The Architectural Basis of Affective States and Processes

Aaron Sloman[*]**, Ron Chrisley**[+] **and Matthias Scheutz**[#]
∗ School of Computer Science, The University of Birmingham, UK
+ Centre for Cognitive Science, The University of Sussex, UK
# Department of Computer Science and Engineering,
University of Notre Dame, USA

A.Sloman@cs.bham.ac.uk
R.L.Chrisley@cogs.susx.ac.uk
Matthias.Scheutz.1@nd.edu

December 14, 2003

Paper for inclusion in Fellous and Arbib (eds),
***Who Needs Emotions?: The Brain Meets the Machine*, Oxford University Press.**
Short title: Architectural basis of affective states & processes
Author to be contacted for correspondence:
Aaron Sloman
School of Computer Science
University of Birmingham
Birmingham, B15 2TT
United Kingdom
Email: A.Sloman@cs.bham.ac.uk

NOTE: a version posted here on 2nd December had to be truncated to meet the publisher's constraints. This is the shortened version. The original is still available as

http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions-long.pdf

Abstract

Much discussion of emotions and related topics is riddled with confusion because different authors use the key expressions with different meanings. Some confuse the concept of "emotion" with the more general concept of "affect", which covers other things besides emotions, including moods, attitudes, desires, preferences, intentions, dislikes, etc. Moreover researchers have different goals: some are concerned with understanding natural phenomena, while others are more concerned with producing useful artifacts, e.g. synthetic entertainment agents, sympathetic machine interfaces, and the like. We address this confusion by showing how "architecture-based" concepts can extend and refine our pre-theoretical concepts in ways that make them more useful both for expressing scientific questions and theories, and for specifying engineering objectives. An implication is that different information-processing architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and so on. We start with high level concepts applicable to a wide variety of types of natural and artificial systems, including very simple organisms, namely concepts such as "need", "function", "information-user", "affect", "information-processing architecture". For more complex architectures, we offer the CogAff schema as a generic framework which distinguishes types of components that may be in a architecture, operating concurrently with different functional roles. We also sketch H-Cogaff, a richly-featured special case of CogAff, conjectured as a type of architecture that can explain or replicate human mental phenomena. We show how the concepts that are definable in terms of such architectures can clarify and enrich research on human emotions. If successful for the purposes of science and philosophy the architecture is also likely to be useful for engineering purposes, though many engineering goals can be achieved using shallow concepts and shallow theories, e.g., producing "believable" agents for computer entertainments. The more human-like robot emotions will emerge, as they do in humans, from the interactions of many mechanisms serving different purposes, not from a particular, dedicated "emotion mechanism".

Contents

1 INTRODUCTION

Many confusions and ambiguities bedevil discussions of emotions. As a way out of this, we attempt to present a view of mental phenomena in general, and the various sorts of things called "emotions" in particular, as states and processes in an information-processing architecture. Emotions are a subset of *affective* states. Since different animals and machines can have different sorts of architectures capable of supporting different varieties of states and processes, there will be different families of such concepts, depending on what the architecture is. For instance if human infants, cats, or robots, lack the sort of architecture presupposed by certain classes of states (such as obsessive ambition, or being proud of one's family), then they cannot be in those states. So the question whether an organism or a robot needs emotions, or needs emotions of a certain type, reduces to the question of what sort of information-processing architecture it has and what needs arise within such an architecture.

1.1 Needs, functions and functional states

The general notion of X having a need does not presuppose a notion of goal or purpose, but merely refers to necessary conditions for the truth of some statement about X, $P(X)$. E.g. in trivial cases $P(X)$ could be "X continues to exist", and in less trivial cases something like "X grows, reproduces, avoids or repairs damage," etc. I.e., all needs are relative to whatever they are necessary for. Some needs are *indirect* insofar as they are necessary for something else that is needed for some condition to hold. A need may also be relative to a context, since Y may be necessary for $P(X)$ only in some contexts. So X *needs* Y is elliptical for something like: *There is a context C, and there is a possible state of affairs $P(X)$, such that in context C, Y is necessary for $P(X)$.* Such statements of need are actually shorthand for a complex collection of counterfactual conditional statements about "'what would happen if..."

Parts of a system have a *function* in that system if their existence helps to serve the needs of the system, under some conditions. In those conditions the parts with functions are *sufficient*, or *part of a sufficient condition* for the need to be met. Suppose X has a need N, in conditions of type C (i.e. there is a predicate P such that in conditions of type C, N is necessary for $P(X)$). And suppose that O is an organ, component, or state, or sub-process of X. We can use $F(O, X, C, N)$ as an abbreviation for *"In contexts of type C, O has the function F of meeting X's need N (i.e. the function of producing satisfaction of that necessary condition for $P(X)$)"*. This actually states:

> *In contexts of type C the existence of O, in the presence of the rest of X, tends to bring about states meeting the need N or tends to preserve such states if they already exist, or tends to prevent things that would otherwise prevent or terminate such states.*

Where sufficiency is not achievable, a weaker way of serving the need is to make the necessary condition *more likely* to be true. This analysis rebuts arguments (e.g. (Millikan, 1984)) that the notion of function has to be explicated in terms of evolutionary or any other history, since the causal relationships summarised above suffice to support the notion of function, independently of how the mechanism was produced.

We call a state in which something is performing its function of serving a need, a *functional* state. Later we'll distinguish *desire-like, belief-like* and other sorts of functional states (Sloman, 1993). The label 'affective' as generally understood seems to be very close to this notion of a desire-like state, and subsumes a wide variety of more specific types of affective states, including the subset we'll define as 'emotional'.

Being able to serve a function by producing different behaviours in the face of a variety of threats and opportunities minimally requires (a) sensors to detect when the need arises, if it is not a constant need, and (b) sensors to identify aspects of the context which determine what should be done to meet the need — for instance, in which direction to move, or which object to avoid, and (c) action mechanisms that combine the information from the sensors and deploy energy so as to meet the need. In describing components of a system as sensors or selection mechanisms we are ascribing to them functions analysable as complex dispositional properties depending on what would happen in various circumstances.

Combinations of the sensor states trigger or modulate activation of need-supporting capabilities. There may in some systems be conflicts and conflict-resolution mechanisms (e.g. using weights, thresholds, etc.). Later we'll see how the processes generated by sensor states may be purely reactive in some cases, and in other cases deliberative, i.e. mediated by a mechanism that represents possible sequences of actions, compares them, evaluates them and makes selections on that basis before executing the actions.

We can distinguish sensors that act as *need-sensors* those that act as *fact-sensors*. Need-sensors have the function of initiating action, or tending to initiate action (in contexts where something else happens to get higher priority), to address a need, whereas fact-sensors do not, though they can modify the effects of need sensors. For most animals, merely sensing the fact of an apple on a tree would not in itself initiate any action relating to the apple. On the other hand, if a need for food has been sensed, then that will (unless overridden by another need) initiate a process of seeking and consuming food. In that case the factual information about the apple could influence which food is found and consumed.

The very same fact-sensor detecting the very same apple could also modify a process initiated by a need to deter a predator – in that case, the apple could be selected for throwing at the predator. In this case we can say that the sensing of the apple has no motivational role. It is a "belief-like" state, not a "desire-like" state.

1.2 Information-processing architectures

The *information-processing architecture* of an organism or other object is the collection of information-processing mechanisms which together enable it to perform in such a way as to meet its needs (or, in "derivative" cases, *could* enable it to meet the needs of some larger system containing it).

Describing an architecture involves (recursively) describing the various parts and their relationships, including the ways in which they cooperate or interfere with one another. Systems for which there are such true collections of statements about what they would do to meet needs under various circumstances can be described as having *control-states*, of which the belief-like and desire-like states mentioned previously are examples. In a complex architecture there will be

many concurrently active and concurrently changing control states.

The components of an architecture need not be physical components: physical mechanisms may be used to implement *virtual machines* in which non-physical structures such as symbols, trees, graphs, attractors, information records, are constructed and manipulated. This idea of a virtual machine implemented in a physical machine is familiar in computing systems (e.g. running word-processors, compilers and operating systems) but is equally applicable to organisms which include things like information stores, concepts, skills, strategies, desires, plans, decisions, inferences, etc. that are not physical objects or processes but are *implemented* in physical mechanisms, such as brains.[1]

Information-processing virtual machines can vary in many dimensions, e.g. the number and variety of their components, whether they use discretely or continuously variable sub-states, whether they can cope with fixed or variable complexity in information structures (e.g. vectors of values *vs* parse trees), the number and variety of sensors and effectors, how closely internal states are coupled to external processes, whether processing is inherently serial or uses multiple concurrent, possibly asynchronous sub-systems, whether the architecture itself can change over time, whether the system builds itself or has to be assembled by an external machine (like computers and most current software), whether the system includes the ability to observe and evaluate its own virtual-machine processes or not (i.e. whether it includes "meta-management" as defined in (Beaudoin, 1994)), whether it has different needs or goals at different times, how conflicts are detected and resolved, and so on.

In particular, whereas the earliest organisms had sensors and effectors directly connected so that all behaviours were totally reactive and immediate, evolution 'discovered' that for some organisms, in some circumstances, there are advantages in having an *indirect* causal connection between sensed needs and the selections and actions that can be triggered to meet the needs: i.e. an intermediate state that 'represents' a need, and is capable of entering into a wider variety of types of information processing than simply triggering a response to the need.

Such intermediate states could (a) allow different sensors to contribute data for the same need, (b) allow multi-function sensors to be re-directed to gain new information relevant to the need (looking in a different direction to check that enemies really are approaching), (c) allow alternative responses to the same need to be compared, (d) allow conflicting needs to be evaluated, including needs that arise at different times, (e) allow actions to be postponed while the need is remembered, (f) allow associations between needs and ways of meeting them to be learnt and used, etc.

This seems to capture the notion of a system having *goals* as well as needs. Having a goal is *having an enduring representation of a need, namely a representation that can persist after sensor mechanisms are no longer recording the need, and which can enter into diverse processes attempting to meet the need*.

Evolution also produced organisms that in addition to having need-sensors also had fact-sensors that produced information that could be used for varieties of different needs, i.e. 'percepts' (closely tied to sensor states) and 'beliefs', which are indirectly produced and can endure beyond

[1]The attribute "virtual" here is in contrast to "physical", i.e., a running "virtual machine" is an abstract machine containing abstract components which may be capable of running on different physical machines. Virtual machine states can have causal powers, for instance the power to deliver email or to detect and prevent access violations.

sensor states that produce them.

The use of the intermediate states *explicitly* representing needs and sensed facts requires extra architectural complexity. It also provides opportunities for new kinds of functionality (Scheutz, 2001). For example, if need-representations and fact-representations can be separated from the existence of sensor states detecting needs and facts, it becomes possible for such representations to be *derived* from other things instead of being directly sensed. The derived ones can have the same causal powers, i.e. helping to activate need-serving capabilities. So we get derived desires and derived beliefs. However, all such derivation mechanisms can, in principle, be buggy (in relation to their original biological function), for instance allowing desires to be derived that if acted on serve no real needs and may even produce death, etc. as happens in many humans.

By specifying architectural features that can support states with the characteristics associated with concepts like "belief", "desire", "intention", we avoid the need for what Dennett calls 'the intentional stance' (Dennett, 1978), which is based on an assumption of rationality, as is Newell's 'knowledge level' (Newell, 1990). Rather we need only what Dennett calls 'the design stance', as explained in (Sloman, 2002). However, we lack a systematic overview of the space of relevant architectures. As we learn more about architectures produced by we are likely to discover that the architectures we have explored so far form but a tiny subset of what is possible.

We now try to show how we can make progress in removing, or at least reducing, conceptual confusions regarding emotions (and other mental phenomena) by paying attention to the diversity of architectures and making use of architecture-based concepts.

2 EMOTION AS A SPECIAL CASE OF AFFECT

2.1 A conceptual morass

Much discussion of emotions and related topics is riddled with confusion because the key words are used with different meanings by different authors, and some are used inconsistently by individuals. For instance, many researchers treat all forms of motivation, or all forms of evaluation, or all forms of reinforcing reward or punishment, as emotions. The current confusion is summarised aptly in (Delancey, 2002) [2]

> There probably is no scientifically appropriate class of things referred to by our term emotion. Such disparate phenomena – fear, guilt, shame, melancholy, and so on – are grouped under this term that it is dubious that they share anything but a family resemblance.

The phenomena are even more disparate than that suggests, for instance insofar as some people would describe an insect as having emotions, such as fear, anger, or being startled, whereas others deny the possibility. Worse still, when people disagree as to whether something does or does not have emotions (e.g. whether a foetus can suffer) they often disagree on what would count as evidence to settle the question. For instance, some, but not all, will take behavioural responses as determining the answer, others require certain neural mechanisms to have developed, some will

[2] There are many variants of this point in the emotions literature: Give a search engine : emotion + "natural kind". Oatley and Jenkins (1996) comment on the diversity of definitions of "emotion" in the psychology literature.

say it is merely a matter of degree and some claim that it is not a factual matter at all but a matter for ethical decision.

Despite all the often-documented conceptual unclarity, many researchers still assume that the word "emotion" refers to a generally understood and fairly precisely defined, collection of mechanisms, processes or states. For them, the question whether (some) robots should or could have emotions is a well-defined question. However, if there really is no clear, well-defined, widely understood, concept it is not worth attempting to answer the question until we have achieved more conceptual clarity.

Detailed analysis of pre-theoretical concepts can make progress using the methods of conceptual analysis explained in chapter 4 of (Sloman, 1978), based on (Austin, 1956)). However, that is not our main purpose.

Arguing about what emotions *really* are is pointless: "emotion" is a "cluster" concept (Sloman, 2002), which has some clear instances (e.g. violent anger) some clear non-instances (e.g. remembering a mathematical formula) and a host of indeterminate cases on which agreement cannot easily be reached. However, something all the various phenomena called emotions seem to have in common is membership of a more general category of phenomena that are often called "affective", e.g. desires, likes, dislikes, drives, preferences, pleasures, pains, values, ideals, attitudes, concerns, interests, moods, intentions, etc., the more enduring of which can be thought of as components of *personality* – as suggested in (Ortony, 2002) and in the chapter by Norman *et al.*.

Mental phenomena that would not be classified as affective include perceiving, learning, thinking, reasoning, wondering whether, noticing, remembering, imagining, planning, attending, selecting, acting, changing one's mind, stopping or altering an action, etc. We shall try to clarify this distinction, below.

It may be that many of the people who are interested in emotions are, unwittingly, interested in the more general phenomena of *affect* (Ortony, 2002). This would account for some of the over-general applications of the label "emotion".

2.2 Towards a useful ontology for a science of emotions

How can emotion concepts and other concepts of mind be identified for the purposes of science? Many different approaches have been tried. Some concentrate on externally observable expressions of emotion. Some combine externally observable eliciting conditions as well as expressions. Some of those who look at conditions and responses focus on physically describable phenomena, whereas others use the ontology of ordinary language which goes beyond the ontology of the physical sciences in describing both environment and behaviour (e.g. using the concepts *threat, opportunity, injury, escape, attack, prevent,* etc.) Some focus more on internal physiological processes, e.g. changes in muscular tension, blood pressure, hormones in the blood stream, etc. Some focus more on events in the central nervous system, e.g. whether some part of the limbic system is activated.

Many professional scientists use "shallow" specifications of emotions and other mental states defined in terms of correlations between stimuli and behaviors, because they adopt an out of date empiricist philosophy of science that does not acknowledge the role of theoretical concepts

going beyond observation. (For counters to this philosophy see (Lakatos, 1970) and chapter 2 of (Sloman, 1978)).

Diametrically opposed to this, some define "emotion" in terms of introspection-inspired descriptions of what it is like to have one (e.g. no(Sartre, 1939) Sartre (1939) claims that having an emotion is "seeing the world as magical"). Some novelists, e.g. (Lodge, 2002), think of emotions as defined primarily by the way they are expressed in thought processes, for instance, thoughts about what might happen, whether the consequences will be good or bad, how bad consequences may be prevented, whether fears, loves, jealousy, etc. will be revealed, and so on. Often these are taken to be thought processes that cannot be controlled.

Nobody knows exactly how pre-theoretical (folk-psychology) concepts of mind work. We conjecture that they are partly architecture-based concepts: people implicitly presuppose an information-processing architecture (incorporating percepts, desires, thoughts, beliefs, intentions, hopes, fears etc.) when they think about others, and they use concepts that are implicitly defined in terms of what can happen in that architecture. For purposes of scientific explanation those naive architectures need to be replaced with deeper and richer explanatory architectures, which will support more precisely defined concepts. If the naive architecture turns out to correspond to some aspects of the new architecture, this will explain how naive theories and concepts are useful precursors of deep scientific theories — as happens in most sciences.

2.3 A design-based ontology

We suggest that "emotion" is best regarded as an imprecise label for a subset of the more general class of *affective* states. We can use ideas in section 1.2 to generate architecture-based descriptions of the variety of states and processes that can occur in different sorts of natural and artificial systems. Then we can explore ways of carving up the possibilities in a manner that reflects our pre-theoretical folk-psychology constrained by the need to develop explanatory scientific theories.

For instance, we'll show how to distinguish affective states from other states. We shall also show how our methodology can deal with more detailed problems, for instance the question whether the distinction between emotion and motivation collapses in simple architectures (e.g., see the chapter by Norman *et al.*). E.g. we'll show that it does not collapse if emotions are defined in terms of one process interrupting or modulating the "normal" behaviour of another.

We'll also see that where agents (e.g. humans) have complex, hybrid information-processing architectures involving a variety of types of sub-architectures, they may be capable of having different sorts of emotions, percepts, desires, preferences, etc. according to which portions of the architecture are involved. For instance, processes in a reactive sub-system may be insect-like (e.g. being startled) while other processes (e.g. long-term grief and obsessive jealousy) go far beyond anything found in insects. This is why, in previous work, we have distinguished *primary*, *secondary*, and *tertiary* emotions,[3] on the basis of their architectural underpinnings: *primary* emotions (such as primitive forms of fear) reside in a reactive layer and do not require representational capacities of possible, but non-actual states of the world and hypothetical reasoning abilities, whereas *secondary* emotions (such as worry, i.e., fear about possible future

[3]Extending terminology used by (Damasio, 1994; Goleman, 1996; Picard, 1997).

events) intrinsically do. For this, they need a deliberative layer. What we call *tertiary emotions* (such as self-blame) need, in addition, a layer (which we call "meta-management"), which is able to monitor, observe, and to some extent oversee processing in the deliberative layer and other parts of the system. This division into three architectural layers is only a rough categorization as is the division into three sorts of emotion (we will elaborate more in section 4.3). Further sub-divisions are required to cover the full variety of human emotions, especially as emotions can change their character over time as they grow and subside (as explained in Sloman (1982)).[4]

This task involves specifying information-processing architectures that can support the types of mental states and processes under investigation. The catch is that different architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and different varieties of mental states in general, just as some computer operating system architectures support states like "thrashing" where more time is spent swapping and paging than doing useful work, whereas other architectures, do not, for instance if they do not include virtual memory or multi-processing mechanisms.

So in order to understand the full variety of types of emotions, we need to study not just human-like systems but alternative architectures, in order to explore the varieties of mental states they support. This includes attempting to understand the control architectures found in many animals and also the different stages in the development of human architectures from infancy onward. Some aspects of the architecture will also reflect evolutionary development (Sloman, 2000a; Scheutz and Sloman, 2001).

3 VARIETIES OF AFFECT

What are affective states and processes? We now attempt to explain the intuitive affective/non-affective distinction in a general way. Like "emotion", the concept "affect" lacks any generally agreed definition. We suggest that what is intended by this notion is best captured by our architecture-based notion of a *desire-like* state introduced in section 1.1, in contrast with *belief-like* and other types of non-affective state. Desire-like and belief-like states are defined below.

3.1 Varieties of control-states

Previously we introduced a notion of a control-state that has some sort function which may include preserving or preventing some state or process. An individual's being in such a state involves the truth of some collection of counterfactual conditional statements about what the individual would do in a variety of possible circumstances.

We have defined "desire-like" states as those which have the function of detecting needs so that the state can act as an *initiator* of action designed to produce changes or prevent changes in a manner that serves the need. This can be taken as a more precise version of the intuitive notion of "affective" state. These are states that involve dispositions to produce or prevent some (internal or external) occurrence related to a need. It is an old point - dating at least back to the philosopher

[4] A similar theory is presented in Minsky's draft book *The Emotion Machine* available online at his web site.

David Hume (1739) – that all action may be based on many beliefs and derivatively affective states, but must have some intrinsically affective component in its instigation. In our terminology, no matter how many beliefs, percepts, expectations, and reasoning skills a machine or organism has, that will not cause it to do one thing rather than another, or even to do anything at all, unless it also has at least one desire or desire-like state.

Another use of "affective" implies that something is being *experienced* as pleasant or unpleasant. We do not assume that connotation, partly because it can be introduced as a special case, and partly because we wish to use a general notion of affect (desire-like state) that is broad enough to cover organisms and machines that would not naturally be described as experiencing states as pleasant or unpleasant, and also to states and processes in humans that they are not conscious of. For instance, one can be jealous or infatuated without being conscious or aware of the jealousy or infatuation. Being conscious of one's jealousy, then, is a "higher order state" that requires the presence of another state, namely that of being jealous. In (Sloman and Chrisley, 2003) our approach is used to *explain* how some architectures support experiential states.

Some people use "cognitive" rather than "non-affective", but that is undesirable if it implies that affective states cannot have rich semantic content and involve beliefs, percepts, etc., as illustrated in the "apple" example in section 2. Cognitive mechanisms are required for many affective states and processes.

3.2 Affective *vs* non-affective (what to do *vs* how things are)

We can now introduce our definitions.

- A *desire-like* state D of a system S is one whose function it is to get S to do something to preserve or to change the state of the world – which could include part of S (in a particular way dependent on D). Examples include preferences, pleasures, pains, evaluations, attitudes, goals, intentions, and moods.

- A *belief-like* state B of a system S is one whose function is to provide information that could, in combination with one or more different sorts of desire-like states, enable the desire-like states to fulfil their functions. Examples include beliefs (particular and general), percepts, memories, and fact-sensor states.

Primitive sensors provide information about some aspect of the world simply because the information provided varies as the world changes. (Another example of sets of counterfactual conditional statements.) Insofar as the sensors meet the need of providing *correct* information they also serve a desire-like function, namely to "track the truth" so that the actions initiated by other desire-like states serving other needs can be appropriate to meeting those needs. In such cases, the state B will include mechanisms for checking and maintaining correctness of B: in which case there will be, as part of the mechanisms producing the belief-like state, sub-mechanisms whose operation amounts to the existence of another desire-like state, serving the need of keeping B true and accurate. In the case of a visual system this could include vergence control, focus control, and visual tracking.

In these cases B has a dual function, the primary belief-like function of providing information, and also a secondary desire-like function of ensuring that the system is in state B only when the

content of B actually holds (i.e., that the information expressed in B is correct and accurate.) The secondary function is a means to the first. Hence, what is often regarded as non-desire-like states can be seen as including a special subclass of desire-like states.

We are not assuming that these states have propositional content in the sense in which propositional content can be expressed as predicates applied to arguments, or expressed in natural language. On the contrary, an insect which has a desire-like state whose function is to get the insect to find food, need not have anything that could be described as a representation or encoding of "I need food". Likewise the percepts and beliefs (belief-like states) of an insect need not be expressible in terms of propositions. Similar comments could be made about desire-like and belief-like states in evolutionarily old parts of the human information processing architecture. Nevertheless the states should have a type of semantic content for which the notion of truth or correspondence with reality makes sense (Sloman, 1996).

In describing states as having functions we imply that their causal connections are to some extent reliable. However, this is consistent with their sometimes being suppressed or over-ridden by other states in a complex information processing system. For instance, although it is the function of a belief-like state to "track the truth", a particular belief may not be removed by a change in the environment if the change is not perceived, or if something prevents the significance of a perceived change being noticed. Likewise the desire to achieve something need not produce any process tending to bring about the achievement, if other stronger desires dominate, or if attention is switched to something else, or if an opportunity to achieve what is desired is not recognized, etc. So all of these notions have interpretations that depend heavily on complex collections of counterfactual conditionals being true: they are inherently *dispositional* concepts (see also the discussion of the belief-desire-intention models of teamwork in Tambe's article).

Our distinction is closely related to the old notion familiar to philosophers that desires and beliefs can both represent states of the world but they differ in the "direction of fit". When there is a mismatch, beliefs tend to get changed to produce a match (fit) and desires tend to cause something else in the world to be changed to produce or preserve a match, thus:

- A change in World *tends to cause* A change in Beliefs
- A change in Desires *tends to cause* A change in World

where "World" can include states of the organism.

Belief-like and desire-like states exhaust the variety of possible information states in *simple* organisms and machines, but in more sophisticated architectures there are sub-systems providing states that are neither desire-like nor belief-like. Examples include states in which possibilities are contemplated, but neither desired nor believed, for instance in planning, or in purposeless day-dreaming (*imagination-like* and *plan-like* states (Sloman, 1993)) or some kinds of artistic activities. Such activities have requirements that overlap with requirements for producing belief-like and desire-like states. E.g. they require possession of a collection of concepts and mechanisms for manipulating representations. Language considerably enhances such capabilities.

In other words, the evolution of sophisticated belief-like and desire-like states required the evolution of mechanisms whose power could also be harnessed for producing states that are neither. Such resources can then produce states that play a role in more complex affective states

and processes even though they are not themselves affective. For instance, the ability to generate a certain sort of supposition might trigger states that are desire-like (e.g. disgust or desire) or belief-like (e.g. being reminded of something previously known). What we refer to as secondary and tertiary emotions can also use such mechanisms.

3.3 Positive versus negative affect

There are many further distinctions that can be made among types of affective states. Among the class of affective (i.e., desire-like) states we can distinguish *positive* and *negative* cases, approximately definable as follows:

- A state N of a system S is a *negatively affective* state if being in N or moving towards being in N changes the dispositions of S so as to cause processes which *reduce* the likelihood of N persisting, or which tend to resist processes that bring N into existence.
- A state P of a system S is a *positively affective* state if being in P or moving towards being in P changes the dispositions of S so as to cause processes which *increase* the likelihood of P persisting, or which tend to produce or enhance processes that bring P into existence or maintain the existence of P.

For example, being in pain is negatively affective since it tends to produce actions that remove or reduce the pain. Enjoying eating an apple is positively affective since that involves being in a state which tends to prolong the eating and tends to resist things that would interfere with eating the apple. In both cases the effects of the states can be overridden by other factors, which is why the definitions have to be couched in terms of *dispositions* not actual effects. For instance, masochistic mechanisms can produce pain-seeking behaviour, and various kinds of religious indoctrination can cause states of pleasure to produce guilt-feelings that interfere with those states.

There are many subdivisions and special cases that would need to be discussed in a more complete analysis of information-processing systems with affective and non-affective states. In particular, various parts of the above definitions could be made more precise. We could also add further details such as defining intensity of an affective state, which might involve things like its ability to override or be overridden by other affective states and perhaps how many parts of the overall system it affects. Here, we mention only three important points.

We can distinguish *direct* and *mediated* belief-like and desire-like states. This amounts to a distinction between states without and with an *explicit* instantiation in some information structure that the system can create, inspect, modify, store, retrieve, remove. If the state is merely *implicit* (i.e. direct, unmediated) then the information state cannot be created or destroyed while leaving the rest of the system unchanged.

In other words, explicit mental states are instantiated in, but are not part of the underlying architecture (although they can be acquired and represented within it), whereas implicit mental states are simply states of the architecture which have certain effects. Note that "explicit" does not mean "conscious", as it is possible for a system to have explicit instantiations of an information structure without being aware of it (i.e., while the information structure is used by some process, there is no process that notices or records its presence).

Secondly, some belief-like states and desire-like states are *derivative* sub-states, in that they result from a process that uses something like premisses (i.e. pre-existing explicit/mediated states) and a derivation of a new explicitly represented state. Others are *non-derivative* sub-states because they are produced without any process of reasoning, or derivation of one representation from others, but merely arise out of activation of internal or external sensors and their effects on other sub-systems. Derivative states, as defined here, are necessarily also *explicit* (but not necessarily conscious). The derivative ones might also be described as "rational", and the derivative ones as "non-rational", insofar as the former but not the latter are produced by reasoning processes.

A third point concerns a causal connection between two states that does not include explicit reasoning, but something more like reinforcement learning. E.g., associative learning may bring it about that a certain kind of action A is the "content" of a desire-like state S, because state S is repeatedly followed by a previously desired state S'. Thus the state S in which A is desired arises because A has been found to be a means to S'. For instance, a rat can be trained to press a lever because that has been associated with acquiring food. This does not require the rat to have an explicit *belief* that pressing the lever causes food, from which it *infers* the result of pressing the lever. Having such a belief would support a different set of possible mental processes from the set supported by the mere learnt desirability of pressing the lever. For instance, the explicit belief could be used in making predictions as well as selecting actions.

Likewise a result of associative learning may be that a particular kind of sensory stimulation produces a belief-like state because the organism has learnt to associate the corresponding situations with those stimuli. For instance, instead of only the sound or smell of food producing the belief or expectation that food will appear, the perception of the lever going down could produce that belief.

In summary, we have distinguished merely *associatively triggered* belief-like and desire-like states from those that are derived by a process of *reasoning* making use of explicit representations rather than simply the causal consequences of implicit desire-like and belief-like states. The distinction between derivative and associative affective states will later be of assistance when attempting to distinguish between different kinds of emotions.

3.4 Positive and negative affect and learning

We have defined positive and negative affective states in terms of tendencies or dispositions to achieve/preserve (positive), or avoid/remove (negative) some state of affairs. It might be thought tempting to define affect in terms of the ability to produce learning, e.g. by defining positive affective states (rewards) as those that tend to increase the *future* likelihood of behaviours that produce or maintain those states and negative affective states (punishments) as those that tend to increase the *future* likelihood of behaviours that prevent or remove those states.

However there is no need to introduce these effects on learning as part of the *definition* of "affective state", since those causal connections follow from the more general definitions given above. If predictive associative learning is possible in an organism, i.e., if it can discover that some state of affairs S tends to produce another state of affairs S', which is positively or negatively affective, then actions that tend to produce, or to avoid S will have the consequence of producing or avoiding a positively or negatively affective state, and will therefore themselves tend to be

supported or opposed (from the definitions of positive and negative affect). Therefore if S' is positively affective so will S be and if S' is negatively affective so will S be.

So states associated with affective states may themselves become associative affective states. Of course, the relationships become far more complex and subtle in more sophisticated organisms with multiple goals, context sensitive conflict-resolution strategies, explicit as opposed to implicit affective states and belief-like states, derivation processes, and so on.

3.5 Complex affective states

Depression would seem to be a counter-example to our analysis of positive and negative affective states.[5] It is clearly a negative affective state, and yet some forms of depression do not prompt action that tends to remove the state, as our analysis of negative affective states requires. Indeed, depression often prompts behaviours that function to perpetuate the state, the defining characteristic of *positive* affect. How can depression be accommodated under our account?

The answer lies in viewing depression as a *complex* affective state. A possible explanation that employs this view is as follows:

> Having an in-built desire to maximize one's possibilities for action is a plausible feature for autonomous systems. Such a system might be capable of having a negative affective state N of the following sort: it goes into N when it perceives that its set of possible actions is being restricted; and when N occurs, a mechanism E is reliably triggered which generates a variety of attempts to escape from N by escaping from the restrictions. So the state N has the function of making the system engage in activity that tends to remove or diminish N.
>
> But now suppose that there are some situations in which an overall damping of action is adaptive: for instance, hibernation, being in the presence of a dominant conspecific, or having a brutal parent who reacts violently on the slightest provocation. The adaptivity of restricting actions in such situations might result in the evolution of a damping mechanism D that, when activated, globally reduces the possibilities for action, via internal controls. So, when the system detects a situation in which such damping would be advantageous, this produces state P (an example of a mood) where P reliably activates D which, in turn both activates or enhances the negative affective state N, and enhances P. While those conditions in which damping is advantages persist, P would be a positively affective state – it can be desirable to lie low in a dangerous situation even though it is not desirable to be in a dangerous situation and lying low is not normally desirable (e.g., when hungry!). So there will be a conflict between P, whose function is to reduce activity and N, whose function is to increase possibilities for action – but P wins in certain circumstances. In some cases, positive feedback mechanisms could make it very difficult to break out of P, even after the initiating conditions have been removed and continuation of damping would no longer be advantageous.

The actual nature of depression is probably far more complex; this explanation sketch is offered only to show that there is no incompatibility, in principle, between complex states like depression and our analysis of affect.

[5] Thanks to Brian Logan for drawing this to our attention.

Incidentally, this outline explanation also shows that what we call positively or negatively affective states, need not be consciously experienced as pleasant or unpleasant. In fact, the state itself need not be recognized even though some of its consequences are.

Crucial to this explanation is the fact that if two affective sub-states co-exist, one positive and one negative (or if there are two positive or two negative affective states that tend to produce conflicting actions) their effects do not in general "sum up" or "cancel out" as if they were coexisting physical forces. It is even possible for one sub-state to have the specific function of *disabling* the normal effects of another, for instance when being paralyzed by fear prevents the "normal" escape behaviour that would reveal one's location. More generally, vector summation is often not suitable either for combining the effects of coexisting affective states or for dealing with conflicts. Instead of summing, it is normally sensible to *select* one from a set of desirable but incompatible actions, since any "summing" could produce disastrous effects, like Buridan's proverbial ass placed half-way between food and drink. More intelligent organisms may invent ways of satisfying two initially incompatible desires, instead of merely selecting one of them.

3.6 Varieties of affective states and processes

Within the context of a sufficiently rich (e,g, human-like) architecture we can distinguish a wide range of affective states, depending on factors such as:

- whether they are directed (e.g. craving an apple) or non-specific (e.g. general unease or depression),
- whether they are long-lasting or short-lived
- how fast they grow or wane in intensity
- what sorts of belief-like, desire-like and other states they include
- which parts of an architecture trigger them
- which parts of the architecture they can modulate
- whether their operation is detected by processes that monitor them
- whether they in turn can be or are suppressed.
- whether they can become dormant and then be re-awakened later,
- what sorts of external behaviours they produce,
- how they affect internal behaviours, e.g. remembering, deciding, dithering, etc.
- whether they produce second-order affective states (e.g. being ashamed of being angry),
- what sorts of conceptual resources they require.

Many of these distinctions, like the distinctions in the taxonomy in (Ortony et al., 1988). cannot be applied to organisms or robots with much simpler architectures than an adult human architecture. For instance it is not clear that the architecture of a new-born human infant can support long-term affective states that are sometimes dormant because attention is diverted, like long-term grief or intense patriotism.

4 ARCHITECTURAL CONSTRAINTS ON AFFECT

The precise variety of mental states and processes (affective and non-affective) that are possible for an individual, or a species, will depend on the information-processing architecture of that individual or species. Insofar as humans at different stages of development, or humans with various kinds of pathology, or animals of different kinds, or robots, have different sorts of architectures, that will constrain the classes of affective and other kinds of states they support.

The fact that different sorts of architectures support different classes of mental states may mean that care is needed in talking about things like desires, emotions, perception, learning, etc. in different sorts of organisms, e.g., insects, rodents, primates, human infants, human adults, robots of various kinds: varieties of emotions, desires, or consciousness in a newborn infant will be different from those possible in adults. Unfortunately there is no agreed terminology for discussing varieties of architectures so that we can pose questions about which sorts of mental states and processes are possible in which sorts of architectures. We therefore propose the Cogaff Schema as partially defining a high level ontology for components in a wide range of information processing architectures.

4.1 CogAff: a schema allowing multiple types of emotions

The generic CogAff architecture schema sketched in Figures 1 and 2 covers a wide variety of types of possible (virtual machine) architectures for organisms or robots, which vary in the types of sophistication in their perceptual mechanisms, their motor mechanisms and their "central" processing mechanisms, and also in the kinds of connectivity between sub-mechanisms.

For instance, central processes can be purely *reactive*, in the sense of producing immediate (internal or external) actions without the use of any mechanisms for constructing alternative possible multi-step futures and comparing. Alternatively they may be *deliberative*, in the sense of using explicit hypothetical representations of alternative possible futures, or possible predictions, or possible explanations, comparing them and selecting a preferred option. This requires highly specialised and biologically costly mechanisms, including short-term stores for temporary structures of varying complexity, which very few animals seem to have, though simple reactive mechanisms in which two inconsistent reactions are simultaneously activated and then one selected by a competitive mechanism could be described as *proto-deliberative*. Another sub-division among central processes concerns *meta-management* mechanisms which use architectural features that allow internal processes to be monitored, categorised (using an appropriate ontology for information-processing states and processes), evaluated and in some cases controlled or modulated.

These are not mutually exclusive categories, since ultimately all processes have to be implemented in reactive mechanism. Moreover, meta-management processes may be either reactive or deliberative.

Corresponding to the different kinds of processing mechanisms and semantic resources available in the central sub-systems, we can also distinguish layers of abstraction in the perceptual and action sub-systems. For instance, a deliberative layer requires perceptual mechanisms that can discretise, or chunk, the environment into categories between which associations can be learnt that play a role in planning and predicting future events. It is not always appreciated that without such discretisation multi-step planning would require consideration of branching continua: which appears to be totally infeasible. Another sort of correspondence concerns the ability of organisms to perceive others as information-users. Doing this requires perceptual processes to use concepts for other agents that are similar to those the meta-management system uses for self-categorisation.[6] Examples might be seeing another as happy, sad, attentive, puzzled, undecided, angry, looking to the left, etc.

Similarly layers of abstraction in an action system could evolve to meet the varying needs of central layers.

Superimposing two three-fold distinctions gives a grid of nine possible sorts of components for the architecture, providing a crude, high-level classification of sub-mechanisms that may be present or absent. Architectures can vary according to which of these "boxes" are occupied, how they are occupied and what sorts of connections there are between the occupants of the boxes. Further distinctions can be made according to

- whether the components are capable of learning or fixed in their behaviour,
- whether new components and new linkages develop over time
- which sorts of forms of representations and semantic contents are used in the various boxes.

In figure 2 we indicate the possibility of a reactive component that gets inputs from all the other components and sends outputs to all of them. This could be a design for an "alarm" system that detects situations where rapid global redirection of processing is required, one of the ways of thinking about the so-called "limbic system" (discussed by Kelley and by Fellous and Arbib in this volume), though there can be many more specialised "alarm" systems in a complex architecture, such as a protective blinking reflex.

It should be clear that by using such a schema to provide a generic framework relative to which particular architectures can be defined by specifying which components of the grid they incorporate, which links exist between components, and which sorts of formalisms and mechanisms are used in the various components, we can subsume a very wide variety of types of architectures. See also (Sloman and Logan, 2000; Sloman, 2000b).

Many architectures that have been investigated in recent years are purely reactive insofar as they allow only components in the reactive layer e.g. (Nilsson, 1994). Some purely

[6] An interesting research question is whether the self-descriptive mechanisms or the descriptions of others as information-users evolved first, or whether they evolved partly concurrently, as suggested in (Sloman and Logan, 2000). The ability to describe something as perceiving, reasoning, attending, wanting, choosing, etc. seems to require representational capabilities that are neutral between self-description and other-description.

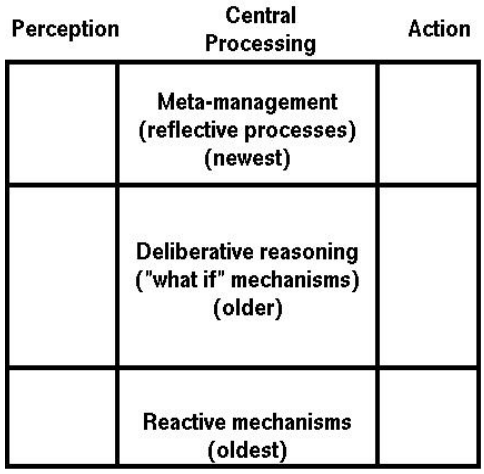

Figure 1: The CogAff schema: two kinds of architectural sub-divisions superimposed. Many information flow-paths between boxes are possible.

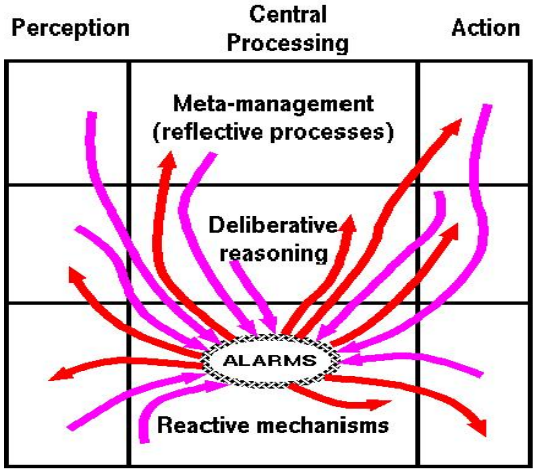

Figure 2: Elaborating the CogAff schema to include reactive alarms – possibly many varieties with different input and output connections.

reactive architectures have layers of control, where all the layers are merely reactive subsystems monitoring and controlling the layers below them (Brooks, 1991). Some early AI systems had purely deliberative architectures, e.g. planners, theorem provers and early versions of SOAR (Laird et al., 1987). Some architectures have different sorts of central processing layers, but do not have corresponding layers of abstraction in their perception and action subsystems. An information flow diagram for such a system would depict information coming in through low-level perceptual mechanisms, then flowing up and then down the central processing tower, and then going out through low level action mechanisms. This sort of flow diagram is reminiscent of a Greek Omega, i.e. Ω, so we call those *omega architectures*, e.g. (Cooper and Shallice, 2000).

4.2 Different architectures support different ontologies

For each type of architecture we can analyse the types of states and processes that can occur in instances of that type, whether organisms or artefacts, and arrive at a taxonomy of types of emotions and other states that the architecture can support. For instance, one class of emotions (primary emotions) might be triggered by input from low level perceptual mechanisms to an "alarm system" (shown in Figure 2), which interrupts "normal processing" in other parts of the reactive sub-system, to deal with emergency situations (we return to this in 5.1). What we are describing as "normal" processing in the other parts, is simply what those parts would do to meet whatever needs they have detected or to perform whatever functions they normally fulfil.

Another class of emotions (secondary emotions) might be triggered by inputs from internal deliberative processes to an alarm system, for instance if a process of planning or reasoning leads to a prediction of some highly dangerous event or a highly desirable opportunity, for which special action is required, e.g. unusual caution or attentiveness. Recognition of this situation by the alarm

mechanism might cause it immediately to send new control signals to many parts of the system, modulating their behaviour (e.g. by pumping hormones into the blood supply). It follows that an architecture that is purely reactive could not support secondary emotions thus defined.

Note, however, that the CogAff framework does not determine a *unique* class of concepts describing possible states, although each *instance* of CogAff does.

A theory-generated ontology of states and processes need not map in a simple way onto the pre-theoretical collection of more or less confused concepts (emotion, mood, desire, pleasure, pain, preference, value, ideal, attitude, and so on). However, instead of simply rejecting the pre-theoretical concepts, we use architecture-based concepts to refine and extend them. There are precedents for this in the history of science: e.g. a theory of the architecture of matter refines and extends our pre-theoretical classifications of kinds of stuff and kinds of processes; a theory of how evolution works refines and extends our pre-theoretical ways of classifying kinds of living things, e.g. grouping whales with fish; and a theory of the physical nature of the cosmos changes our pre-theoretical classifications of observable things in the sky, even though it keeps some of the distinctions, e.g. between planets and stars. See also (Cohen, 1962).

The general CogAff framework should, in principle, be applicable beyond life on earth, to accommodate many alien forms of intelligence, if there are any. However, as it stands it is designed for agents with a located body and some aspects will need to be revised for distributed agents, or purely virtual or otherwise disembodied agents.

If successful for the purposes of science and philosophy, the architecture schema is also likely to be useful for engineering purposes, though many engineering goals can be achieved using shallow concepts (defined purely behaviourally) and shallow theories (linking conditions to observable behaviours). For instance, this may be all that is required for production of simple but effective "believable" agents for computer entertainments.

Intermediate cases may, as pointed out in (Bates, 1994), use architectures that are "broad" in that they encompass many functions, but "shallow" in that the individual components are not realistic. Exploring broad and initially shallow, followed by increasingly deep implementations, may be a good way to understand the general issues. In the later stages of such research we can expect to discover mappings between the architectural functions and neural mechanisms.

4.3 When are architectural layers/levels/divisions the same?

Many people produce layered diagrams indicating different architectural slices through a complex system. However, close textual analysis reveals that things that look the same can actually be very different. For example, there is much talk of "three layer" models, but it is clear that not all three-layered systems include the same sorts of layers! The 3R model presented (by Norman *et al.*) in this volume has three layers: reactive, routine, and reflective, but none of their three layers map directly onto the three layers of the CogAff model. E.g., their middle layer, the *routine* layer, combines some aspects of what we assign to the lowest layer, the reactive layer (e.g., learnt, automatically executable strategies), and their *reflective* layer (like Minsky's reflective layer) includes mechanisms that we label as part of the deliberative layer (e.g., observing performance of a plan and repairing defects in the plan – whereas our third layer would contain only the ability to observe and evaluate internal processes, such as the planning process itself and to improve

planning strategies, like Minsky's "self-reflective" layer). Moreover, what we call "reactive" mechanisms occur in all three layers in the sense that everything ultimately has to be implemented in purely reactive systems.

More importantly, in the 3R model, the reflective layer receives only pre-processed perceptual input, and does not do any perceptual processing itself, whereas CogAff allows for perceptual and action processing in the meta-management layer, for instance seeing a face as happy, or producing behaviour that expresses a high level mental state, such as indecision.

Even when people use the same labels for their layers they often interpret them differently: e.g., some people use "deliberative" to refer to a reactive system which can have two or more simultaneously triggered competing reactions, one of which wins over the other (e.g. using a "winner takes all" neural mechanism). We call that case "proto-deliberative", reserving the label "deliberative" for a system that is able to construct and compare structured descriptions with compositional semantics, where the descriptions do not have a fixed format but can vary according to the task (e.g. planning-trees, theories, explanations of an observed event, etc.). Another example is the tendency in some researchers to use "reactive" to imply "stateless." Unfortunately we do not yet have a good theoretical overview of the space of possible designs comprising both purely reactive and fully deliberative designs. There are probably many interesting intermediate cases that need to be studied if we are to understand both evolution and individual development.

4.4 H-Cogaff: a special case of CogAff

Based on CogAff, we are currently developing a first-draft version of a specific architecture, called H-Cogaff (depicted in Figure 3), which is a special case of the CogAff schema, and is conjectured to cover the main features of the virtual information-processing architecture of normal (adult) humans, though there are still many details to be worked out.

This architecture allows us to define a variety of classes of human emotions, which differ as regards which component of the architecture triggers them and which components they affect: in addition to primary and secondary emotions defined above we distinguish tertiary emotions which perturb or have a disposition to perturb the control of attention in the meta-management subsystem, as explained at length in (Wright et al., 1996). The layers in H-CogAff are also intended to mark significant evolutionary steps. For example, the architecture of H-CogAff assumes that the evolution of the meta-management layer made possible evolution of additional layers in perceptual and action systems related to the needs and capabilities of the meta-management layer (e.g., using the same ontology for labelling internal states and perceived states of others). (See Chapter 9 of (Sloman, 1978) and (Sloman, 1989; Sloman, 2001b; Sloman and Chrisley, 2003).)

4.5 Architectural presuppositions

Our conjectures in sections 2.2 and 3 imply that our folk-psychological concepts and theories all have architectural presuppositions. However, since those presuppositions are sometimes unclear, inarticulate, confused, or inconsistent that will undermine the clarity and consistency of our use of concepts like "emotion", "attention", "learning", etc. So, scientists, engineers and philosophers who use those concepts to ask questions, state theories, or propose practical goals, are likely to

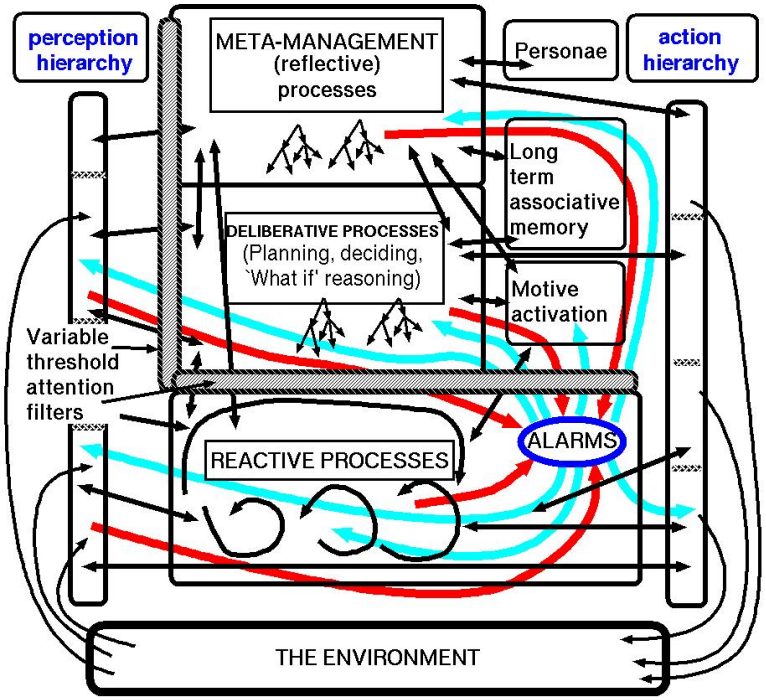

Figure 3: The H-CogAff architecture. The central layer relates to different functional layers in perception and action hierarchies. Not all possible links between boxes are shown. Meta-management may be able to inspect intermediate states in perceptual layers, e.g. sensory qualia.

be at least confused, or unclear. Clear architectural theories can help us avoid such confusion and unclarity, if we use architecture-based concepts.

By defining new more precise versions of our old mental concepts in terms of the kinds of processes supported by an underlying architecture, we may hope to avoid arguing at cross-purposes because of conceptual unclarity and confusion. (Similar comments may be made about using architecture-based analysis to clarify some technical concepts in psychology, e.g. “drive”, “executive function”).

4.6 Where to begin?

We agree with (Turner and Ortony, 1992) that the notion of “basic emotion” involves deep muddles. Searching for a small number of basic emotions from which others are composed is a bit like searching for a small number of chemical reactions from which others are composed. It is the wrong place to look. To understand a wide variety of chemical processes a much better strategy is to look for a collection of “basic” physical processes in physical mechanisms *underlying* chemical reactions and see how they can be combined. Likewise, with emotions, it is better to look for an underlying collection of processes in information-based control systems (a mixture of virtual machines and physical machines) that implement a wide variety of emotional (and other affective) states and processes, rather than trying to isolate a subset of emotions to provide the basis of all

others, e.g. by blending or vector summation.

The kinds of architectural presuppositions on which folk-psychology is based are too vague and too shallow to provide explanations for *working* systems, whether natural or artificial. Nevertheless, folk-psychology is a useful starting point, as it is very rich and includes many concepts and implicit theories that we use successfully in everyday life. However, as scientists and engineers, we have to go beyond the architectures implicit in folk-psychology, adding breadth and depth.

Since we do not know enough yet to get our theories right first time, we must be prepared to explore alternative architectures. In any case there are many types of organisms with many similarities and differences in their architectures. And different applied systems will need different architectures. So there are many reasons for not attending exclusively to any one kind of architecture.

Many alternative conjectured architectures can be inspired by empirical evidence regarding biological systems (including the fact that humans still have many sub-systems that evolved long ago and still exist in other animals, perhaps in different forms). We should also be open to the possibility of biological discoveries of architectures that do not fit our schema, for which the schema will have to be extended. Moreover, we are not restricted to what is biologically plausible. We can also consider architectures for future possible robots.

5 EXAMPLES OF ARCHITECTURE-BASED CONCEPTS

We are attempting to extend folk-psychological architectures in the framework of the CogAff schema (figure 1), which supports a wide variety of architectures. An example is our tentatively proposed special case, the H-CogAff architecture offered as a first draft theory of the human virtual information processing architecture. In the context of H-CogAff we can distinguish more varieties of emotions than are normally distinguished (and more varieties of perceiving, learning, deciding, attending, acting, etc. too). However it is likely that the ontology for mental states and processes that will emerge from more advanced versions of H-CogAff (or its successors) will be far more complex than anyone now imagines.

We shall offer some examples of words normally regarded as referring to emotions and show how to analyse them in the context of an architecture. We start with a proposal for a generic definition of emotion that might cover many states that are of interest to psychologists who are trying to understand emotions in humans as well as to roboticists intending to study the utility of emotional control in artifacts. This is an elaboration of ideas originally in (Simon, 1967).

5.1 Towards a generic definition of “emotion”

We start from the assumption that in any information-processing system there are temporally extended processes that sometimes require more time to complete a task than is available, because of the speed with which external events occur. For example the task of working out how to get some food that is out of reach may not be finished by the time a large, fast approaching object is detected, requiring evasive action. An operating system might be trying to write data to a memory

device, but the user starts disconnecting the device before the transfer is complete. It may be useful to have a process which detects such cases and interrupts normal functioning, producing a very rapid default response, taking high priority over everything else, to avoid file corruption. In figure 2 we used the label "alarm mechanism" for such a fast-acting system which avoids some danger or which grasps some short-lived opportunity.

In an animal or robot, this sort of alarm mechanism will have to use very fast pattern-triggered actions, and may be less sophisticated in its reasoning, and therefore more likely, in some cases, to produce an inappropriate response, than the mechanism which it interrupts and overrides would have produced if it had had sufficient time to complete its processing. However, the frequency of wrong responses might be reduced by training in a wide variety of circumstances. This notion can also be generalised to cases where instead of interrupting the alarm mechanism merely *modulates* the "normal process" (e.g., by slowing it down or turning on some extra resources which are normally not needed such as mechanisms for paying attention to details).

We can use the idea of an alarm system to attempt a very general definition of "emotion": An organism is in an *emotional state* if it is in an episodic or dispositional state in which some part of it whose biological function is to detect and respond to 'abnormal' states has detected something and is either

1. *actually* (episodic) interrupting, preventing, disturbing, or modulating one or more processes which were initiated or would have been initiated independently of this detection,

   or

2. *disposed* (under certain conditions) to interrupt, prevent, disturb, etc. such processes, but currently suppressed by a filter (Figure 3) or priority mechanism.

We have given examples involving a speed requirement, but other examples may involve detection of some risk or opportunity which requires an ongoing action to be altered but not necessarily at high speed, for instance noticing that you are going to be near a potentially harmful object if you do not revise your course.

This architecture-based notion of emotion (involving actual or potential disruption or modulation of "normal" processing) falls under the very general notion of "affective" (desire-like) state or process as analysed in section 3.2. It encompasses a large class of states that might be of interest to psychologists and engineers alike. In the limiting cases, it could even apply to relatively simple organisms such as insects, like the fly whose feeding is aborted by detection of the fly-swatter moving rapidly towards it, or the woodlouse that quickly rolls up into a ball if touched by a pencil. For even simpler organisms, e.g. a single-celled organism, it is not clear whether the information processing architecture is rich enough to support the required notions.

This generic notion of emotion as "actual or potential disturbance of normal processing" can be subdivided into many different cases, depending on the architecture involved, and where in the architecture the process is initiated, what it disturbs, and how it does so. There is no implication that the disturbance will be externally visible, or measurable, though often it will be, if the processes that are modified include external actions.

Previous papers, e.g. (Sloman, 2001a), elaborated this idea by defining "primary" emotions as entirely triggered within a reactive mechanism, "secondary" emotions as those triggered within a

deliberative system, and "tertiary" emotions (referred to as "perturbances" in the analysis of grief in (Wright et al., 1996)) as states and processes that involve actual or dispositional disruption of attention-control processes in the meta-management (reflective) system. That is just a very crude, inadequate, first draft high level subdivision which does not capture the rich variety of processes colloquially described as "emotions" or "emotional".

Within the framework of an architecture as rich as H-Cogaff many more subdivisions are possible, including sub-divisions concerning different time-scales, different numbers of interacting sub-processes, different aetiologies, different sorts of semantic content etc. This overlaps with the taxonomy in (Ortony et al., 1988).

5.2 An architecture-based analysis of "being afraid"

Many specific emotion concepts (e.g. fear, joy, disgust, jealousy, infatuation, grief, obsessive ambition, etc.) share some of the polymorphism, and indeterminacy of the general concept. For example, "fear" and "afraid" cover many types of states and processes. Consider being:

1. afraid of spiders
2. afraid of large vehicles
3. afraid of a large vehicle careering towards you
4. afraid of a thug asking you to hand over your wallet
5. afraid your favourite party is going to lose the next election
6. afraid you have some horrible disease
7. afraid of growing old
8. afraid that your recently published proof of Goldbach's conjecture has some hidden flaw.

Each of these different forms of "being afraid" requires a minimal set of architectural features (i.e., components and links among them) to be present in the architecture of the individual concerned. For example, there are instances of the first four forms which involve perceptions that directly cause the instantiation of the state of being afraid, while the other four do not depend on perception to cause their instantiation. E.g. merely remembering that your proof has been published might be sufficient to cause fear that the proof has a hidden flaw. There are states that inherently come from mental processes other than current perception, e.g. embarrassment about what you said yesterday.

Furthermore, the above states vary in cognitive sophistication. The first, for example, might only require a reactive perceptual process that involves a matcher comparing current perceptions to a innate patterns (i.e., those of spiders), which, in turn, triggers an alarm mechanism. The alarm mechanism could then cause various visceral processes (such as release of hormones, the widening of the pupils, etc.) in addition to modifications of action tendencies and dispositions (e.g., the disposition to run away or to scream – compare LeDoux (1996)).

The second, for example, could be similar to the first in that large objects cause anxiety, or it could be learnt e.g., because fast approaching vehicles in the past have caused state 3

to be instantiated, which in turn formed an association between it and large vehicles, so that the presence of large vehicles alone can instantiate state 3. State 2 then involves a permanent dispositional state by virtue of the learnt associative connection between large vehicles and state 3. State 2 ceases to be dormant upon perceiving a large vehicle, regardless of whether it is approaching or not.

The fourth involves even more in that it requires projections into the future and is instantiated because of possible negative outcomes. Consequently, a system that can instantiate state 4 will have to be able to construe and represent possible future states and maybe assess their likelihood. Note, however, that simple forms of state 4 might be possible in a system that has learnt a temporal association only (namely that a particular situation, e.g., that of a thug asking for one's wallet, is always preceded by encountering a thug). In that case, a simple conditioning mechanism might be sufficient.

For the remaining examples, however, conditioning is not sufficient. Rather, reasoning processes of varying complexity are required that combine various kinds of information. In the case of state 6 this may be evidence from one's medical history, statements of doctors, common sense knowledge, etc. The information needs to be corroborated in some way (whether the corroboration is valid or not does not matter) to cause the instantiation of these states. For the last three, it is likely that additional reflective processes are involved, which are capable of representing the very system that instantiates them in different possible contexts and evaluate future outcomes with respect to these contexts and the role of the system in them (e.g., a context in which the disease has manifested itself and how friends would react to it, or how colleagues would perceive one's failure at getting the proof right).

The above paragraphs are, of course, only very sketchy outlines that hint at the kind of functional analysis we have in mind, which eventually leads to a list of functional components that are required for an affective state of a particular kind to be instantiable in an architecture. Once these requirements are fixed, then it is possible to define the state in terms of these requirements and also ask whether a particular architecture is capable of instantiating the state. For example, if reflective processes that observe, monitor, inspect, and modify deliberative processes are part of the last three states, then architectures without a meta-management layer (as defined in CogAff) will not be capable of instantiating any of them.

This kind of analysis is obviously not restricted to the above states, but could be done for any form of anger (Sloman, 1982), fear, grief (Wright et al., 1996), pride, jealousy, excited anticipation, infatuation, relief, various kinds of joy, schadenfreude, spite, shame, embarrassment, guilt, regret, delight, enjoyment (of a state or activity) etc. Architecture-based analyses are also possible for other non-emotional, affective states such as attitudes, moods, states like surprise, expectation, and the like.

6 DISCUSSION

Our approach to the study of emotions in terms of properties of agent architectures can safely be ignored by engineers whose sole object is to produce "believable" mechanical toys or displays that present appearances that trigger, in humans, the attribution of emotional and other mental

states. Such "emotional models" are based on *shallow concepts* that are exclusively defined in terms of observable behaviours and measurable states of the system. This is in contrast to deep concepts, which are based on theoretical entities (such as mechanisms, information structures, types of information, architectures, etc.) postulated to generate those behaviours and states, but not necessarily directly observable or measurable (as most of the theoretical entities of physics and chemistry are not directly observable).

Implementing *shallow models* does not take much, if, for example, the criteria for success depend only on human ratings of the "emotionality" of the system, for we, as human observers, are predisposed to confer mental states even upon very simple systems (as long as they obey basic rules of behavior, e.g., Disney cartoons). At the same time, shallow models do not advance our theoretical understanding of the functional roles of emotions in agent architectures as they are effectively silent about processes internal to an agent. Shallow definitions of emotions, for example, would make it impossible for someone whose face has been destroyed by fire, or whose limbs have been paralysed, etc. to have various emotional states that are *defined* in terms of facial expressions and bodily movements. In contrast, architecture-based notions would allow people (or robots) to have joy, fear, anguish, despair, relief, etc. despite lacking any normal way of expressing them.

The majority view in this volume seems to be that we need explanatory theories including theoretical entities whose properties may not be directly detectable, at least using the methods of the physical sciences or the measurements familiar to psychologists (including button-pushing events, timings, questionnaire results, etc.). This is consistent with the generic definition of "emotion" proposed in this chapter based on internal processes that are capable of modulating other processes (i.e., initiating or interrupting them, changing parameters that give rise to dispositional changes, etc.). Such a definition should be useful for psychologists interested in the study of human emotions and for engineers implementing deep emotional control systems for robots or virtual agents. While the definition was not intended to cover *all aspects* of the ordinary notion use of the word "emotion" (nor could it cover them all given that "emotion" is a cluster concept), it can be used as a guideline that determines the minimal set of architectural features necessary to implement emotions (as defined). Furthermore, it allows us to determine whether a given architecture is capable of implementing such emotions, and if so of what kinds (as different emotion terms are defined in terms of architectural features). This is different from much research in AI, where it is merely taken as "obvious" that a system of a certain sort is indeed emotional.

More importantly, our definition also suggests possible roles of mechanisms generating what are described as "emotions" in agent architectures (e.g., as interrupt controllers, process modifiers, action initiators or suppressors, etc.), and hence, when and where it is appropriate and useful to employ such control systems. This is crucial for a general understanding of the utility of what is often referred to as "emotional control" and consequently the adaptive advantage of the underlying mechanisms in biological systems, even though many of the emotions they produce may be dysfunctional.

6.1 Do robots need emotions and why?

One of the questions some robot designers address is whether there is any principled reason why their robots need "emotions" to perform a given task (assuming some clear definition of "emotion"). However there is a more general question, namely whether there is any task that cannot be performed by a system that is not capable of having emotional states.

The answer to this question is certainly non-trivial in the general case. For simple control systems satisfying a particular definition of 'emotional', it may be possible to define a finite state machine, which has exactly the same input-output behavior, but does not instantiate any emotion in the specified sense. Most so-called 'emotional' agents currently developed in AI would probably fall under this category.

While this idea applies in principle to agents of all levels of complexity, in practice there are a limits to the approach, and the situation will already be very different for more complex agents. For one, implementing the control system as a finite state controller will not work as the number of states of a complex agents (e.g., with thousands of condition-action rules involving complex representations) will likely be too large for the state table to fit into a standard computer. Hence, the control system needs to be implemented in a virtual machine that supports multiple finite state machines with substates and connections among them. In short, a complex architecture with complex states will have to be implemented in a virtual machine that supports the complexity. While transitions are immediate in finite state machines, many steps may be required for a complex virtual machine transition (like a computer updating a simulated neural net). Finite state machines do not need alarm systems to interrupt normal processing in order to react to unforeseen events: they simply transit into a state where they deal with the circumstance. Complex systems with multiple finite state machines with complex substates, however, need a way of coordinating state transitions (especially if they have different lengths, might take different amounts of time, or might even occur asynchronously). In that case, special mechanisms need to be added to improve the reactivity of the system (i.e., the time it takes to respond to critical environmental changes).

Following this reasoning, one would expect to find something like alarm mechanisms in complex agents that need to react quickly in real-time to unforeseen events. Such systems might lead to internal interactions instantiating emotional states as defined above which the designers did not intend (e.g., an operating system with a mechanism that terminates processes, limits and reallocates resources, etc. in response to an overload, might delete processes urgently required for some sub-task).

Returning to the question whether robots need or should have emotions, the answer will depend on the task and environment for which then robot is intended. This "niche", i.e., the set of requirements to be satisfied, will, in turn, determine a range of architectures able to satisfy the requirements. The architectures will then determine the sorts of emotions that are possible (or desirable) for the robot. Here are some examples of questions designers may ask:

- Will the robot be purely for entertainment?
- Will it have a routine practical task, e.g. on a factory floor or in the home (cleaning carpets)?
- Will it have to undertake dangerous tasks in a dynamic and unpredictable environment (as in the Robocup Rescue project)?

- Will it have to cooperate with other agents (robots and humans/animals)?
- Will it be a long term friend or helper for one or more humans (e.g. robots to help the disabled or infirm)?
- Will its tasks include understanding humans with whom it interacts?
- Will it need to fit into different cultures or sub-cultures with different tastes, preferences, values, etc.?
- Will the designers be able to anticipate all the kinds of problems and conflicts that can arise during the 'life' of the robot?
- Will it ever need to resolve ethical conflicts on its own, or will it always refer such problems to humans? (Maybe there won't be time, or communication links, if it's down a mine or in a space-craft on a distant planet....)
- Will it need to be able to provide explanations and justifications for its goals, preferences, decisions, etc.?
- Is the design process aimed primarily at scientific goals, i.e. trying to understand how human (and other animal) minds work, or are the objectives practical, i.e. to get some task done? We are mainly interested in the science, whereas some people are primarily interested in practical goals.)

A full treatment will require a survey of niche-space and design-space and the relationships between them. (This is also required for understanding evolutionary and developmental trajectories.)

To say that certain mechanisms, forms of representation, architectural organisation, are required for an animal or robot is to say something about the niche of that animal or robot and what sorts of information processing capabilities, behaviours, etc. are well suited to doing well (surviving, flourishing, reproducing successfully, achieving individual goals etc.) in that niche.

6.2 How are emotions implemented?

Another important, recurring question raised in the literature on emotions (in AI) is whether a realistic architecture needs to include some particular, dedicated "emotion mechanism". Our view (e.g., as argued in (Sloman and Croucher, 1981; Sloman, 2001a)) is that in realistic human-like robots, emotions of various types will *emerge*, as they do in humans, from various types of interactions between many mechanisms serving different purposes, not from a dedicated "emotion mechanism".

Another issue is whether emotions are necessarily tied to visceral processes, as assumed in biological theories that construe notions like "emotion", "affect", "mood" as characterising physical entities (animal bodies, including brains, muscles, skin, circulatory system, hormonal

systems, etc.). If the presence of an emotion requires a body of a particular type (e.g., with chemical hormones), then there will never be (non-biological) robots with emotions.

Alternatively, one could take emotion terms to refer to states and processes in virtual machines that happen to be implemented in these particular physical mechanisms but might in principle be implemented in different mechanisms. In that case, non-biological artefacts may be capable of implementing emotions as long as they are capable of implementing all relevant causal relationships that are part of the definition of the emotion term. The above alternatives are not mutually exclusive, for there is nothing to rule out the combination of

- deep, implementation-neutral, architecture-based concepts of emotion, definable in terms of virtual machine architectures without reference to implementation-dependent properties of the physical substratum

- special cases (i.e. sub-concepts) that are implementation-dependent and defined in terms of specific types of bodies and how they express their states (e.g., snarling, weeping, grimacing, tensing, changing colour, jumping up and down, etc.).

LeDoux (1996) Panksepp (1998) are examples of such "special cases", where emotions are defined in terms of particular brain regions and pathways. These definitions are intrinsically dependent on a particular bodily make-up (i.e., anatomical, physiological, chemical, etc.). Hence, systems that do not possess the respective bodies cannot, by definition, implement them.

The conceptual framework of Ortony et al. (Ortony et al., 1988), on the other hand, is an example of an implementation-neutral conception, where emotions are defined in terms of an ontology distinguishing events, objects, and agents and their relationship to the system implementing the emotion. It is interesting to note that if emotions are reactions to events, agents, or objects (as Ortony et al. (Ortony et al., 1988) claim), then their agent-based emotions, i.e., emotions elicited by agents, cannot be instantiated in architectures that do not support representation of the ontological distinction between objects and agents. Such systems could consequently never be jealous (as being jealous involves other agents). This is a virtual machine design constraint, not an implementation constraint.

6.3 Comparison with other work

There is now so much work on emotions in so many disciplines that a comparison with alternative theories would require a whole book. Readers of this volume will be able to decide which of the other authors have explicitly or implicitly adopted definitions of 'emotion' that take account of the underlying architecture and the processes that the architecture can support, which have assumed that there is a clear and unambiguous notion of 'emotion' and which have not, which are primarily interested in solving an engineering design problem (e.g. producing artefacts that are entertaining, or demonstrate how humans react to certain perceived behaviours) and which are attempting to model or explain naturally occurring states and processes. One thing that is relatively unusual that we have attempted is producing a generic framework that should be able to accommodate a wide variety of types of organisms and machines. We hope that more researchers will accept that challenge, and the challenge of attempting to come up with a useful ontology for describing and

comparing different architectures so that our work can grow into a mature science instead of a large collection of ad hoc and loosely related studies that are hard to compare and contrast.

The view we have propounded contradicts some well known theories of emotions, in particular Jamesian theories (James, 1890; Damasio, 1994) according to which having an emotion involves sensing some pattern in one's physiological state. The claim that many emotions involve changes to physiological states (e.g. blood pressure, muscular tension, hormones in the blood stream) is perfectly consistent with what we have said about emotions, but not the claim that such processes are *necessary* conditions for emotions. Theories of this sort have a hard problem accommodating long term emotional states that are often temporarily suppressed by other states and processes, for instance long term grief, long term concern about a threat to one's job, intense long term devotion to a political project, etc.

On the other hand (Barkley, 1997) presents architectural ideas partly similar to our own, though arrived at from a completely different standpoint (he is a neuropsychiatrist). Our emphasis on the link between the concept of emotion and mechanisms that produce strong dispositions to disrupt and redirect other processing also fits much folk psychology and also features of emotions that make them the subject of novels. Changes in blood pressure, galvanic skin responses, levels of hormones are not usually of much interest to readers of great literature, compared with changes in thought processes, in preferences, in evaluations, in how much people can control their desires, in the extent to which their attention is strongly held by someone or something, and the consequences thereof etc. These are features of what we have called 'tertiary' emotions, which usually involve rich semantic content as well as strong control states. It is arguable that only linguistic expression is capable of conveying the vast majority of tertiary emotions, whereas most current research on detecting emotions focuses on such "peripheral" phenomena as facial expression, posture and other easily measurable physiological states.

When a robot first tells you *in detail* why it is upset by your critical analysis of the poems it has written, you will be far more likely to believe it has emotions than if it merely blushes, weeps, shakes its head, etc. Even ducking to avoid being hit by a large moving object might just be a simple planned response to a perceived threat, in a robot whose processing speeds are so great that it needs no alarm mechanism.

7 THE NEXT STEPS

Emotions, in the sense defined in Section 5.1, are present in many controlled systems, where parts of the control mechanism can detect abnormal states and react to them (causing a change in the normal processing of the control system, either directly through interruption of the current processing or dispositionally through modification of processing parameters). Emotions thus defined are not intrinsically connected to living creatures, nor are they dependent on biological mechanisms — e.g., operating systems running on standard computers have several emotions in our technical sense, although they lack many of the detailed features of the sorts of emotions to which our folk concepts are applied.

What *is* special about at least a subset of emotions so defined (compared to other non-emotional control states) is that it can be shown that they (1) form a class of *useful* control

states that (2) are likely to evolve in certain resource-constrained environments and, hence, (3) may therefore also prove useful for certain AI applications (e.g., robots that have only limited processing resources, which impose severe constraints on the kinds of control mechanisms that can be implemented on them).

Useful affective control mechanisms are likely to evolve if there are many evolutionary trajectories that, given various sets of well specified initial conditions and fitness functions, will lead to those control systems (e.g., (Scheutz, 2001; Scheutz and Schermerhorn, 2002). A subset of those will be control mechanisms that can produce emotional states suited to coping with emergencies or unexpected situations as they occur in dynamic, unpredictable real-world environments.

In the case of more subtle and complex long term emotional states, such as grief, ambition, jealousy, infatuation, and obsession with a difficult problem, it is not yet clear which of them are merely side-effects of desirable mechanisms and which are states that can be shown to be useful in relation either to the needs of individuals or needs of a social group, or a species. Human aberrations make it clear, however, that machines containing useful mechanisms are capable of getting into highly dysfunctional states through the interactions of those mechanisms. As machines become more human-like we can expect some undesirable emotional states, to be hard to avoid in certain contexts, if the machines have affective control mechanism that interact in complex ways.

Detailed studies of design and niche space, in which the relationships between classes of designs and classes of niche for these designs in a variety of environments are investigated, should clarify the costs and benefits. For this, we need experiments with agent architectures that complement theoretical, functional analyses of control systems by systematic studies of performance-cost trade-offs, which will reveal utility or disadvantages of various forms of control in various environments.

Finally, the main utility in AI of control systems producing states conforming to our suggested definition of "emotional" does not lie in systems that need to interact with humans or animals (e.g., by recognizing emotions in others and displaying emotions to others). There is no reason to believe that such control mechanisms (where something can modulate or override the normal behaviour of something else) are necessary to achieve "believable interactions" among artifacts and humans. Large sets of condition-action rules, for example, may produce convincing behavioral expressions giving the appearance of sympathy, surprise, etc. without implementing the kinds of control mechanisms which we called "emotional". Hence, such systems may appear to be emotional without actually having emotions in our sense. But appearances will suffice for many applications, especially in computer games and entertainments, as they do in human stage performances and in cartoon films.

In contrast, control mechanisms capable of producing states conforming to our proposed definition of "emotional" will be useful in systems that need to cope with dynamically changing, partly unpredictable and unobservable situations where prior knowledge is insufficient to cover all possible outcomes. Specifically, noisy and/or faulty sensors, inexact effectors, and insufficient time to carry out reasoning processes are all limiting factors that real world, real time systems have to deal with. As argued in (Simon, 1967; Sloman and Croucher, 1981) architectures for such systems will require mechanisms able to deal with unexpected situations. In part, this trivialises

the claim that emotional controls are useful, since they turn out to be instances of very general requirements that are obvious to engineers who have to design robust and "failsafe" systems to operate in complex environments. What is non-trivial is which varieties of such systems are useful in different sorts of architectures, and why.

There is much work in computer science and robotics that deals with control systems that have some features in common with what we call affective mechanisms, from real-time operating systems that implement timers, alarm mechanisms, etc. to be able to achieve time critical tasks, to robot control systems that drive an autonomous unmanned vehicle and need to react to and correct different kinds of errors at different levels of processing (e.g., (Albus, 2000)).

As our field matures it should be possible to explicate this practical wisdom developed in the engineering sciences and compare it to findings in psychology and neuroscience about the control architectures of biological creatures in a coherent way. For this, we need a conceptual framework in which we can express control concepts useful in the description of neural circuits, in the description of higher level mental processes, and in control theory and related fields. Such a conceptual framework will allow us to see the commonalities and differences in various kinds of affective and non-affective control mechanisms found in biological systems or designed into machines. Systematic studies of architectural trade-offs will help us understand the kinds of situations where emotional control states should be employed, because they will be beneficial, situations where they should be avoided because they are harmful and situations where they arise unavoidably out of interactions between mechanisms that are useful for other reasons.

8 ACKNOWLEDGEMENTS

This work is funded by grant F/94/BW from the Leverhulme Trust, for research on 'Evolvable virtual information processing architectures for human-like minds'. The ideas presented here were inspired especially by the work of Herbert Simon, and developed with the help of Luc Beaudoin, Ian Wright, Brian Logan, Marvin Minsky, Ruth Kavanagh, and also many students, colleagues and friends. We are grateful for comments and suggestions from the editors, and for their patience (i.e. lack of emotion).

9 REFERENCES

References

Albus, J. S. (2000). 4-D/RCS Reference model architecture for unmanned ground vehicles. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation*.

Austin, J. (1956). A plea for excuses. In Urmson, J. O. and Warnock, G. J., editors, *Philosophical Papers*, pages 175–204. Oxford University Press, Oxford.

Barkley, R. A. (1997). *ADHD and the nature of self-control*. The Guildford Press, New York.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

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/).

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

Cohen, L. (1962). *The diversity of meaning*. Methuen & Co Ltd, London.

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

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

Delancey, C. (2002). *Passionate Engines: What Emotions Reveal about the Mind and Artificial Intelligence*. Oxford University press, Oxford.

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

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

Hume, D. (1739). *A Treatise of Human Nature*. Oxford University Press, New York. 2nd Ed 1978.

James, W. (1890). *The Principles of Psychology*. Henry Holt, New York.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64.

Lakatos, I. (1970). *Criticism and the Growth of Knowledge*. Cambridge University Press, New York.

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

Lodge, D. (2002). *Consciousness and the Novel: Connected Essays*. Secker & Warburg, London.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. MIT Press, Cambridge.

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.

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

Ortony, A. (2002). On making believable emotional agents believable. In Trappl, R., Petta, P., and Payr, S., editors, *Emotions in Humans and Artifacts*, pages 189–211. MIT Press, Cambridge, MA.

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

Panksepp, J. (1998). *Affective neuroscience-The Foundations of Human and Animal Emotions*. Oxford University Press, Oxford.

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

Sartre, J.-P. (1939). *The Emotions: A Sketch of a Theory*. Macmillan.

Scheutz, M. (2001). The evolution of simple affective states in multi-agent environments. In Cañamero, D., editor, *Proceedings AAAI Fall Symposium 01*, pages 123–128, Falmouth, MA. AAAI Press.

Scheutz, M. and Schermerhorn, P. (2002). Steps towards a systematic investigation of possible evolutionary trajectories from reactive to deliberative control systems. In Standish, R., editor, *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..*, editor, *Intelligent Agent Technology: Research and Development*, pages 200–209. World Scientific Publisher, New Jersey.

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

Sloman, A. (1978). *The Computer Revolution in Philosophy*. Harvester Press (and Humanities Press), Hassocks, Sussex. 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. (1989). On designing a visual system (Towards a Gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337.

Sloman, A. (1993). 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. (1996). 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. (2000a). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M.Schoenauer, *et al.*., editor, *Parallel Problem Solving from Nature – PPSN VI*, Lecture Notes in Computer Science, No 1917, pages 3–16, Berlin. Springer-Verlag.

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

Sloman, A. (2001b). Evolvable biologically plausible visual architectures. In Cootes, T. and Taylor, C., editors, *Proceedings of British Machine Vision Conference*, pages 313–322, Manchester. BMVA.

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 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*, pages 197–202, Vancouver.

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

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.