What Are Emotion Theories About?

Aaron Sloman
School of Computer Science, The University of Birmingham
Birmingham, B29 7LF, UK
A.Sloman@cs.bham.ac.uk

Abstract

This is a set of notes relating to an invited talk at the cross-disciplinary workshop on Architectures for Modeling Emotion at the AAAI Spring Symposium at Stanford University in March 2004. The organisers of the workshop note that work on emotions "is often carried out in an *ad hoc* manner", and hope to remedy this by focusing on two themes (a) validation of emotion models and architectures, and (b) relevance of recent findings from affective neuroscience research. I shall focus mainly on (a), but in a manner which, I hope is relevant to (b), by addressing the need for conceptual clarification to remove, or at least reduce, the ad-hocery, both in modelling and in empirical research. In particular I try to show how a design-based approach can provide an improved conceptual framework and sharpen empirical questions relating to the study of mind and brain. From this standpoint it turns out that what are normally called emotions are a somewhat fuzzy subset of a larger class of states and processes that can arise out of interactions between different mechanisms in an architecture. What exactly the architecture is will determine both the larger class and the subset, since different architectures support different classes of states and processes. In order to develop the design-based approach we need a good ontology for characterising varieties of architectures and the states and processes that can occur in them. At present this too is often a matter of much ad-hocery. We propose steps toward a remedy.

Validation *vs* testing

It is good to ask how theories can be validated, though I would rather ask how they can be *tested*, and how they can be *compared*, in various dimensions, such as depth, clarity, generality, precision, explanatory power, etc., since most theories are incomplete, provisional, premature, vague, or just false. So validation is rarely to be expected, even when a theory is the best one available and provides a good basis for further research, a point that is familiar from the writings of Karl Popper (Popper 1934; Magee 1985), and work of Lakatos cited by Dean Petters in this symposium.

There is, however, a kind of validation of theories of a type Popper did not (as far as I know) admit as science, and many scientists do not acknowledge, partly because of Popper's influence, namely theories about what is possible (what can occur). Simple logic shows that even a *single* example validates such a theory. The importance of theories of what is possible and how they are presupposed by the search for laws and regularities which constrain what is possible, was discussed in Sloman (1978, Ch 2). See also (Boden 1990). In particular, contrary to wide-spread views on scientific method, the truth of a statement that something can occur is established by a *single* example – which could be world-shaking (e.g. the discovery that light of low intensity and small wavelength can trigger an electric current when high intensity large wavelength light fails (the photo-electric effect), the discovery in 1919 that light from a star can be deflected by passing near the sun, or Newton's earlier discovery that white light passing through a prism can be split into beams different colours). However, checking that the description of the example is correct may be non-trivial, especially if it requires the use of concepts that are not yet generally understood or theories that are not yet accepted.

For present purposes the study of architectures and the phenomena they support is part of the study of what *can* exist or occur (deep science) and why, rather than an attempt to discover any new laws about *what happens when* or *what correlates with what* (shallow science). But we need to know what we are talking about.

What are we talking about?

It is sometimes forgotten that the question whether a theory is true or false presupposes an answer to whether it makes sense at all. All theories use *concepts*, for these are the building blocks of propositions, but insofar as the concepts are obscure, confused, or vague, the theories, and even the questions to which the theories are answers, will be flawed. For instance, if our concept of 'emotion' is ambiguous then so also will many questions about emotions be ambiguous, e.g. how emotions evolved, what their functions are, which animals have them, which brain mechanisms produce them, what types there are, whether they are needed for intelligence, whether a foetus has them, etc.

Alas, our everyday concept of 'emotion' mixes up motivations, attitudes, moods, and other affective states and

processes, and is therefore is too deeply flawed to be a useful component of scientific questions and theories for reasons recounted elsewhere.[1] But that does not prevent scientists assuming that these concepts can be used to formulate clear, unambiguous scientific questions or theories. For instance, sometimes people who argue that emotions are needed for intelligence are merely defending the truism that motivation is needed for action, and preferences are needed for selecting between options. However, not everyone would call a motive or preference, however important, an emotion.

Wishful thinking isn't science

Sometimes over-generalising the notion of 'emotion' is related to a desire to argue that emotions are important in ways not previously acknowledged, e.g. that they are a pre-requisite for intelligence. This can be wishful thinking or a trivial truism. If 'emotion' is construed so broadly that it covers all goals and preferences the claim that emotions are needed for intelligence is vacuous. On the other hand if it refers more narrowly to the sorts of processes in which one subsystem interferes with or disrupts the normal functioning of another, as happens in many of the states in which people are described as being 'emotional' then it is false that emotions are *required* for intelligence: on the contrary, emotions of that sort can get in the way of sensible decisions and actions.

Monica Croucher and I once argued in 1981 that mechanisms required for intelligence in *resource-limited* robots in fast-changing environments would inevitably produce the *possibility* of emotional states, involving interruption or modulation of one process by another (where the latter is often a fast and therefore relatively stupid process) that detects an urgent need for some change, e.g. using reactive mechanisms labelled 'alarms' in Fig. 1. But saying that states of type X can occur as a side-effect of the operation of some mechanism M that is required for intelligence does not imply that states of type X are themselves required for intelligence. Many wishful thinkers misinterpreted that paper as claiming that emotions are *required* for intelligence, just as they fail to see the obvious flaw in Damasio's widely quoted reasoning (1994) from the premiss: *Damage to frontal lobes impairs both intelligence and emotional capabilities* to the conclusion *Emotions are required for intelligence*. A moment's thought should show that two capabilities could presuppose some common mechanisms without either capability being required for the other.

A research community with too much wishful thinking does not advance science. Of course, if I have misread such people and they are merely proposing the truism (noted by Hume, which needs no empirical evidence) that motivation and preferences are required for intelligent thought and actions then that is another manifestation of the ambiguity of the word 'emotion'.

[1] E.g. in (Sloman 2002b; 2002a; 2001; Sloman, Chrisley, & Scheutz To Appear).

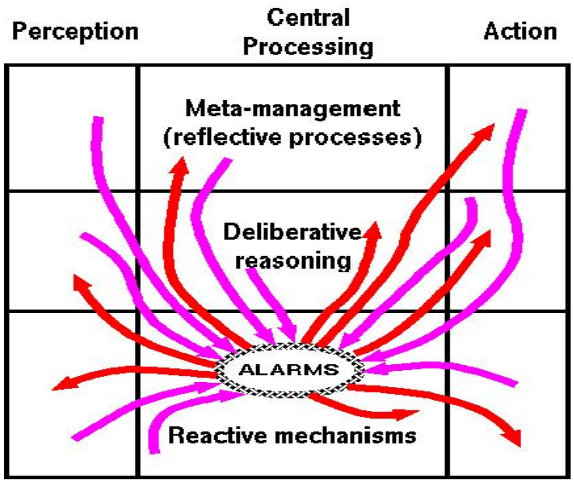

Figure 1: *The CogAff schema defines a crude first-draft division of mechanisms into 9 categories. A particular type of fast, pattern-driven, reactive central mechanism, with inputs from and outputs to many other components of the architecture could function as an alarm mechanism able to produce global reorganisation very quickly. Different sorts of alarm systems can produce different sorts of emotions, depending on where the inputs come from where the outputs go, what kinds of decisions are taken and what kinds of output signals are sent. Slower, longer lasting, more easily suppressed* dispositional *mechanisms can produce long term emotions, such as grief or jealousy.*

We need finer-grained ontologies

We should not put both (i) a general preference for saving effort, and (ii) fear produced by a stampeding herd, in the same conceptual basket when they have so very many differences, including the (relative) permanence of the first and the transience of the second. Or rather, we can put them in a more general basket labelled 'affect' which includes sub-categories which might be aptly labelled 'emotion', 'desire', 'preference', 'attitude', 'value', 'mood', etc.

I am not claiming that all emotions are short-term (though many emotion theories have that consequence, often not noticed by their proponents). In Wright *et al.* (1996) we tried to show, at least in outline, how long-term emotions such as grief, could exist in the form of dispositions which only rarely manifest themselves, either because of external triggers (reminders) or because other competing attention-grabbers subside. Many states generally regarded as important human emotions that form the stuff of plays, novels and garden-gate gossip are long term largely dispositional states, including jealousy, vengefulness, family love, obsessive ambition, infatuation, fascination with a mathematical problem, etc. There are other long term affective states such as preferences, and attitudes that are not normally called emotions.

Of course, someone who defines an emotion as an episodic state in which there are particular sorts of bodily changes or sensed changes in a body state map will not include some of these long term states as emotions. But that's just another example of the terminological disarray.

Non-vicious cycles of defining relationships

We cannot avoid the problems by defining 'emotion' in some clear, simple formula – for deep theoretical concepts must at least partly be defined implicitly by mutual relationships with other concepts in a good theory, as concepts like 'force', 'mass', 'acceleration', 'energy', 'momentum' and the like are largely defined by their mutual relationships in a physical theory, and change when the theory changes (e.g. with the advent of relativity theory).

But that suggests a vicious circle: we cannot articulate clear, precise theories without using clear, precise concepts and we cannot specify concepts clearly and precisely except in the context of a theory.

The circle is not vicious if we accept that there is a gradual process of increasing clarity and depth: as we extend our theories we make our concepts more precise, and sometimes this includes relating the topics that interest us to concepts used in a deeper, more general theory, as the theory of the atomic structure of atoms and the behaviour of sub-atomic particles is deeper and more general than the theory of chemical compounds and how they behave.

In particular, there is a way of extending and refining our confused and indeterminate pre-scientific concepts of mental phenomena in a systematic and productive manner, if we can provide a deep theory of an architecture that generates a space of possible mental phenomena, as happened in physics and chemistry during the last two centuries when new theories of the architecture of matter provided a basis for developing improved concepts of types of matter and concepts of types of processes involving them.

The need for a broad view

In mathematics facts about integers could not be fully understood until they were studied first as special cases of reals, and then of complex numbers. This is a general point. Consider a research chemist who investigated only a particular subset of molecules (e.g. oxides) without investigating the broader class including components of oxides, molecules containing oxides, and studying interactions between molecules that can produce oxides and molecules with which oxides can interact. Such a chemist would be missing important facts about oxides.

Likewise research on a specific sub-class of affective phenomena (e.g. emotions) which is not treated as part of a more general and varied class of *affective* states and processes along with the cognitive processes that produce, interact with, or are part of them would be missing important facts about the sub-class. I am not saying that everyone has to study everything, but that study of a part of an integrated functioning system needs to take account of the whole of which it is a part.

So, although many of the terms used to characterise the field, including 'affect' and 'emotion' are riddled with ambiguity and unclarity which limits their usefulness for the purposes of science or engineering, as argued in the papers previously cited, we can nevertheless "triangulate" onto more precise meanings by using a broader context, and considering:

- the variety of architectures in which emotions and other affective states can occur,
- the mechanisms that can produce them,
- the mechanisms with which they can interact,
- their substructures,
- how they develop and decay (Sloman 1982).

This requires us to consider complete systems (agents), as well as the environments in which they evolve or develop, and the other kinds of agents with which they interact. Thinking about a complete system requires thinking about its *architecture* in which different kinds of mechanisms concurrently perform different tasks, use different forms of representation, learn in different ways, and interact with one another. That's almost a truism, but how do we know whether we have considered a rich enough architecture? And how can we describe what components are available in a manner that is not question-begging, as architectures with boxes labelled 'emotion' may be?

Varieties of architectural nomenclature

Unfortunately, although there has been much discussion of architectures in the last two decades there is no agreed terminology for describing them in a manner that is neutral between different architectures, as the language of sub-atomic particles is common to the description of atoms and molecules of different structures.

Even when people studying architectures use the same labels (e.g. 'reactive', 'deliberative', 'reflective'), they don't necessarily understand them the same way. For instance, some may be inclined to describe any component that can evaluate two possibilities for action and then select one as 'deliberative', whereas others would require that to be done in the context of a mechanism that can consider, evaluate, and select between alternative multi-step actions, using a formalism with compositional semantics. The latter is what I thought 'deliberative' meant, until I met people using the label for selection processes in a reactive mechanism e.g. a neural net using a 'winner-takes-all' decision strategy. I have tended to label the latter 'proto-deliberative'.

This is an example of a general problem: if people do not all notice the same range of intermediate possibilities they may inadvertently draw boundaries in different places, and end up talking at cross-purposes. (I am not saying that there are 'correct' places to draw boundaries, only that agreement on conceptual boundaries is needed for fruitful communication, but often lacking.)

Towards an ontology for mental architectures

In principle, architectures can be specified in terms of: (a) types of components, labelled using agreed terminology, (b) how those components are connected (e.g. how control information and other sorts of information flow between the components), (c) the forms of representation they use, and (d) their functions.

The sorts of components can be classified according to various subdivisions, one of which is whether their primary function is perception, action, or some kind of central processing. (See Nilsson's discussion of architectures

in his 1998 book). Each of those categories can be further subdivided according to whether the processes are primarily concerned with the *external* environment or with observation, evaluation and control of *internal* processes, e.g. keeping records of percepts, or planning decisions or actions selected. The former involve *outward-oriented* and the latter *inward-oriented* processes and mechanisms.

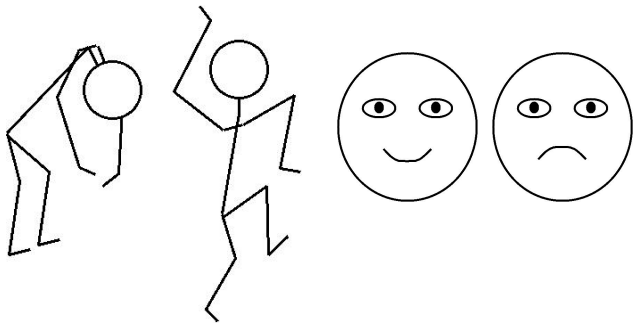

Figure 2: *Multi-layer perception: some high level perceptual mechanisms, linked to central mechanisms using an ontology of information states, may be dedicated to detection of affective states. Contrast perceiving 2-D and 3-D structure and motion, recognizing types of physical objects, identifying individual entities.*

We can also distinguish the ontologies and formalisms used in different perceptual, motor, or central sub-systems. For instance sophisticated perception and action systems may be layered: the lowest layers are concerned with low-level sensory or motor details and higher levels with different sorts of abstractions used in different ways. E.g. different mechanisms and ontologies are needed for, and different purposes served by, detecting a flash, measuring optical flow, seeing meaningful words on a page, seeing, a face, seeing happiness in a face or gesture, seeing danger in a rotten tree, seeing something as graspable. Likewise contracting a muscle, hitting someone, exploring a cavity with a stick, greeting a friend and parking a car require different sorts of action mechanisms. Fig. 1 crudely indicates different levels of abstraction in perception and action mechanisms, and Fig. 2 indicates some percepts possibly linked to a central meta-management layer.[2]

The forms in which information, including control information, can be expressed differ from one component to another. In primitive mechanisms, only the current state of sensory mechanisms expresses what is sensed, whereas in more sophisticated systems, enduring internal records can survive their causes, allowing more complex and indirect relations between sensing and acting. Likewise, in primitive mechanisms, goals, drives, needs, preferences, etc. are expressed only implicitly in the current states of interacting sub-systems, whereas in more sophisticated cases they can be 'summarised' in explicit, enduring records that can enter into more varied interactions with other kinds of information across different time-scales.

Architectures may vary according to whether and how they develop over time: some architectures do not change, whereas the architecture of a newborn human infant does not include many of the capabilities of a normal adult, so human architectures must develop.

Some of the forms of information processing evolved many millions of years ago and can be found in all animals, including microbes and insects. Others require very specific and biologically costly machinery and are found only in a relatively small number of large-brained species of animals.

All organisms use some information expressed only as scalar values, or vectors of scalar values, whereas a subset can in addition cope with information items that have rich structures of variable depth and complexity, such as plans, parse trees, or percepts of kinds of machinery. The thoughts you have as you read this document vary in complexity, structure, internal and external relationships. Much of early AI focused on formalisms supporting structural variability, as that appears to be central to some peculiarly human capabilities (e.g. mathematical reasoning, language understanding, planning), but we can now see that other forms of representation are also needed.

It's an unfinished ontology

The CogAff schema depicted in Fig 1 gives only a very crude indication of the variety of possible architectures. There's still work to be done to produce an adequate ontology. This will lead to a very much richer schema representing a wider range of possibilities in more detail.

When we have done that we can analyse the different sorts of states and processes (short term and long term) that are possible in different instances of the schema. For instance, which of the various forms of information representation and processing are available in a sub-system will determine what kinds of decisions can be taken, what kinds of conflicts can arise and what mechanisms are available for dealing with them and also the time-scales over which causal relations can operate.

Not every species has the kind of information-processing capability that supports a long term, often temporarily dormant, desire for revenge against a particular individual for a particular action performed by that individual.

When architectures include multiple concurrently active mechanisms with different capabilities there may be complex mixtures of states and processes, some of them evolutionarily very old and shared with many other species and some dependent on relatively rare and specialised capabilities. A human's fear produced by an approaching wildebeest stampede can have complexities not available to a frightened rabbit in the same place.

A deliberative mechanism able to use varying structures with compositional semantics to represent possible actions will provide capabilities lacking in a purely reactive proto-deliberative system that merely supports competing activation patterns in a winner-takes all neural net. However, such differences in internal processing may be obscured by similarities in external behaviour: for a suitably evolved

[2]Compare neural nets etc. trained to label images as 'happy', 'sad', etc. for such systems have no idea what happiness and sadness are: they are mere labels not linked to a useful ontology, even if wishful thinkers consider them to be emotion recognisers.

or trained non-deliberative system can produce behaviours achieved more creatively in a deliberative system requiring far less storage for special cases.

Kinds of affect in kinds of minds

A theory of possible architectures for natural and artificial behaving systems provides a basis for developing an ontology for the kinds of mental states and processes supported by each architecture. We can do this in a systematic, generative way – recommended in (Sloman, Chrisley, & Scheutz To Appear) – e.g. by considering different combinations of mechanisms in the different 'boxes' in the CogAff schema and different ways of connecting them, then work out what sorts of states and processes can arise in the resulting system.[3]

The results of this 'generative' approach can be used as a collection of hypotheses that can drive research to investigate the capabilities of different sorts of animals (including humans) both to investigate the architectures they have and to investigate the kinds of states and processes that can occur in them. This is likely, in turn to reveal gaps and errors in our theories about possible mechanisms and the architectures that can contain them.

An example of this kind of mechanism-based analysis of possibilities is the well-known book (Ortony, Clore, & Collins 1988), though it focuses mainly on the capabilities of a fairly sophisticated deliberative system producing mental states with rich semantic contents, and is as much about attitudes as about emotions.

In the long run, the kind of theory-driven (some would say bottom-up) systematically generated overview of varieties of minds and what they can do (something like the periodic table of chemical elements, but very much more complex) could be a major achievement of science, to be tested by the variety of phenomena it is able to explain and the precision with which it explains them (e.g. how many distinct kinds of fear or learning it accounts for).

Using conceptual analysis top down

Instead of simply striving for that, we can also start from what we already know about the considerable variety of affective and other states and processes, and do conceptual analysis (Ch. 4 of Sloman 1978). Affective phenomena already known to common sense include diverse familiar states and processes referred to in colloquial words and phrases: attitudes, emotions, moods, desires, preferences, inclinations, interests, hopes, ambitions, obsessions, values, standards, ideals, etc., and new ones found in counselling rooms, clinical laboratories (e.g. studies of effects of various kinds of drugs or brain damage), and also anthropological studies which show how some of the possibilities are culture-dependent, even when the underlying architecture is the same.

[3]Where that is too difficult to establish analytically, it can be done by running simulations. Matthias Scheutz has been doing this with relatively simple affective and proto-deliberative agents, e.g. (Scheutz & Sloman 2001; Scheutz & Schermerhorn 2002).

Our general notion of 'affect' and our special case colloquial concepts all suffer from vagueness and confusion because our ordinary language did not evolve to meet the purposes and standards of science and engineering, but to serve the multifarious purposes of everyday life, including gossiping, collaboration, squabbling, educating, story-telling and generally living with others. However, if we attempt to relate our analysis of these informal colloquial concepts to the ontologies generated systematically on the basis of architectural theories, we soon find the need to subdivide the familiar notions into more precise categories, as well as perhaps introducing new broad-distinctions (e.g. between emotions that do and do not involve the deliberative or meta-management architectural layers).

So research can go both ways: thinking about and running examples of an architecture can shed new light on old concepts, and investigating architectural requirements for some well known but complex human phenomena, e.g. long term grief, schadenfreude, infatuation, or political ambition can help to determine requirements for adequate architectural theories. (Much of my own research into architectures over the last 30 years has taken the latter form, along with using results from psychology and neuroscience, especially clinical studies.)

Exploratory development of working systems often produces illuminating surprises: in the process of doing that we discover important gaps in our thinking about what is and is not possible, which is why in AI generally the nature of the problems investigated cannot be understood when the investigation starts. Only when solutions have been found can we see what problems they solve. Only when we have many good working models can we understand what they are models of, partly because we'll then learn that they are models of different things that we had not previously distinguished.

Linking the research to brain science can also yield surprises. If a particular virtual machine architecture that is implementable on computers could not possibly be implemented on brains then it cannot be the architecture used by humans or any other animals: and we need to think again. Alternatively we might find that we have not understood brain mechanisms.

Brain science can also draw attention to types of mechanisms that we might never have thought of, which can, in turn lead us to consider types of architectures we might not otherwise have thought of (e.g. complex networks of neural networks some monitoring or modulating others, with some global quantitative and qualitative control by chemical mechanisms).

Varieties of biological minds

Different architectures will support different classes of possible states and processes. If the architectures of human infants, toddlers and adults are different, then different sets of concepts may be required for describing them. Even more different will be the range of possible affective states in insects and other animals very different from humans. Neither human infants or insects have the ability to be obsessively politically ambitious: the

architectural requirements, the conceptual requirements, and the knowledge of the world required for such a state are beyond their reach. Why can't a goldfish long for its mother?

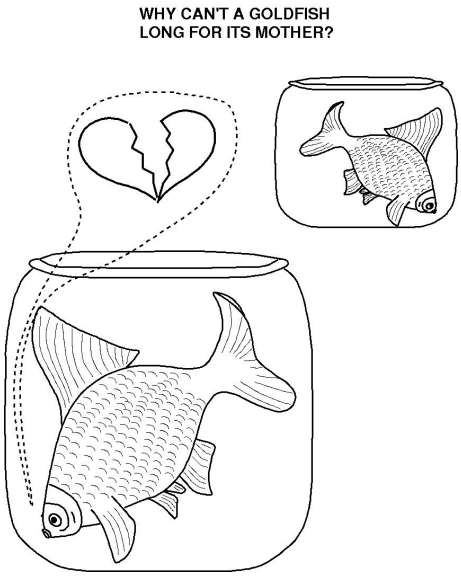

Figure 3: *Does the inability of a fish to long for its mother depend on missing behavioural capabilities, or on something deeper?*

What about fear? If a fly escapes my attempt to swat it, or a woodlouse touched with a finger suddenly rolls up in a ball, are their behaviours produced by emotional responses? Are those examples of fear? It is not clear that our ordinary concepts are sufficiently definite for such questions to have answers. Similar indeterminacy afflicts questions about whether an unborn infant feels pain when it responds to a prod.

Architecture-neutral affective concepts?

This raises the question whether we can go beyond colloquial language and define schematic notions of 'affect', 'emotion', 'mood' etc. that transcend architectural differences? Perhaps, but only at a very general level. Even if it is possible, this does not support the common notion of 'basic emotion' as might be thought (see the critique by (Turner & Ortony 1992)). For instance in my forthcoming paper with Chrisley and Scheutz we try to distinguish affective and non-affective states, and then positive and negative affect in an architecture-neutral way.

There may be an architecture-neutral notion of 'fear' as a state in which mechanisms detect a need to avoid, escape, or prevent something harmful that might happen, but the actual manifestations of that type of state in different systems will be so different (e.g. in insects, rabbits and humans) that to use a notion of *basic emotion* for this sort of thing is totally misleading if it suggests that all occurrences of fear involve some common mechanism as all occurrences of oxides include oxygen atoms.

All this can be expected to undermine many cherished assumptions, such as: the assumptions that we all understand what we mean by 'emotion', that there are 'basic' emotions, that emotions are needed for intelligence, that emotions are classifiable on the basis of their external manifestations, to name a few. Incidentally, those are not assumptions of common sense or colloquial language, but the inventions of scientists trying to fit phenomena into the prevalent scientific methodology.

Starting again from an architecture-based theory may upset some scientists, but may also reveal the true depths of some of our intuitive, pre-scientific theories of mind, such as those used by novelists, playwrights, poets and gossips.

Where next?

The approach to the study of affect (including emotions) discussed here is clearly a long term affair, with the end, if there is an end, nowhere in sight. But if people at least try to work with a common framework for talking about architectures and the states they support, it may be possible for more collaboration and more critical comparative analysis to accelerate progress.

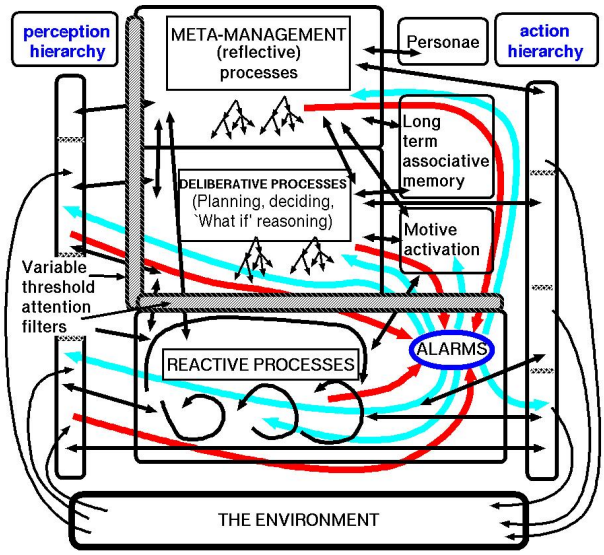

Figure 4: *The H-Cogaff architecture, depicted only in crude outline here, is a special case of the CogAff schema in which all the 'boxes' have many mechanisms performing different sorts of tasks concurrently, with complex interactions between them. At least three different classes of emotions (primary, secondary and tertiary) can arise in this sort of architecture. There are also many kinds of learning, decision making, perception, action, self-understanding, communication, and a wide variety of types of affect, not yet fully documented. See also Dean Petters' paper.*

In this spirit we invite criticisms of both the CogAff schema and the H-Cogaff architecture which is a special, rather complex, case of the general CogAff schema, reflecting many aspects of human minds (hence the 'H'). It has inspired work on automatic intrusion detection (Kennedy & Sloman 2003), intelligent game agents (Hawes 2002; 2003), forensic text analysis (Kotov 2004), and modelling of attachment in infants reported in this workshop. It overlaps considerably with the architecture

presented in Minsky's Emotion Machine (see (Minsky, Singh, & Sloman 2004 Forthcoming)). The architecture is crudely depicted in Fig 4 and discussed in papers available on the Birmingham Cognition and Affect web site. But it is clearly too simple to be taken seriously as a model of human (child or adult) minds and much more work is needed.

Acknowledgements

I have benefited from interaction with many people over the last 30 years or so, including colleagues at Sussex and Birmingham Universities, especially those who worked in the Cognition and Affect project, some of them funded by a grant from the Leverhulme Trust between 1999 and 2003. There are close connections with work on 'The Emotion Machine' by Marvin Minsky. Many seeds for this work can be found in Herb Simon's pioneering paper written over three decades ago.

References

Boden, M. A. 1990. *The Creative Mind: Myths and Mechanisms*. London: Weidenfeld & Nicolson.

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

Hawes, N. 2002. An anytime planning agent for computer game worlds. In *Workshop on agents in computer games, 3rd Int. Conf. on Computers and Games (CG'02)*, 1–14.

Hawes, N. 2003. *Anytime deliberation for computer game agents. PhD Thesis*. Ph.D. Dissertation, School of Computer Science, University of Birmingham. Available at http://www.cs.bham.ac.uk/research/cogaff/.

Kennedy, C. M., and Sloman, A. 2003. Autonomous recovery from hostile code insertion using distributed reflection. *Journal of Cognitive Systems Research* 4(2):89–117.

Kotov, A. 2004. D-scripts model for speech influence and emotional dialogue simulation. In Lee, M., ed., *Proceedings 7th Annual CLUK Research Colloquium*. (Available http://www.cs.bham.ac.uk/~mgl/cluk/papers/kotov.pdf).

Magee, B. 1985. *Popper*. London: Fontana (Modern Masters Series).

Minsky, M.; Singh, P.; and Sloman, A. 2004 Forthcoming. The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine*. http://web.media.mit.edu/~push/StThomas-AIMag.pdf.

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

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

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

Scheutz, M., and Schermerhorn, P. 2002. Steps towards a systematic investigation of possible evolutionary trajectories from reactive to deliberative control systems. In Standish, R., ed., *Proceedings of the 8th Conference of Artificial Life*. MIT Press.

Scheutz, M., and Sloman, A. 2001. Affect and agent control: Experiments with simple affective states. In Ning Zhong, *et al*., ed., *Intelligent Agent Technology: Research and Development*. New Jersey: World Scientific Publisher. 200–209.

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

Sloman, A., and Chrisley, R. 2003. Virtual machines and consciousness. *Journal of Consciousness Studies* 10(4-5):113–172.

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

Sloman, A.; Chrisley, R.; and Scheutz, M. To Appear. The Architectural Basis of Affective States and Processes. In Arbib, M., and Fellous, J.-M., eds., *Who Needs Emotions?: The Brain Meets the Machine*. Oxford, New York: Oxford University Press. Online at http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf.

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

Sloman, A. 1982. Towards a grammar of emotions. *New Universities Quarterly* 36(3):230–238. (http://www.cs.bham.ac.uk/research/cogaff/0-INDEX96-99.html#47).

Sloman, A. 2001. Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science* 2(1):177–198.

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

Sloman, A. 2002b. How many separately evolved emotional beasties live within us? In Trappl, R.; Petta, P.; and Payr, S., eds., *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press. 35–114.

Turner, T., and Ortony, A. 1992. Basic Emotions: Can Conflicting Criteria Converge? *Psychological Review* 99:566–571. 3.

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.