

Discussion of some of the diversity of approaches and definitions can be found in Oatley and Jenkins (1996). Although there are many excellent surveys of issues concerning emotions,[2] it is difficult for newcomers to the field to achieve a balanced overview, and in consequence there is sometimes a tendency to present simplistic AI programs and robots as if they justified epithets like "emotional", "sad", "surprised", "afraid", "affective", etc. without any deep theory justifying these labels. This, for instance, is why Boden referred to PARRY, the simulated paranoid program, as a "fraud" (Boden, 1978) (though this was not intended as a criticism of its author, Colby, who was always open about what the program could and could not do). Likewise, McDermott (1981) lambasted the tendency of AI researchers to use terms like "goal", "plan", "learn", simply because there are procedures or variables with these names in a program. His criticism was directed at symbolic AI programs, but similar comments can be made about labels applied to neural and other models.

In previous papers[3] we have recommended analysing mental concepts on the basis of the types of states and processes supported by particular virtual machine architectures, and below we illustrate this approach. However starting from over-simple architectures can lead to shallow concepts, for example, assuming that emotional states are implemented in one or more emotional state variables (e.g. happiness, sadness, anger, fear, etc.), with either boolean values that can be toggled or numerical or "qualitative" ranges of values. There may be some biological states that involve such explicit state representations, but in general such models (e.g. anger in PARRY) are grossly inadequate as accounts of typical human social emotions which are rich in semantic content, for instance being angry with a particular person about a particular action performed by that person, or feeling humiliated because some silly mistake you made was pointed out by a famous person in a large public lecture. An interactive artificial counsellor which assumed that anger was simply some sort of continuously variable global state (like some moods) rather than a semantically directed state might make inappropriate comments to its clients.

Of course, there are human states that may vary in degree or intensity, but from that it does not follow that a good explanatory model of such a state should simply use a variable with a numerical value, to represent such a state: the change in intensity might be an emergent feature of both the number and the variety of processes of certain sorts that become activated. Likewise the fact that more or less thrashing can occur in an operating system does not imply that the operating system includes a numerical variable whose value is the degree of thrashing. A self-monitoring operating system *might* measure the ratio of useful computation to time spent paging and swapping and use that ratio to take some decision, e.g. disabling new logins or killing very large processes. But there does not have to be any such explicit numerical representation for the thrashing to exist and to vary in amount.

In short, it is important not to assume that the forms of representation that are useful for scientists and others to use when *describing* a complex system or predicting its behaviour are to be found in the system itself. Moreover, when a system with the meta-management capabilities described below does monitor itself and detect aspects of its own behaviour, the existence of the process detected and the existence of the process of detection and categorisation should not be confused. The detected process might be far more complex than the detecting process.

4 How to make progress

There are several different strategies for dealing with these conceptual confusions. One is to ignore them and proceed as if everything were clear, as may occur when new graduates in subjects like computer science or mathematics embark on AI projects, assuming that they know what emotions are, and without any knowledge of philosophy, psychology, linguistics, etc.

Another strategy, to be found in many psychology departments, is to search for operational definitions of various states in terms of measurable aspects of behaviour, physiological changes, etc. This approach often uses empirical correlations between such measurables and intuitive judgements about emotions elicited from experimental subjects in simple situations. (E.g. if people who are thought to be angry often frown then frowning might be taken as part of the definition of anger.) The development of non-invasive brain scanning devices will probably lead to new variants of this type of definition based on correlations.

A very different approach is to do surveys of linguistic usage, either using questionnaires or analysis of published texts to attempt to extract rules for the use of words like "emotion", "feeling", etc. Because of individual variations in usage this may come up with probabilistic rules (e.g. a person with such and such a facial expression and such and such behaviour has probability X of being angry).

Yet another approach (e.g. recommended by Oatley) is to study the role of emotions in literature and to try to derive therefrom a theory consistent with the role of emotions and the references to emotions in stories, plays,

[2]E.g. (Ortony et al., 1988; Goleman, 1996; LeDoux, 1996; Picard, 1997)

[3]E.g. Sloman (1984, 1985, 1987, 1992, 1994, 1998, 1999)

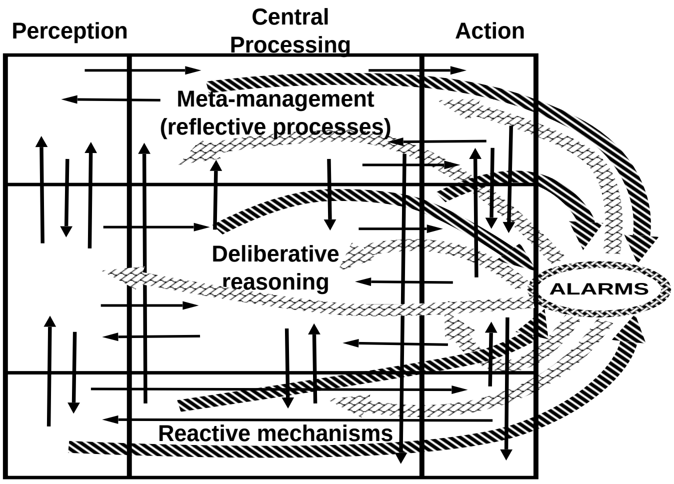

Figure 1: *The CogAff Schema: pillars, layers and alarms*
If we consider a system in which both the division between perceptual, central and motor systems can be made, and also the division between reactive, deliberative and meta-management layers, and if we assume that the perceptual and motor systems include components related to the needs of all three central layers, then we have a three by three grid of architectural components with different sorts of functionality. The nature of each component is defined by its functional connections to all the others. If some of the internal processing is slow relative to the speeds at which things happen in the environment, then it may be useful to have inputs from many parts of the system to a fast pattern driven reactive "alarm" mechanism that can redirect the whole system. Solid arrows are as before. The shaded arrows represent information flowing to and from the alarm mechanism. The alarm mechanism being purely reactive and pattern driven will typically be stupid and capable of mistakes, but may be trainable. CogAff covers a wide variety of architectures containing various subsets of the schema. Fig. 2 shows a particular example H-Cogaff.

poems. This might include analysing emotions as aspects of evolving patterns of social relationships for instance. Since stories vary from one culture to another this can lead to a theory of emotions as largely culture-relative, whereas many psychologists regard emotions as universal, at least in humans. Both views are partly correct, but this point will not be discussed here.

Subtly different from such empirical investigations are philosophical attempts at conceptual analysis which start from the assumption that we cannot reliably articulate the rules by which we use most of our concepts, so that analysing concepts requires a cycle of conjectures, testing with examples, and then modifying the conjectures. This was widely used by analytical philosophers in the second half of the 20th century, e.g. J.L.Austin, G.Ryle, L.Wittgenstein. A summary of the techniques was presented in Sloman (1978), ch 4.

An approach favoured by some evolutionary theorists is to attempt to understand the biological value of many of the kinds of behaviours regarded as emotional and on that basis to define different kinds of emotions in terms of their biological functions. Darwin was a major contributor to this approach (Darwin, 1872). An extreme view (not held by Darwin) would be that all emotions have functions. This does not allow for a type of emotion that is a result of interactions between functional components of an organism but which does not in itself have any useful function. For instance, grief and embarrassment might be such "emergent" states. It is also likely that many emotions may have social functions insofar as they are social control mechanisms, even though they do no good for the individual concerned, e.g. feeling guilty about some alleged sin. Proponents of a biological approach sometimes also differ as to whether there is some special biological module that produces all emotions, or whether some or all of them are states that arise out of interactions between other modules (Sloman, 1992).

5 Architecture-based concepts of mind

Although there is something to be learnt from all of those approaches we feel that most of them suffer by not construing an animal (e.g. a human) or a robot as employing an information processing architecture containing various kinds of coexisting interacting sub-mechanisms whose states, processes and interactions account for its mental states and processes. The precise combination of mechanisms will vary from species to species (Dennett, 1996), and possibly also between individuals within a species.

The varieties of types of mental states and processes possible will vary from one architecture to another, and therefore the sets of concepts applicable to different species will be different, except insofar as they share certain aspects of their architectures, or certain functions achieved by their architectures. Architecture-relative versions of mental concepts allow us to transform ill-defined questions into questions on which we can make progress[4]

Probably everyone would agree that a flea cannot wonder how many prime numbers there are (why?) but whether it might be in pain would be a matter for endless debate, because of the indeterminacy and confusion in the concept of "pain". (See Dennett's discussion of pain in Dennett (1978).) Architecture-based concepts of "pain" allow such debates to switch to using precise concepts, so that precise, answerable questions can be formulated. Explicitly distinguished concepts will then lead to different questions with different answers.

This is related to the standpoint of Simon's seminal paper. It also partly reflects the standpoint of Ortony et al. (1988) who eschew arguments about what particular emotion words and phrases actually mean and instead attempt to survey a space of possible concepts, which they (implicitly) base on a theory of the human cognitive architecture, insofar as they assume that agents have beliefs, desires, intentions, uncertainty, etc. We can generalise that approach by not restricting ourselves to a single architecture.

In our own work we have been developing an architecture schema, called CogAff, shown in Fig. 1, which provides a framework for describing different kinds of architectures and sub-architectures, and which, to a first approximation, is based on superimposing two sorts of distinctions between components of the architecture: firstly the distinction between perceptual, central and action components, and secondly a distinction between types of components which evolved at different stages and provide increasingly abstract and flexible processing mechanisms within the virtual machine (Sloman, 2000; Sloman and Logan, 2000; Sloman, (to appear).

By analysing some of the types of states and processes that can occur within different variants of the architecture schema we find that our intuitive notions of affect, emotion, perception, belief, and other mental states and processes, correspond, in a not very determinate manner, to many different, precisely definable, concepts related to particular classes of architectures. This is something like the way in which concepts of kinds of physical stuff correspond loosely to the concepts of types of elements and compounds that are definable on the basis of the architectures of atoms and molecules.

6 Three levels

A first crude sub-division of architectural components arises out of three levels of sophistication in biological information processing architectures, which can also be found in artificial architectures. We conjecture that these levels, depicted in Fig. 1, emerged at different times in biological evolution.[5]

6.1 Level 1 (Reactive mechanisms)

Reactive systems can be defined mainly negatively: they are systems which lack the ability to represent, evaluate and compare possible actions, or possible future consequences of actions. They sense internal or external conditions and then respond by producing internal or external state changes (or some combination). There may be competing reactions but these will be resolved by some mechanism that does not involve deliberation or making inferences. E.g., it could use vector addition to produce a combination or compromise response, or the selection between options might be controlled by a state variable that is modified by some other reactive mechanism. Systems built entirely out of reactive components may be capable of producing extremely complex behaviour, and as insects and simpler organisms demonstrate, they can be biologically very successful, if success is measured in terms of biomass, numbers of individuals, generations of existence.

Many reactive systems use an information-processing architecture with a fixed collection of condition-action associations. However they may be capable of changing by modifying weights, or even by generating new associations through something like Hebbian learning. They can be implemented in a variety of mechanisms, including neural nets, symbolic condition-action rules, chemical mechanisms, and so on. Such purely reactive organisms would be driven largely by genetically determined mechanisms along with minor changes produced by learning.

In principle any desired combination of competences can be produced by purely reactive systems, but at the cost of potentially explosive requirements for storage and for training or evolution times. It is this trade-off that probably led to the evolution of deliberative mechanisms.

Within a purely reactive architecture it is possible to distinguish what might be described as "normal" operation from states produced by detection of threats or opportunities requiring rapid and speedy redirection of processing. The organism (or robot) need not have the concept of a "threat" or "opportunity" merely (possibly innate) mechanism which in fact detect instances (possibly sometimes erroneously). The detection could lead to appropriate behaviour even though the organism has no conception of the purpose of the behaviour.

These reactions could be described as proto-emotions, which we would expect to find in insects and other purely reactive organisms. They are primitive, evolutionary precursors, of the more familiar types of states and processes found in humans and other more complex animals. Re-

[4]Without throwing away the substance of the original question, like looking for lost keys only in the lamplight. However, arguing that is beyond the scope of this paper.

[5]The mechanisms are also likely to be relevant to some applications of AI e.g. because the mechanisms will be useful in certain sorts of robots and software agents, just as they are in animals.

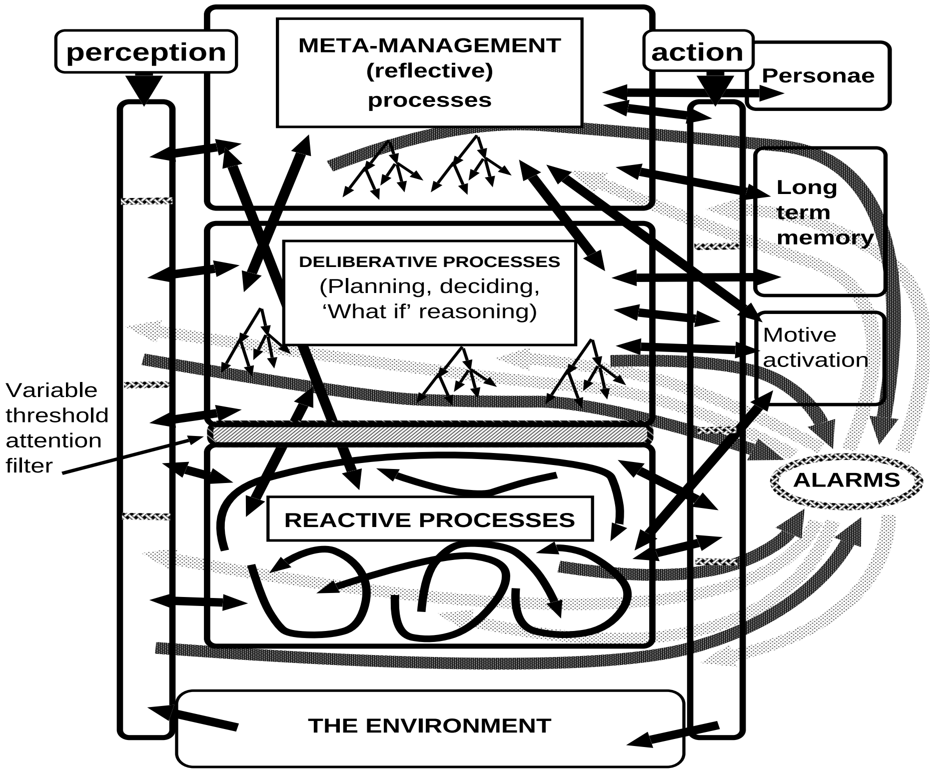

Figure 2: *H-Cogaff – a three layer architecture.*
Fig. 1 shows a schema containing a collection of permitted *components. This figure displays components* required *for H-cogaff, the proposed human-like architecture. The meta-management layer provides the ability to attend to, monitor, evaluate, and sometimes change internal processes and strategies used to produce internal processes. However, all the layers and the alarm system(s) operate concurrently, and none is in total control. A collection of high level culturally determined "personae" may be available, turned on and off by different contexts and causing global features of the behaviour to change, e.g. switching from bullying to servile behaviour. Note that some of the divisions between layers are a matter of taste: some authors e.g. Davis (1996) prefer to separate out reflexes from the reactive layer, and some e.g. Minsky (2000) prefer to separate out some of the high level functionality of the meta-management layer as an extra layer.*

active systems cannot have the kinds of emotions that depend on Levels 2 and 3, described below.

In particular, lacking the types of semantic apparatus involved in level 2, and lacking the self-monitoring capabilities provided by level 3, they would have no understanding of what they are doing or why they are doing it.

Within the reactive framework, it is also possible to distinguish states in which different kinds of needs dominate processing. These could be described as proto-desires or proto-motives: i.e. primitive, evolutionary precursors of the more familiar varieties. If the organism is capable of reinforcement learning we might describe the states in which it receives negative reinforcement ("punishment") as a type of pain, or proto-pain. But such an animal would not know that it is in pain, it would merely react externally by aversive behaviour and internally by changing contingencies for future behaviours.

More global states of a reactive system which are less goal directed but change the quality of processing in some general way might be described as proto-moods, e.g. where behaviour tends to be very cautious or very aggressive. These might be triggered by features of the environment that change some aspect of internal state (e.g. concentrations of some chemical) which modulate a wide range of behaviours. Again this could occur without the organism having the representational capability to describe such a state nor the self-monitoring mechanisms to detect it.

6.2 Level 2 (Deliberative mechanisms)

As the architecture becomes more complex, and deliberative capabilities are added which provide the opportunity to represent, analyse, compare, evaluate and react to descriptions of *hypothetical* future scenarios or possible explanations of previously observed phenomena, a new class of architecture-based states and processes will become possible.

Exactly which ones are supported will depend on the sophistication of the deliberative mechanisms. For instance, human deliberative capabilities include being able to short term and long term desires of varying levels of abstraction, high level ideas, and support more or less abstract and complex deliberations, predictions, and hypotheses. We can construct plans or conjectures of varying structure and complexity, requiring the ability to manipulate structures with sufficiently rich syntax to express these contents, along with something like a compositional semantics, and a very flexible and powerful re-usable memory for thoughts, conjectures, partial plans, and various kinds of reasoning. By contrast, some simpler organisms (and many current robots) may be able to do 'what if' reasoning only about possibilities with fixed flat structures ("food that way", "danger this way", "find food", "find drink", "avoid obstacle", "hit that" etc.) There is still much to be learnt about the nature of human deliberative capabilities and how they are implemented in brains, or how they might be implemented in computers.

However, it is clear that these capabilities can interact with emotional processes. The realisation that some highly valued result is easily achievable in the near future, or the discovery that a selected plan is fraught with danger, could produce kinds of affective states and processes that are not possible in a purely reactive architecture, even though partly similar, much simpler states can exist. A reactive organism may have a kind of fear in its response to a presently perceived threat, whereas a deliberative mechanism permits apprehension about possible remote consequences of actions being contemplated. The semantic complexity of the varieties of hope, anticipation, apprehension, available to such an organism would depend on the type of representational apparatus supported by the architecture.

Insofar as the operations of a deliberative mechanism involve use of structured representations with a compositional semantics, many of the affective states that can arise in such a system will have rich and varied semantic contents, unlike those supported by a purely reactive architecture. Some of this richness is illustrated in the classification of emotions and attitudes by Ortony, Clore and Collins. This is the sort of thing that has led many philosophers to argue that emotions cannot be separated from cognition, whereas some psychologists are inclined to treat emotions as semantics-free, purely reactive states, a view that is more appropriate to organisms with only reactive architectures.

Context-dependent global modulation of goal-generating processes, goal-comparisons, plan-construction, plan-evaluation, plan-execution, provides a basis for a further family of concepts referring to affective states and processes not possible in a purely reactive architecture. For instance the kind of caution manifested by a deliberative agent that has thought of the possibility of being detected by a predator is different from the kind of caution (proto-caution?) observable in a purely reactive organism whose innate reactive rules are triggered by the smell of a certain predator to modulate normal reactive behaviours, without the animal having any knowledge of what might happen if it did not move cautiously.

Of course, although an architecture-based distinction can be made between what we have described as purely reactive proto-caution and deliberation-based knowledge-rich caution, the externally observable behaviours produced by those states may be indistinguishable. So determining which state should be attributed to an organism (or robot) will require finding out something about its information processing architecture. That will in general be a difficult task. (E.g. I may be wrong in assuming that insects are purely reactive!)

Another class of processes that can occur if there is a deliberative layer present, involves various types, frequencies, and strengths of interruptions of deliberative processes, arising out of processes in the reactive layer, or arising out of perceptual processes, or even some triggered by deliberative processes themselves. If there is no deliberative layer these "perturbances" (Wright et al., 1996) cannot occur.

If the need to limit such disruptions is addressed by the evolution (or design) of some kind of variable-threshold filtering mechanism, as suggested in Sloman (1992) and Beaudoin (1994) then an additional class of states and processes corresponding to modifications of the attention filter threshold can be distinguished. The ability of humans to be more or less *absorbed* in what they are doing seems to be related to varying interrupt thresholds for such attention filters.

6.3 Level 3 (Metamanagement mechanisms)

Beaudoin (1994) suggested that in addition to the first two levels, a human-like architecture requires a "reflective" or "meta-management" layer, shown both as a permitted component in the CogAff schema in Fig. 1 and as a required part of the H-Cogaff architecture in Fig. 2. This permits self-observation or self-monitoring of a wide variety of internal states, along with categorisation and evaluation of those states, linked to high level mechanisms for learning and for controlling future processes. Examples of the operation of meta-management might be:

- The ability to think about and answer questions about one's own thoughts and experiences, e.g. noticing that a rectangular surface looks like a parallelogram from certain viewpoints, even though it is still perceived as rectangular (i.e. the *qualia* change but not the perceived 3-D shape).

- The ability to notice and report on circularity in one's thinking ("I decided to B in order to achieve A. I decided to do C in order to do B. I decided to do A in order to do C. I then noticed that I was thinking in circles.").

- The ability to notice that one is not attending to a task judged as important ("I really should be reading this student exercise, not thinking about what happened last night").

- The ability to notice opportunities for changing one's thinking ("I solved this problem much faster than the previous one: so what exactly did I do this time?")

Where such a layer is present yet another family of concepts becomes applicable for describing states and processes involving the third layer.

The ordinary usage of some of these concepts might refer to states and processes that can occur without this layer but become enriched when the layer is present.

For instance the kinds of apprehension or anticipation that might occur in a system with reactive and deliberative layers could also be detected, evaluated, and produce a second-order reaction in a system with the third layer, so that the states of apprehension or anticipation have extra dimensions, e.g. combining whatever sorts of positive or negative evaluations the deliberative system achieves with additional evaluations linked to self-awareness.

It is also possible for processes in other layers to disrupt the third layer and to over-ride some of its decisions, leading to yet more complex states and processes which are possible only when the third layer is present. For instance if the meta-management layer attempts to direct deliberation and other processes at a particular task and other processes manage to divert attention from that task, then this loss of control, which is common in many familiar human emotions, is a type of state that is impossible without the third layer: you cannot lose control that you've never had. A deliberative system might be constantly diverted by non-deliberative processes but not detect that this is what is happening to it.

Further architecture-based conceptual distinctions could be related to different modes of operation of the third layer, e.g. which sorts of internal processes it is capable of detecting, which modes of categorisation and evaluation it is can use, and which sorts of control it has over other processes. As with the deliberative layer we can distinguish varying degrees and kinds of sophistication in the representational apparatus available to the third layer. It may or may not be similar to the forms of representation used in the second layer.[6]

7 Pleasure and pain

It should now be clear that some of the ambiguities in our ordinary concepts of mind may be due, in part, to the fact that they sometimes refer to relatively simple states that are supported by relatively simple architectures and manifested in behaviours that require only those architectures, whereas they sometimes refer to far more complex states, especially when used in discussing human emotions connected with social relationships or self-awareness.

For example, consider purely reactive organisms (or robots) with aversive or seeking behaviours, with tendencies to avoid or reduce certain states and to achieve and preserve others, and with reinforcement learning mechanisms that support positive and negative reinforcement. It would be possible to use the words "pain" and "pleasure" to refer to states of such an organism, and perhaps that is what happens when people think of an insect as being in pain if exposed to a noxious chemical or having some sort of pleasure when feeding.

But those states are extremely primitive in comparison with the states that also include explicit recognition of goals as having being subverted or achieved, or harm being done, or needs fulfilled. Even those can occur without awareness that they are occurring, in organisms with the first two levels but lacking the third. When the third level is present the additional explicit characterisation and evaluation of the state, along with internal high level reactions triggered by that, begin to reach the sort of complexity involved in many human pains and pleasures.

Other authors Damasio (1994); Goleman (1996); Picard (1997) have distinguished primary and secondary emotions. I have tried to show elsewhere Sloman (1998, 2000, 1999); Sloman and Logan (2000) that those ideas can be both explained and generalised by relating primary emotions to the capabilities of the reactive layer of H-Cogaff (also found in simpler architectures), relating secondary emotions to disturbances triggered by events in the deliberative layer, and introducing *tertiary* emotions as perturbances involving partial loss of control of the metamanagement layer, for instance when a person who is infatuated or embarrassed finds it hard to think about tasks unrelated to the cause of the infatuation or embarrassment. Within the H-Cogaff framework we can begin to introduce far more refined distinctions between different types of emotions and other affective states, according to which components are involved and how they interact. My guess is that most of the emotions that are of interest to humans, and therefore figure in plays, novels and gossip, involve the third layer, whereas emotions primarily involving ancient reactive mechanisms are the ones that are easiest to study in laboratories, and therefore get more attention in the scientific literature. They are also easiest to simulate on computers! However, some simple simulations involving all three layers have been and are being developed using our SimAgent toolkit, e.g. Wright (1977); Scheutz and Logan (2001)

[6] The idea of meta-management is related to Minsky's "C-brain" idea in his Minsky (1987), and to the "commentary" idea in Weiskrantz (1997). Catriona Kennedy is exploring a type of mutual meta-management in secure software systems, using our toolkit. See http://www.cs.bham.ac.uk/~cmk. The common notion of "executive function" in psychiatry and psychology does not clearly distinguish the deliberative and meta-management capabilities. Much early AI work was on systems with level 2 but no level 3.

8 Conclusions

This paper attempts to show how the variety of affective and cognitive states of which an organism or robot is capable can vary according to which of the three architectural levels is present and which sorts of capabilities (e.g. which representational and semantic capabilities) are available within each level. This provides a framework for analysing, refining, and extending many of our ordinary concepts of mind. Although we have focussed primarily on concepts concerned with affective states it should be clear that the analytical framework provided is far more general. For instance, within this framework far more varieties of learning and development can be separated out than are normally distinguished.

Our own motivation for this work is primarily the scientific and philosophical goal of understanding how humans and other animals work and also what sorts of robots and software agents are possible. But the same considerations could be relevant to a variety of practical applications of AI, as indicated in the introduction.

It must be stressed that it is not only the information processing architecture that determines what sorts of affective states and other mental states are possible. It is clear that social and other external factors are relevant also. For instance, in a social system without any notion of marriage or commitment to a sexual partner it will not be possible for an individual to be ashamed or feel guilty about being unfaithful. Moreover, there are many emotional states that depend on the existence of other agents, including embarrassment, shyness, envy, gloating, etc. However, many of these presuppose the sorts of architectures that we have been discussing.

A more extensive discussion, for which this paper does not provide space, would explain in more detail the ideas underlying the CogAff architecture schema and show how a very wide variety of concepts referring to what would intuitively be described as "affective" states and processes can be defined in terms of the various types of information processing and control states supported by different variants of the architecture, in which different subsets of the architecture are present.

In particular, this will help to show that both the subdivision of emotions into "primary" and "secondary" emotions in the works of Damasio, Goleman, Picard, etc. and the extension to include "tertiary" emotions in Sloman (1998, 2000); Sloman and Logan (2000) merely scratch the surface of a far more complex and varied space of phenomena.

From this viewpoint, arguing about which definitions of the various types of mental concepts are *correct* is pointless, like arguing over whether the mathematician's concept of ellipse (which includes circles) or the wheelwright's concept of ellipse (which excludes circles) is correct. The important point is to understand the space of possibilities and the implications of the different architectural underpinnings of different sorts of concepts. We can even use this approach to investigate different types of consciousness[7] supported by different sorts of architectures, and perhaps provide new clarity in debates about consciousness.

Acknowledgements

This work was funded by a grant from the Leverhulme trust. Many of the ideas here grew out of the PhD work by Luc Beaudoin (1994) and later Ian Wright (1997). I am grateful for help received more recently from Brian Logan and Matthias Scheutz. The debts to many others, e.g. Simon, Minsky, Dennett, will be obvious. The bibliography gives only a sample.

The SimAgent toolkit used in our simulation work can be found at the Free Poplog Site:
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
It is described here
http://www.cs.bham.ac.uk/~axs/cogaff/simagent.html
It owes much to contributions from Riccardo Poli, Jeremy Baxter (DERA), Richard Hepplewhite (DERA), Darryl Davis (Hull), Brian Logan (Nottingham), Catriona Kennedy, Matthias Scheutz.

References

M.B. Arnold, editor. *The Nature of Emotion*. Penguin Books, Harmondsworth, England, 1968.

L.P. Beaudoin. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham, 1994. (Available at http://www.cs.bham.ac.uk/ research/ cogaff/).

Margaret A. Boden. *Artificial Intelligence and Natural Man*. Harvester Press, Hassocks, Sussex, 1978. Second edition 1986. MIT Press.

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

Charles Darwin. *The Expression of the Emotions in Man and Animals*. Harper Collins, London, 1872. (Reprinted 1998).

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

[7]For instance, a housefly is, in a sense, conscious of something moving rapidly towards it, which is why it escapes the fly-swat, but it is not conscious that it is conscious. That requires something like meta-management. Being conscious of dangers in a proposed plan requires a deliberative layer. Being conscious of features of your sensory percept, i.e. having qualia, requires a meta-management layer with links to intermediate stages in perceptual mechanisms.

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

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

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

Joseph E LeDoux. *The Emotional Brain*. Simon & Schuster, New York, 1996.

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

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

M.L. Minsky. Future Models for Mind-Machines. In A.Sloman et al., editor, *Proceedings Symposium on How to Design a Functioning Mind AISB00 Convention*, pages 124–129, 2000.

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

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

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

M. Scheutz and B.S. Logan. Affective vs. deliberative agent control. In C. Johnson et al., editor, *Proceedings Symposium on Emotion, cognition and affective computing AISB01 Convention*, York, 2001.

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

L. Sizer. Towards a computational theory of mood. *British Journal for the Philosophy of Science*, 51:743–769, December 2000. ISSN 0007-0882. 4.

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

A. Sloman. The structure of the space of possible minds'. In S. Torrance, editor, *The Mind and the Machine: philosophical aspects of Artificial Intelligence*. Ellis Horwood, Chichester, 1984.

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

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

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

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

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

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

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

A. Sloman. How many separately evolved emotional beasties live within us? In Robert Trappl and Paolo Petta, editors, *Emotions in Humans and Artifacts*. MIT Press, Cambridge MA, (to appear).

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

A. Sloman and B.S. Logan. Evolvable architectures for human-like minds. In G. Hatano, N. Okada, and H. Tanabe, editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam, 2000. ISBN 0-444-50418-4.

L. Weiskrantz. *Consciousness Lost and Found*. Oxford University Press, New York, Oxford, 1997.

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

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