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 2, 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

Tel: +44 121 414 4775
Fax: +44 121 414 4281 (not reliable – use email if possible)
Email: A.Sloman@cs.bham.ac.uk

Abstract

Much discussion of emotions and related topics is riddled with confusion because the key words are used with different meanings by different authors. For instance, some fail to distinguish the concept of "emotion" from the more general concept of "affect" which covers other things besides emotions, including moods, attitudes, desires, preferences, intentions, dislikes, etc. Others simply treat all forms of motivation as emotions. Moreover researchers do not all have the same research goals: some are primarily concerned with understanding natural phenomena, in humans and other animals, while some are more concerned with producing useful artefacts, e.g. synthetic entertainment agents, sympathetic machine interfaces, and the like. The work in our group has aimed, over several years, to address the confusion by showing how "architecture-based" concepts can extend and refine our pre-theoretical concepts in directions that make them more useful as a basis both for expressing scientific questions and theories, and for specifying engineering objectives. The catch is that different information-processing architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and so on.

As a step towards a high level overview of the variety of types of architectures, we analyse some very basic notions such as "need", "function", "information-user", "affect", "information-processing architecture" and proceed to offer the CogAff schema, which distinguishes types of components that may be in a architecture, operating concurrently with different functional roles. We also offer a first-draft sketch of H-Cogaff, a conjectured type of architecture which is an instance of the CogAff schema required to explain or replicate human mental phenomena, and show how the concepts that are definable in terms of such an architecture can clarify and enrich research on human emotions – our primary objective. 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 realistic 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".

[Temporary table of contents]

Contents

1 Introduction

There are many confusions and ambiguities that 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 that arise within an information-processing architecture. 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.

Concepts like "belief", "desire", "emotion" are then construed as *architecture-based*.[1] In this framework, 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 needs.

Answering such questions requires us to investigate not only mechanisms built in to an architecture to provide specific sorts of functionality, but also more global patterns of processing that emerge from interactions between components. So some robot emotions may emerge, as they do in humans, from the interactions of many mechanisms serving different purposes, not from a particular, dedicated "emotion mechanism". Of course, in an architecture whose components all serve useful purposes, the interactions may also have undesirable side-effects (like deadlocks, or thrashing, in an operating system), and those dysfunctional processes could include certain classes of emotions. So some sorts of emotions, while not needed, may be unavoidable in certain architectures under certain circumstances — as pointed out in (Sloman & Croucher 1981).

Another implication of these ideas is that different architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and so on. Is there any reason for using the same labels to describe agents with very different architectures: e.g. can we use the same concept of "emotion" in discussing a fly's response to a fly-swatter and a human's response to a falling rock? Similar questions can be asked about other concepts concerned with affect, e.g. moods, desires, pleasures, pains, attitudes, values, preferences, etc.

Although cognitive scientists and AI researchers have written much about architectures in recent years there is no clear agreement concerning what sorts of architectures are possible what their costs and benefits are, or how they should be described. Thus even though labels like "reactive", "deliberative" and "reflective" are employed in characterising some architectures or architectural mechanisms, people do not necessarily mean the same things by these words (as will be found by comparing the papers in this volume, for instance). As a step towards remedying this situation we shall try in section 5.1 to present a general schema, the CogAff schema, for describing a wide variety of architectures, and then later, in section 5.4 sketch a special case of that called H-Cogaff, conjectured as a first-draft theory of the architecture of a typical human adult, and capable of supporting a wide variety of types of affective states and processes.

In addition to communication problems caused by conceptual confusions, there are some that arise because researchers do not all have the same research goals. For instance, some researchers are primarily concerned with understanding natural phenomena in humans and other animals,

[1] Compare (Sloman 2002*a*, Sloman 2000*c*, Sloman 2001*a*).

while some are more concerned with producing useful artefacts, e.g. synthetic entertainment agents, sympathetic machine interfaces or robots able to cope with uncertainty and danger. So the question whether a machine needs to have either emotions or behaviours that give the appearance of having emotions (or both!), will depend not only on what "emotions" refers to but also on what the machines are for.

The work in our group has aimed, over several years, to address confusions about "emotion" and other mental concepts by showing how *architecture-based* concepts can extend and refine our pre-theoretical concepts in directions that make them more useful as a basis for expressing scientific questions and theories, and for specifying engineering objectives. We shall summarise the ideas in the next few sections. We also show how the concepts that are definable in terms of an architecture like H-Cogaff can clarify and enrich research on human emotions – our primary objective. Moreover, it could also be relevant to the design of human-like robots and other kinds of synthetic agents. We'll suggest in section 5.6 that from this viewpoint the attempt to identify a set of "basic" emotions from which others are composed is misguided somewhat like trying to identify a set of basic chemical reactions from which others are composed. Rather we need to look for *underlying mechanisms* and understand the variety of states and processes they can generate.

However, before we can explain those ideas we need to sort out some very general notions including *need*, *serving a need* and *having a function*. We shall use those ideas in talking about organisms or machines that use information to meet their needs.

2 The general notion of 'need'

2.1 Needs and functions that serve needs

In what follows we shall make use of the notion of an *information-processing architecture*. A full analysis of this notion is beyond the scope of this paper, but we shall try in this and the next section to make the key ideas clear. We start by slightly generalising the ordinary notion of 'need' to cover a wider variety of cases, some of them trivial some not. Later we introduce more specific types of needs and discuss needs served by the use of information.

This 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. In this sense all needs are relative to whatever they are necessary for. Any statement that X needs Y presupposes that there is some actual or possible state, event or process involving X, here expressed as $P(X)$, for which Y is necessary.[2] Moreover, the need may also be relative to a context, since in some contexts Y may not be necessary because something else suffices to produce or preserve the state, event, or process.

So just as the statement 'X is married' is elliptical for 'there is someone to whom X is married', so also any statement that 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 $P(X)$*. Here P is a predicate expressing what Y is needed for. $P(X)$ may or may not already

[2] So there are no 'absolute' (i.e. non-relative) needs even though we often use grammatical constructs suggesting that there are.

be true when a need statement is true. We can also regard a *context* as specified by a statement that may or may not be true. We shall see that such statements of need are actually shorthand for a complex collection of counterfactual conditional statements about what would happen if...

If X needs Y and in certain contexts Y' is necessary for Y to exist, then in those contexts X also needs Y'. We can call Y a *direct* need and Y' an *indirect* or *derivative* need, or say that the former is needed *directly* the latter *indirectly* or *derivatively*. Many more sub-cases could be distinguished in a full discussion of the notion of 'need', but this is not the place.

We shall explain below what it is for an object to include mechanisms or features which have *functions* in relation to such needs, e.g. the function of meeting a need or, less directly, supporting other things that meet or help to meed the need. In some cases the functions, like the needs, will be trivial, in others not.

We can then talk about information-users, namely things that use information or have components that use information, in order to meet their needs (in this general sense of need). So both information and information-processing mechanisms can have functions for information-users.

We are not here using the technical Shannon/Weaver notion of information, which is a purely syntactic notion. Instead we are using the commonplace notion of information which is *about* something. Information in this sense has content, has implications, may be consistent or inconsistent, and in some cases may be true or false, or obeyed or disobeyed in the case of instructions or rules with information content. In this general sense all biological organisms are information processors: They use information about themselves and their environments in deciding what to do, and they use genetic information in building and repairing themselves.[3] All of this requires access to a store of energy, and in some cases also materials that the organism can use.

All of these notions are analysable in terms of what the individual or some sub-mechanism *would do if ...*, i.e., the notions depend on truth and falsity of more or less complex sets of counterfactual conditionals expressing causal relationships. Explaining why talking like this has a place in science, requires more detailed analysis of concepts like "function", "need", "representation", "information" and "cause".[4]

States of an object can change depending on which mechanisms are active and what they are doing. X may have a need that is not being met, and then it may switch to a state in which the need is being met, e.g. because a part of X is doing something it was not doing previously. We call a state in which something is performing its function of serving a need, a *functional* state.[5] Later we'll distinguish desire-like, belief-like and other sorts of functional states of an information-using organism or machine. The label 'affective' as generally understood seems to be very close to our notion of a desire-like state, and subsumes a wide variety of more specific types of affective states.

Information-users can vary in the number and variety of mechanisms they have that contribute, directly or indirectly, to meeting their needs. We can make further distinctions that depend on how the information is used, e.g. whether it is used directly and immediately or used indirectly, mediated by a representation, i.e. a sub-state that can play different roles in different contexts.

[3] They need not do it deliberately, or even be aware that they are doing it.

[4] Some of these ideas are explained in more detail in http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

[5] This is a more restricted notion than the notion of 'state' mentioned in philosophical functionalism, which does not presuppose the existence of needs of any kind.

Where many functional components form part of an integrated system we can talk about an *architecture*. In the cases that are of interest to us, these are generally virtual machine architectures. Different sorts of architectures will provide a basis for different classes of functional states and processes. What we normally call mental states and processes can be subsumed by these, including what we describe as different classes of emotions, as explained below.

2.2 Trivial and non-trivial needs and functions

Among the general class of physical systems, which have needs of the trivial sort described above, there is a sub-class, including biological organisms, that we can usefully described as *having needs related to survival and reproduction* and having components and states that can be said to have *functions* insofar as they serve those needs. We do not have space to offer a fully developed analysis of the notion of "function" here. However we can offer a sketch of an analysis that is probably compatible with a variety of more detailed analyses.

Needs of an entity can be defined in terms of requirements for survival or reproduction or, *recursively*, requirements for support for something that supports needs.

If some object would be destroyed by contact with water, then in this sense it has a need to kept away from water, and to have water kept away from it, since that is a requirement for survival. In a context in which water is moving towards it, it would have a need for a barrier to be created if that's the only way of preventing the water reaching it. Alternatively there could be a large disjunctive need, to have a barrier or a water-diversion or a means of transport away from the path of the water, etc. The existence of such needs does not imply the existence of parts or components or behaviours which have the function of meeting those needs.

Any physical object trivially has needs and parts with functions: including the trivial need to survive and components that trivially have the function of meeting that need. For instance insofar as a rock trivially has the need to remain whole, the molecular forces that hold it together serve that need, and to that extent trivially have a function. What makes it trivial is the fact that nothing more than the laws of physics and general features of solid physical objects are involved, and no special store of energy or any other resource has to be deployed in meeting that need.

In contrast, some physical objects have parts or states that serve needs such as survival in a less trivial way, because they are special-purpose mechanisms that serve those needs by doing different things in different circumstances, and using up resources in the process, e.g. chemical energy. Such parts or states non-trivially have the function of meeting those needs.

Moreover, the need may be temporarily present or absent, depending on the state of the object. In that sense an animal's legs meet the need of locomotion (which serves various other needs including finding food, mates, shelter, etc.) The legs have to produce specific behaviours tailored both to the current content (e.g. terrain that is rocky or smooth, horizontal or inclined, etc.) and the current need, and in so doing they use up energy. They therefore have a distinctly non-trivial function, compared with the molecular forces holding a rock together.

Parts of a system, or states and processes in a system, can be said to 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. Then, to say that

> *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 $P(X)$), which we can abbreviate as $F(O, X, C, N)$*

means that

> *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.*

As explained previously, bringing about states meeting the need N, amounts to bringing about certain conditions that are necessary for $P(X)$ to be true, for some predicate P. Then something serves or meets X's need N in context C if, in that context, it is sufficient for another state that is necessary for the truth of $P(X)$). I.e. meeting a need is being sufficient for something that is necessary. We can generalise this to being sufficient for a disjunct in a disjunction that is a necessary condition as a whole, though none of the disjuncts is. E.g. having a car meets the need to be able to travel to work even though going by car is not necessary for getting to work, though going by car, or going by bus, or walking or cycling or, ... is necessary.

Sufficiency, that is a guarantee of success, is often hard to achieve. In that case, a weaker way of serving or meeting the need is to make the necessary condition *more likely* to be true.

Relativisation to a context is important because the same need might be met by different sub-systems in different contexts. For instance an animal that needs to avoid approaching predators might use visual sensors by day and smell or hearing at night. An otter's need to move is met in different ways on land and in water. So different parts and different capabilities may have the function of producing motion in different contexts.

On our account, statements such as "O tends to bring about $P(X)$", "O preserves $P(X)$", and "O prevents things which make $P(X)$ less likely" are true by virtue of the existence of true counterfactuals involving O and $P(X)$. Thus, the larger and more diverse the set of true counterfactual conditionals about what would be done by O to perform function F if various things were to happen, the less trivial, or the more significant, the functional role of O is, and likewise the less trivial, the need that is met. So the need of a rock to stay in one piece is in that sense trivial compared with the need of an insect to move to where food is.

There are subtleties that we shall not discuss such as components that are not *absolutely* necessary because of redundancy in the system. Such components might become absolutely necessary if other components were damaged. Describing precisely which collection of true counterfactual conditional statements defines the function of a component can be a subtle and complex task though special cases like thermostats and homeostatic sub-systems are well understood.

Being able to serve a function by producing different reactions in the face of a variety of threats and opportunities requires a specialised design, including, for instance, (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, as indicated above, ascribing to them complex dispositional properties that can be analysed in terms of what would happen in various circumstances. For instance, in simple cases, what makes something a sensor that detects a shortage of fluid would be the way its effects on other parts of the system increase as fluid levels drop, so as to initiate or intensify fluid-seeking behaviour.

An object that has components which have such functions can be described as having (at least partial) self-control or self-organisation, which contrasts with being entirely controlled by the operations of external physical forces, like a marble rolling down a helter-skelter.[6] Having such functions is sufficient, though not necessary for having control: there might be a use for the notion of control in simpler cases. Note that the notion of a system exercising "self-control" does not presuppose an entity called "the self": it merely contrasts with control by *something else*.

Biological organisms contain myriad mechanisms performing very specific functions, some of them requiring only restricted actions, e.g. homeostatic functions, others requiring far more complex and varied actions, e.g. tissue repair, locomotion, digestion. Many of the homeostatic functions of components of biological systems are in that sense relatively trivial, though they are all more complex than a typical artificial thermostat. Components that perform relatively trivial functions may, together, form a system that performs very sophisticated functions in complex and diverse ways.

Saying that parts or aspects of a system have functions does not imply that the whole system has a function (though a robot may have a function for its designers, and individual organisms may have functions in a family or a colony). Neither does our use of the word "function" presuppose any kind of designer, though evolution could be loosely described as a "designer" of organisms, since it produces instantiated designs. Other designs are produced by processes like growth, learning, or geological processes that produce rivers which have functions for ecosystems. Something produced by random accidents can have a function in the sense defined above.

The notion of "function" has received much discussion in the philosophical literature. We do not agree with those (e.g. (Millikan 1984)) who argue that the notion has to be explicated in terms of history, especially evolutionary history, since the causal relationships summarised above suffice to support the notion of function defined here, independently of how the mechanism was produced.

Compare the argument in (Young 1994) that a robot not produced by evolution could have various mental functions, such as perception, learning, reasoning. In talking about whether something has certain functions we are talking about what it would or would not do in various conditions, not about its origins. In that sense the functions of an object or a component or aspect of an object are among its *dispositional properties*, which include things like brittleness, elasticity, electrical resistance, rigidity, etc.

Familiar physical dispositional properties, e.g. modulus of elasticity, are easily defined in terms of invariant relationships between physical measures (e.g. between force and change in length). This summarises a large collection of counterfactual conditionals about how changes in one value (e.g. force) cause changes in another (e.g. compression). In contrast, most biological dispositional properties are much harder to specify. Far more complex formulae are required to specify the dispositional properties of legs that justify saying that their function includes

[6]The notion of "autopoesis" of Maturana and Varela (Maturana & Varela 1980) analysed in (Boden 2000) is a special case of this general notion of self-control.

providing locomotion. Various programming languages provide formalisms for specifying such relationships, though it is not easy to do, and we may need to invent new formalisms.

2.3 More on information-users

The previous section described mechanisms whose functions involve production of information that is used in deciding what to do, i.e. sensors. We call objects containing mechanisms with such functions *information-users*.[7] These entities (a) have needs, e.g., in non-trivial cases, their needs might be necessary conditions for survival, for reproduction, for growth, for development, and indirect needs derived from those needs, (b) have a store of energy available to mechanisms that can use the energy to meet those needs (usually mostly chemical energy in the case of organisms), and (c) are able to acquire information on the basis of which they select ways of using energy to meet their needs. For this, as explained above, they have sensors for needs and sensors for other changing aspects of the context. 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 make a distinction between sensors that act as *need sensors* for a machine or organism and those that act as *fact sensors* for it. 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. (We define these two notions in section 4.2.)

In more complex cases perceptual sub-systems may include mechanisms for checking and improving accuracy, because the need for truth and accuracy exists insofar as they are useful in supporting other needs, such as the need for survival or reproduction. So a sub-system that in itself has a non-motivational function may have components that serve the need of that sub-system to avoid error and imprecision. These components will include need-sensors, for instance sensors that detect a need for an eye to re-focus. This would be a side-effect of the detection of something that might be edible, or a useful weapon, and might have to be grasped quickly. So in some cases, existence of mechanism $M1$ concerned with producing belief-like states may lead to evolution of a sub-mechanism $M2$ producing desire-like states

[7] Some artefacts are 'derivatively' information-users because they don't themselves have needs (as defined above) but they have the potential to play the same roles as things that are parts of systems that do have needs. For instance we can talk about the functions of a spare thermostat that happens still to be in a box, and not installed in a central-heating system where its functions would be used. Even though the thermostat is not connected, it may still be detecting and responding to temperature changes, e.g. using a bi-metallic strip. It is an information-user because of what it could do if installed. So it is derivatively an information user. Perhaps we should call it a *potential* information user. In that sense, any object could be a potential information user.

that are capable of supporting $M1$ in its function. Nevertheless $M1$ itself does not thereby become a need-sensor.

As should be clear from the above analysis, saying that something has mechanisms that serve its needs involves a complex collection of implicit claims about what those mechanisms would or would not do under various conditions when those needs are threatened or when opportunities are available for meeting or anticipating the needs. These are claims about the truth of complex and varied sets of counterfactual conditional statements. And in the case of non-trivial functions meeting non-trivial needs, the collections of counterfactual conditionals may be very large and complex. Usually this is because we are investigating a object with a complex architecture.

Our notion of an "information-user" refers to the special case where some of the mechanisms acquire, manipulate, store, compare, transform, combine, and use information, for instance in detecting the presence of needs, opportunities and constraints, in deriving consequences, in resolving conflicts, forming and executing plans, learning, etc. Information-users can vary enormously in the variety and complexity of the processes that occur in them. Many different architectures are possible.

2.4 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: in many cases 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 are all virtual machines) 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 (in that, e.g., they cannot be observed using the usual techniques of the physical sciences, and cannot be described using the language of the physical sciences) but are *implemented* in physical mechanisms, such as brains.[8]

Information-processing virtual machines can vary in many dimensions, e.g. the number and

[8]The attribute "virtual" here is in contrast to "physical", i.e., a running "virtual machine" is an abstract machine containing abstract components related to the software defining the virtual machine, which may be capable of running on different physical machines. Although the idea is now very familiar, explaining exactly how these virtual machines are related to physical machines is a non-trivial problem, e.g., see (Scheutz 1999, Scheutz 2001*a*, Sloman & Scheutz 2001), and http://www.cs.bham.ac.uk/research/cogaff/talks/#super . In particular, in virtue of being referred to in sets of true counterfactual conditional statements, virtual machine states can have causal powers, for instance the power to deliver email or to detect and prevent access violations.

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 very 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. This indirectness involves the use of an intermediate state that 'represents' the need, and is capable of entering into a wider variety of types of information processing than simply triggering a response to the need.

An enduring intermediate state could have uses like (a) allowing different sensors to contribute data for the same need, (b) allowing 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) allowing alternative responses to the same need to be compared, (d) allowing conflicting needs to be evaluated, including needs that arise at different times, (e) allowing actions to be postponed while the need is remembered, (f) allowing associations between needs and ways of meeting them to be learnt and used, etc.

In such cases we may say that in addition to having needs the system has *goals*. Having a goal, in this technical sense, 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*. In the next section we note that there can also be "buggy" derivations of representations of needs.

Evolution also produced organisms that in addition to having sensors related to particular needs also had sensors that produced information that could be used for varieties of different needs. For instance knowing where a fig tree is could be useful whether there is a need for food or a need to climb out of reach of a predator. Information structures that have that kind of diversity of function could be called 'percepts' or 'beliefs' depending on their precise causal relationships and whether they can endure beyond the sensor states that produce them.

2.5 Direct and mediated control-states, and representations.

We have distinguished two sorts of control-states: (a) the kind of *direct* control-state in which either a need or contextual information is sensed and which immediately tends to initiate action appropriate to the need in that context, and (b) a *mediated* control-state in which something corresponding to a need or to the current context exists that helps to produce whatever actions would serve that need in that context.

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*b*). 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.

For people who are familiar with recent developments in computing and AI none of this will be news: we have already learnt a great deal about varieties of information-processing architectures and their costs and benefits. Moreover, this gives us a framework for discussing how it is possible for some machines to be information-users, i.e. able to acquire and manipulate information with semantic content, without having states like explicit beliefs, desires or intentions, because their architectures are too primitive.

We can also now begin to see what sorts of additional architectural features can support states with more of the characteristics associated with concepts like "belief", "desire", "intention" and so on, and in doing this we do not need to adopt what Dennett calls 'the intentional stance' (Dennett 1978), which is based on an assumption of rationality. Rather we can base these concepts on an architectural specification as suggested above – requiring only what Dennett calls 'the design stance', as explained in (Sloman 2002*a*).

However, we lack a systematic overview of the space of relevant architectures, and it is also likely that the more we learn about the naturally occurring architectures produced by evolution, the more we shall 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.

3 Emotion construed as a special case of affect

3.1 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. For instance, many researchers treat all forms of motivation, or all forms of evaluation, or all forms of reinforcing reward or punishment, as emotions. It is not always realised that the concept of "emotion" is a special case of the more general concept of "affect" which covers other things besides emotions, including moods, attitudes, desires, dislikes, preferences, values, standards, intentions, etc., the more enduring of which can be thought of as making up a "personality" – as suggested in (Ortony 2002) and in the chapter by Norman *et al.*.

The current confusion is summarised aptly in (Delancey 2002) [9]

> 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.

[9] 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.

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 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.

Moreover researchers studying emotions do not all have the same research goals: some are primarily concerned with understanding natural phenomena, in humans and other animals; some seek to create artefacts which themselves have genuine emotions; while still others are solely concerned with producing useful artefacts, e.g. synthetic entertainment agents, sympathetic machine interfaces, and the like. For the latter it does not matter so much exactly what emotions in humans or animals are, as long as humans react appropriately to the artificial agents (Bates 1994, Picard 1997, Breazeal 2002).

Arguing about what emotions *really* are in the context of so much confusion and ambiguity is pointless: it is a "cluster" concept 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". This is also a difficult notion to define precisely, especially if the notion is to be applicable not just to humans but to a variety of types of animals and robots. Conceptual analysis techniques developed by philosophers can help us with the task, especially when supplemented by the design-based approach sketched above.

3.2 A multitude of concepts – and relations between them

"Emotion" is but one of several affective concepts used in everyday language, forming part of so-called folk psychology. Related affective phenomena that we refer to in thinking and talking about ourselves and others include desires, likes, dislikes, drives, preferences, pleasures, pains, values, ideals, attitudes, concerns, interests, moods, intentions, etc. Other kinds of 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.

No theory or model of emotions can be adequate that does not include relations between emotions and other kinds of mental states and processes (both affective and non-affective) that either initiate, or occur in, or are produced by the affective ones. As we shall see in section 5, in different parts of a mental architecture different sorts of causal relationships can occur between such states and process.

A good theory will also have to explain both the differences and the commonalities between affective and other processes. For example, concepts can play a role in desires and emotions as well as in percepts and beliefs, and any theory that does not link existence of emotions and other affective processes to use of concepts is to that extent inadequate. A belief that apples can be eaten, a desire to eat an apple, annoyance that someone else has eaten your apple, concern that the apple you have just eaten may have been poisoned, all involve use of the concepts "eat" and "apple", which require the agent concerned to have an *ontology*, i.e., a set of concepts for

sorts of things that can exist or occur in the world. This ontology would include kinds of actions of which "eating" would be a special case, and varieties of objects of which "apple" would be a special case. Furthermore, this ontology, and the associated concepts, would be common to many mental states. So a deep theory of the commonalities in human mental processes must address general questions about how agents acquire and use ontologies (systems of concepts). A theory of human emotions which merely states which brain processes are involved and which behaviours are produced, when emotions occur, would fail on this test.

There may, however, be some types of agents (purely reactive agents, for instance, including insects) whose architectures do not support the use of such ontologies and concepts. Therefore, insofar as they have emotions, percepts, beliefs, or desires (or merely drives?) these mental states and processes will be of a different kind and may not have contents that are expressible using the semantic apparatus for describing "propositional" contents of human mental states. Some of these states may be simple control states (such as a "hunger level", which is linked to the value of a homeostatic variable), where some might say that the distinction between emotion and motivation collapses (e.g., see the chapter by Norman *et al.*). Nevertheless we shall offer a different conceptual framework, where the distinction does not collapse: if emotions are defined in terms of one process interrupting or modulating another, so that it no longer proceeds "normally".

Moreover, some agents (including humans) have complex, hybrid information-processing architectures involving a variety of types of sub-architectures. In that case 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.

Further, if processes in several architectural "layers" are active simultaneously, as is often the case in humans, the resulting emotional states may be very complex and very varied. This is why, in previous work listed in the references, we have distinguished *primary*, *secondary*, and *tertiary* emotions,[10] 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 events) intrinsically do. For this, they need a deliberative layer. What we call *tertiary emotions* (such as self-blame) need, in addition, a reflective 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 5.3). 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 the case of anger in Sloman (1982)).[11]

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

[11] Minsky's draft book *The Emotion Machine* presents a similar view. It is available online at his web site http://www.media.mit.edu/~minsky/

3.3 Do we really know what we are talking about?

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 or are likely to have emotions is a well-defined question. Likewise some researchers will ask whether unborn, or new-born infants can have emotions, as if that were a factual question. However, if there really is no clear, well-defined, widely understood, concept of 'emotion' it is not worth attempting to answer the question about robots, or the question about infants, until we have achieved more conceptual clarity.

We can construe the claim that there are no *right* ways to divide phenomena into emotions, moods, etc, as analogous to the claim that there is no *right* way to divide up a part of the earth's surface into countries or continents. Nevertheless, before making such divisions, for whatever purposes, it is necessary to know what there is to divide. Likewise, when debating how to classify mental phenomena we need a way of understanding what philosophers have called the "logical geography", namely the variety of possible phenomena that can exist and their relationships. In order to avoid an infinite regress we must assume that descriptions are available that do not have the problems of ambiguity and indeterminacy, or at least have them to a lesser degree. In the case of 'logical geography' that includes mental states and processes we can start from language that describes information processing mechanisms and architectures.

Using the ideas about varieties of architectures presented in section 2.4, we propose that this understanding of what there is to classify can be based on an analysis of types of information-processing architectures that can exist in animals and machines, and the varieties of states and processes that can occur in different architectures. This collection of possible states and processes defines the "logical geography" which we can divide up in different ways depending on what our purpose is, e.g. merely explicating folk concepts, or rationally reconstructing them, or extending and refining folk psychology towards a deeper more general and precise theory. The last is our goal.

So by defining various types of mental states and processes (e.g. believing, perceiving, attending, desiring, having emotions) in terms of the kinds of architectures in which they can occur and what those architectures can do, we replace ill-defined pre-theoretical concepts with more precise architecture-based concepts (as explained in (Sloman 2002*a*) and other papers in the bibliography). We shall try to illustrate this in what follows, building on our previous work developing this theme. We will start with some remarks on conceptual analysis (discussed in more detail in chapter 4 of (Sloman 1978), based on (Austin 1956)).

3.4 Empirically-guided conceptual analysis

A proposed new analysis of a concept can be empirically-guided in two different ways which may sometimes be in conflict. First of all, the analysis should be guided by how the concept is currently actually used, though there may be disagreements about this if usage is inconsistent or vague. Second, in the case of concepts to be used by scientists and engineers, the analysis should be guided by the requirement that concepts be useful in formulating explanations and predictions, or more generally that they should be consistent with the best available scientific theories. The first requirement is consistency with linguistic usages (which may not be consistent themselves!), whereas the second is concerned with facts about the subject matter

referred to.

For instance, an analysis of the concept denoted by a term (e.g., "fish") should be extensionally congruent as much as possible with actual usage of the term, at least as regards generally agreed *clear* cases, for instance implying that trout, bass and mackerel are fish, and that elephants, eagles and ants are not fish.

On the other hand, where current usage is confused and inconsistent, a useful analysis may diverge from current usage, but be empirically superior, for instance if it is useful for explanation, prediction, or control. Scientific value often overrides current usage, resulting in a change of extension. The concept of "fish" that we find useful today excludes whales, even though at one time whales were considered fish. A good new version of an old concept will often turn what were previously regarded as difficult borderline cases into clear cases. E.g. water-snakes, porpoises and dolphins, will definitely not be fishes. Sometimes this happens because the new theory-based concepts use deeper criteria than the older pre-theoretical versions, for instance if the older version uses observable habitat, shape and behaviour, while the new concept is based on internal structure and bodily function, or, in the case of organisms, on evolutionary origin and genetic makeup.

Similarly, although we take current usage of "emotion" and other labels for mental states and processes as points of departure, we do not require a useful analysis to match up *precisely* with current usage, since current usage is so inconsistent and confused that a precise match is impossible. We can be *guided* by current usage without being a *slave* to it.

As with "fish", consistency with current usage can be restricted to *clear cases*: the set of things that "clearly" fall under the term being analysed, and the set of things that clearly do not. For example, we could put *pleasure at solving a difficult problem, anger at being overtaken in traffic, fear of being caught by a predator* in the set of clear instances of emotions, and *remembering what you had for breakfast, knowing there's someone at the door, deciding to eat the asparagus instead of the potatoes* in the clear set of non-instances of emotions.

However, we must allow for the possibility that not everyone will agree about which cases are clear. For instance, some will say that there are sensory pleasures, e.g. enjoying the taste of good food, that clearly do not involve any emotional state, merely an evaluation of the experience, whereas others will regard that as a clear case of a positive emotion. At another extreme, some may think that every decision must involve an emotion to provide the motivation, for instance deciding which type of sandwich to select for lunch.[12]

Moreover, the clear cases can be construed differently by different people. For instance two scientists may agree that someone is embarrassed but disagree on the implications: one may say that the assertion refers only to observable behaviour and facial colouring, whereas the other takes it to refer primarily to mental states and processes. Our analysis will favour the second, but in the context of a theory about the nature of mental states and processes that may not be acceptable to everyone who wishes to talk about them.

Such disagreements and inconsistencies in ordinary usage may force us to come up with several alternative technical definitions, each of which is more precise than ordinary usage, though all of them only partly conform to ordinary usage. For instance, some might use the phrase 'cold emotions' to cover mild evaluations and 'hot emotions' for intense states that produce disturbances of behaviour. (Compare (Cohen 1962) on "precisification".)

[12] A reader of an early draft suggested that 'hunger' was an emotion, whereas we would regard that as merely a desire-like state, except, for example, in the case where the hunger is so intense that it disrupts thinking in a manner that is hard to control.

3.5 Design-based conceptual analysis

How can emotion concepts and other concepts of mind be identified, apart from listing positive and negative examples? 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)).

At another extreme, some researchers regard emotions as experientially defined, and attempt to use introspection-inspired descriptions of what it is like to have emotions in order to define them — a notorious example being Sartre's analysis of having an emotion as "seeing the world as magical", in (Sartre 1939). Novelists often use notions of emotions and related states defined primarily, as pointed out by (Lodge 2002), 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, a feature of what we refer to as *tertiary* emotions discussed in sections 5.4 and 6.1.

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 other people and they use concepts that are implicitly defined in terms of what can happen in that architecture. This conjecture is controversial, but even if it is correct, for purposes of scientific explanation the naive architectures of folk-psychology need to be replaced with deeper and richer explanatory architectures, which will support more precisely defined concepts. Note that this does not mean that implicit naive theories about architectures are necessarily wrong or that they are wrong about all the details of the architecture, only that they are wrong about some, and those theories need to be clarified and corrected. Some features of naive architectural theories may be useful precursors of deep scientific theories — as happens in most sciences.

The work in our group has aimed, over several years, to show how "architecture-based" concepts can extend and refine our pre-theoretical "cluster concepts" in directions that make them more useful as a basis for expressing scientific questions and theories, and for specifying engineering objectives.

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 2000*b*, Scheutz & Sloman 2001).

4 Varieties of Affect

It may be that many of the people who have recently become 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". What are affective states and processes? In this section, we attempt to explain the intuitive affective/non-affective distinction in a general way.

Like the concept "emotion" the concept "affect" lacks any generally agreed definition. However, the latter concept appears not to be used as much in the scientific literature (especially the recent AI literature) so perhaps that gives us more freedom to recommend definitions. Nevertheless we wish to avoid the charge of usurping a familiar word or distorting its meaning, so, when pressed, we use the label "desire-like state". This can be contrasted with a "belief-like state", a particular variety of non-affective state. Both types of states are defined below. Both of are required in information-users such as biological organisms or robots. (This terminology was proposed in (Sloman 1993, Scheutz & Sloman 2001), though we have refined the definitions here.) We show in section 4.2 that there are additional classes of states with semantic content in organisms or machines with more complex architectures.

4.1 Varieties of control-states

Previously, in section 2.5, we introduced a notion of a control-state of an individual which is a state that has some sort of control function which may include preserving or preventing some state or process. In other words, an individual's being in such a state involves the truth of some collection of counterfactual conditional statements about what the individual would do if any of a variety of things were to happen. In section 2.3 we introduced two sub-classes: belief-like and desire-like control-states. We also claimed that these could be either unmediated or mediated control states, the latter involving an explicit representation with more possible functional roles than unmediated control states.

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 folk notion "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 David Hume[13] – that all action may be based on many beliefs and derivatively

[13] "Reason is, and ought only to be, the slave of the passions, and can never pretend to any other office than to

affective states, but must have some intrinsically affective component in its instigation. This is related to the notion of desire-like state that we are referring to. No matter how many beliefs, percepts, expectations, and reasoning skills you have, you will not have any basis for doing one thing rather than another, or for doing anything at all, unless you have at least one desire or desire-like state. This state could be positive or negative, i.e., an preservation of an action, object, state or event could be desired or its absence or non-existence could be desired.[14]

However, there is another use of "affective" which 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 the another state, namely that of being jealous.[15]

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 3. So cognitive mechanisms are required for many affective states and processes, and we should avoid terminology that drives a wedge between affective and cognitive processes.

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

We can now introduce our definitions[16]:

- 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, and sensory information states.

When primitive sensors provide information about some aspect of the world, that is because the information provided varies as the world changes. (Another example of sets of

serve and obey them" in (Hume 1739)

[14]There is an ambiguity here between the claim that there is no *reason* to act if only belief-like states are present and the claim that there is nothing that will *cause* action if only belief-like states are present. The former presupposes that we are referring to rational agents, as assumed in Dennett's "design stance" and Newell's "knowledge level" (Newell 1990). Our concern is more general and more basic, for it includes organisms and machines that are incapable of being rational or irrational, e.g. microbes and insects. So our version of the Humean point is the purely causal one.

[15]We wish to leave open the possibility that our approach might be used eventually to *explain* experiential affect, and therefore should not presuppose it. See (Sloman & Chrisley 2003)

[16]These formulations conform to the spirit of the definitions offered in our earlier publications, but avoid some objections and complications.

counterfactual conditional statements.) In that case, 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 some cases described in section 2.3 the state B will include additional 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 we can say that 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 it is often regarded as non-desire-like states can be seen as including a special subclass of the more general notion of desire-like state.

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's 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:

- Change in World $\rightarrow$ Change in beliefs
- Change in Desires $\rightarrow$ Change in World

where "$\rightarrow$" means "causes (or tends to cause) change in", and "World" can include states of the organism.

Belief-like and desire-like states may exhaust the variety of possible states with semantic content in *simple* organisms and machines, but in more complex architectures there are sub-systems providing more sophisticated capabilities whose states 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, or some kinds of artistic activities. Such activities depend on possession of a collection of concepts, or grasp of a language, and other resources that can be used in producing belief-like or desire-like states, or which can be used in processes generated by them.

The development of concepts, linguistic forms, and other such representational resources that are capable of being used in belief-like and desire-like states can also produce capabilities to generate states that are themselves neither belief-like nor desire-like. In particular, as pointed out in (Sloman 1993) there are also *imagination-like* states involving possibilities that are thought about but not believed or desired (which could also be called "supposition states") and *plan-like* states in which the individual has constructed a possible plan of action, but has neither adopted nor rejected it.

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). Later we explain that what we call secondary emotions also use such mechanisms.

4.3 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 and elaborated on in different theoretical frameworks. We could also add further details such as analysis of the notion of 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. has in different parts of the system. Here, we mention only three important points.

Our distinction in section 2.5 between direct and mediated belief-like and desire-like states corresponds 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 intact.

Another way of phrasing this difference is that 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.[17]

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 S state 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.

[17]The derivative states are described as "rational" because of the processes that produce them. This concept of rationality refers to the "internal" processes producing the state, not to the "external" facts that determine success or usefulness of the result. Reasoning can lead to disastrous mistakes.

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 (and to that extent are rational). The distinction between derivative and associative affective states will later be of assistance when attempting to distinguish between different kinds of emotions.

4.4 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 that are learnt to be 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.

4.5 Complex affective states

Depression would seem to be a counter-example to the analysis of positive and negative affective states offered above.[18] 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

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

has the function of making the system engage in activity that tends to remove or diminish the state 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,[19] 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 the state P, whose function is to reduce activity and the state N, whose function is to increase the 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.

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 a suitable method 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.

4.6 Varieties of affective states and processes

So far we have discussed belief-like and desire-like states with semantic content. The latter are intrinsically affective states, whereas belief-like states may or may not have mechanisms driving them that produce affective states concerned, for instance, with maintaining the truth or precision of their information content.

Our account provided a very general characterisation of "affective state", which, it should be clear, includes mild and strong likes and dislikes, desires aversions, goals, and intentions. We

[19] This would be an example of a mood.

think it covers a much wider range of what are normally regarded as affective phenomena, but this is not the occasion to demonstrate that.

We have also shown that *complex* affective states can exist in which two or more affective states interact, where one alters the effects of the other, for instance temporarily suppressing another. In that case the second state still exists, but does not have its normal consequences. This may not be possible for very simple organisms, but is common in humans, for instance in long term grief, or love of one's family which continues even though other concerns temporarily have one's full attention, e.g. driving a car safely.

Theories of emotion and other mental states and processes which do not allow for such concurrency will to that extent be too simple. We will offer an architecture-based analysis of notions of emotion based on this general idea, including the special case of an "alarm" sub-system which can rapidly interrupt or modulate a wide range of processes.

All of this assumes that organisms or machines capable of having such complex states have an architecture in which different sub-mechanisms co-exist and interact. In simple cases these could be physical mechanisms, such as a domestic thermostat controlling a heater and an external temperature sensor causing the temperature setting on the thermostat to vary.

More generally, however, we need to include architectures for *virtual machines* as discussed in section 2.4. We assume that long before human engineers discovered the need for virtual machines to provide adequate flexibility in sophisticated control systems, biological evolution produced more complex virtual machine architectures than any we have yet devised, and implemented them in brains with more varied and complex physical mechanisms than any we have so far devised. Although it would be desirable to explain exactly how such virtual machines are implemented in brains, that is far beyond the scope of this paper.

In the remainder of this paper we shall present some ideas about how to think about a wide class of architectures relevant to humans, other animals, and robots, and will try to explain how various architecture-based concepts of emotion can be defined within that general framework. We will also sketch a particular architecture, which seems to cover some of the main requirements for human-like systems, and which is capable of supporting a wide variety of types of affective states found in humans, including what we have previously[20] called primary, secondary and tertiary emotions.

Within the context of a 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 are experienced?)
- whether they in turn can be or are suppressed.
- whether they can become dormant and then be re-awakened by new triggers, or by removal of something that suppresses them.
- what sorts of external behaviours they produce,

[20](Sloman 1998, Sloman 2002*b*, Sloman & Logan 2000, Sloman 2001*a*)

- how they affect internal behaviours, e.g. remembering, deciding, attending, deliberating, dithering, etc.
- in particular, whether they produce second-order affective states (e.g. being ashamed of being angry),
- what sorts of conceptual resources they require, e.g. whether they use an ontology including other intelligent agents or not.

All of this can be further refined by the taxonomy in (Ortony, Clore & Collins 1988). Like the distinctions in that taxonomy many of the cases listed above would be inapplicable in 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.

5 Architectural constraints on emotion and 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. As a first step towards addressing this problem we propose the Cogaff Schema as partially defining a high level ontology for components in a wide range of information processing architectures.

5.1 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 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.[21] Examples might be seeing another as happy, sad, attentive, puzzled, undecided, angry, looking to the left, etc.

Similar comments could be made about the possibility of layers of abstraction in an action system being related to layers of processing in the central mechanism.

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 and obtain a first crude sub-division.[22]

[21] 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 & 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.

[22] The CogAff schema is described more fully in (Sloman 2002*b*, Sloman & Logan 2000, Sloman 2000*c*, Sloman 2002*a*, Sloman 2001*b*, Sloman & Chrisley 2003)

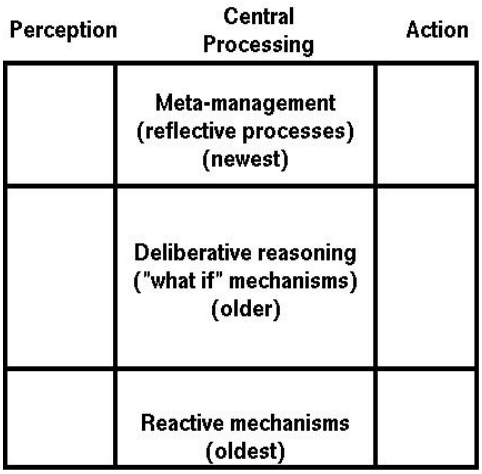

Figure 1: The CogAff schema: two kinds of architectural sub-divisions superimposed. Many information flow-paths between boxes are possible.

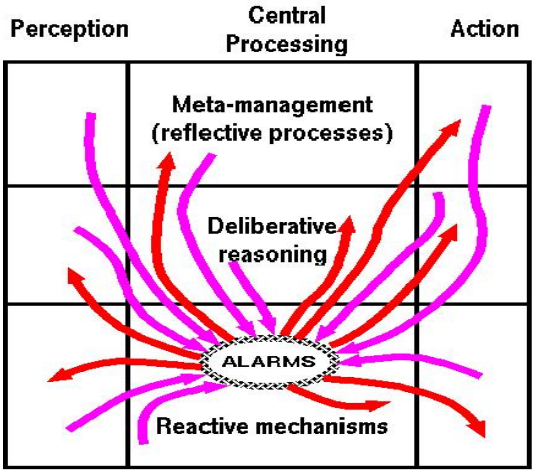

Figure 2: Elaborating the CogAff schema to include reactive alarms – possibly many varieties with different input and output connections.

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 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, Newell & Rosenbloom 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*. An example is the contention-scheduling architecture (Cooper & Shallice 2000).

5.2 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 will get back to this in section 6.1).[23] 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 the alarm system, for instance if a process of planning or reasoning

[23]Note that the use of the word 'alarm' is merely intended as a mnemonic aid. The normal use of the noun 'alarm' is more specific than the collection of functional roles we are postulating for alarm-mechanisms.

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 lead to a *unique* class of emotion concepts, although each instance of the framework can.

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.

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

People produce layered diagrams indicating different architectural slices through a complex system. However close textual analysis suggests 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 includes mechanisms that we have assumed are 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). By the same token, what we call 'reactive' mechanisms may include processes that are part of all three layers in the sense that everything ultimately has to be implemented in purely reactive systems.

Most importantly, their reflective layer only receives pre-processed perceptual input, but does not 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.

Another problem with layered architectures is that even when people use the same labels for their layers they often interpret them differently: e.g., some people are happy to 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.

5.4 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 sub-system.[24]

The layers in H-CogAff are also intended to mark significant evolutionary steps. For example, the architecture of H-CogAff implies 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 2001*b*))

5.5 Architectural presuppositions

Our conjectures in sections 3 and 5 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

[24] In (Sloman & Chrisley 2003) we show how it also accommodates important aspects of human consciousness and enables us to say what qualia are (and why robots will have them!).

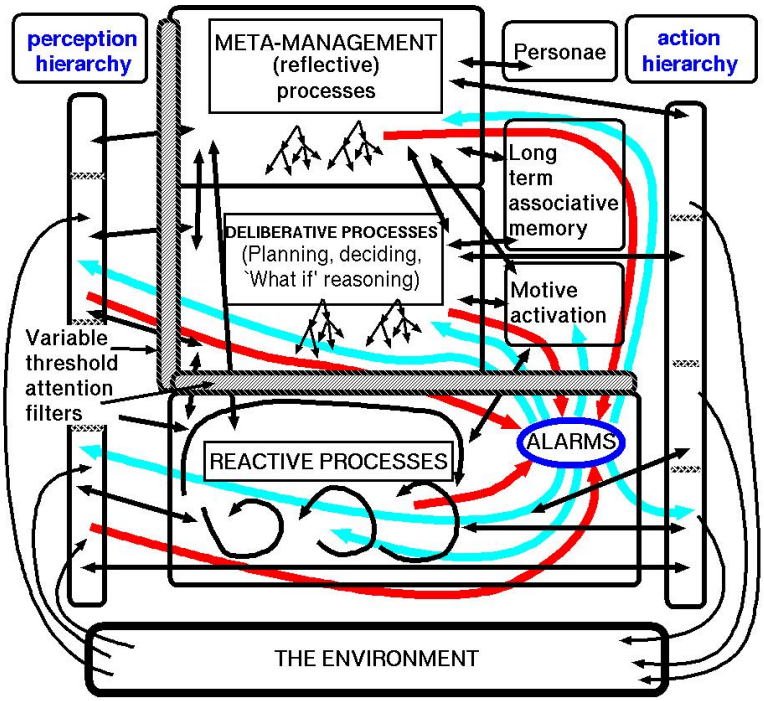

Figure 3: The H-CogAff architecture. The central layer relates to different functional layers in perception and action hierarchies. Not all possible links between boxes are shown. Meta-management may be able to inspect intermediate states in perceptual layers, e.g. sensory qualia.

of our use of concepts like "emotion", "attention", "learning", etc.[25]

In that case, scientists, engineers and philosophers who ask questions about, state theories about, or propose practical goals involving emotions and other mental states, are likely to 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").

5.6 Are there basic emotions?

We suggest, as pointed out in (Turner & Ortony 1992), that this question leads to 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 looking in the wrong place. A much better strategy is to look for a collection of physical processes in physical mechanisms that implement a wide variety of chemical processes.

Likewise, with emotions, it is better to look for a collection of processes in information-based control systems (a mixture of virtual machines and physical machines) that implement a

[25]The kinds of indeterminacy found in cluster concepts are discussed in (Sloman 2002*a*) and (Sloman 2001*a*), though a paper in preparation attempts to give a more thorough analysis.

wide variety of emotional (and other affective) states and processes, rather than trying to isolate a clear specifiable subset of emotions.

5.7 Where to begin?

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

Since we do not know enough yet to get such specifications right the 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.

These alternative architectures should be inspired, where appropriate, 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 or which require something not yet included in our schema. But we are not restricted to what is biologically plausible. We can also consider architectures for future possible robots.

6 An example of architectural analysis of emotion and affect

We have already published and are continuing to develop some suggestions for extending folk-psychological architectures in the framework of the CogAff schema (figure 1), which supports a wide variety of architectures. An example is our proposed (but still developing) 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.

In the following, we shall take 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 robotics who intend to study the utility of emotional control in artifacts.

6.1 Towards a generic definition of “emotion”

For some time we have been exploring the suggestion that the highest level architecture-based (see below) definition of emotion that is likely to be of use in scientific contexts and also consistent with a wide range of non-scientific uses, is based on the assumption that in any

information-processing system there are temporally extended processes sometimes that take more time than is available, because of the speed with which external events occur. E.g., an operating system is trying to write data to disk, but the user has already pressed the eject button before the data was completely written. In that case it may be useful to have a process, call it "watch dog", that detects such cases and interrupts normal functioning, producing a rapid a default response that normally avoids some danger or grasps some short-lived opportunity. We have previously called the mechanism implementing the "watch dog" process "alarm mechanism" (figure 2).[26]

Because of the need for speed, this sort of alarm mechanism will generally be less sophisticated in its reasoning, and less likely to produce a good response than the "normal" mechanism (which it interrupts and overrides) would have produced if it had had sufficient time (and information) to complete its processing. (There may also be cases where there just is not enough information for "normal" processing to produce a good solution, so the role of the alarm mechanism may be to terminate attempts to find the information, etc.) In the case of the disk, the alarm might store the file in special location in memory or on the hard drive to protect it from being erased and notify the user that the disk write operation failed.[27].

Alarm systems can also be generalised to cases where the alarm mechanism does not interrupt, but instead *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). In any case, 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.

Note that this architecture-based notion of emotion (involving something actually disrupting or modulating "normal" processing, or a (possibly temporarily overridden) tendency to do so) falls under the very general notion of "affective" state or process as analysed in §4.2. 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,

[26] Although our diagrams display a single "alarm" module (located within the reactive layer) with lots of arrows going into and coming out of it connected to many different parts of the system, there is no implication that there has to be a unitary alarm mechanism. On the contrary, there are probably many specialised alarm systems, including, for instance, the blinking reflex that protects eyes.

[27] This idea of a "normal process" and an interrupting, possibly simpler, sometimes much faster, process that can interfere with or even terminate the normal process is not new, but can already be found in Simon's proposal for architectures for systems with emotions and motivations (Simon 1967), though we do not think his proposals were sufficiently general. The idea is also consistent with the etymology of the word "emotion" reflected in the English phrase "being moved". That by itself, of course, is only an added bonus, but does not prove anything

e.g. a single-celled organism, it is not clear that 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, though often it will be.

Previous papers (e.g. (Sloman 2000*a*, Sloman 2002*b*)) illustrated this 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, Sloman & Beaudoin 1996)) as states and processes that involve disruption of, or a disposition to disrupt, attention-control processes in the meta-management (reflective) system. That is just a very crude, inadequate, first draft high level subdivision. (Some readers may recognize Le Doux's subdivisions as mapping onto the distinction between the first case and the rest.)

Within the framework of an architecture as rich as H-Cogaff many more subdivisions are possible, including sub-divisions concerning different time-scales, different aetiologies, different sorts of semantic content (when there is semantic content) etc.[28]

We could argue about the relative merits of architecture-based definitions of emotion and the more common definitions of emotion in terms of externally observable or measurable behaviours or physiological changes. For example, the second analysis would link up more closely with older philosophical analyses, while the "alarm"-based notion is easier to apply as an engineer or scientist studying working systems. Perhaps our ordinary notion encompasses aspects of both, in which case arguing about which definition is "correct" is pointless. People who wish to use alternative definitions may do so as long as they understand both what they are doing and what they are not doing, and make things very clear to their audience.

6.2 An architecture-based analysis of "being afraid"

It is interesting to note that specific emotion concepts (e.g. fear, joy, disgust, jealousy, infatuation, grief, obsessive ambition, etc.) do not seem to be less indeterminate than the general concept of "emotion"? In English, for example, "fear" and "afraid" cover a huge variety of 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 an architecture (i.e., for an architecture

[28] Some of this overlaps with the categorisation in (Ortony et al. 1988).

to be able to instantiate them). 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 might involve perception, but do not depend on them to cause their instantiation. E.g., a memory recall process that retrieves the fact that the proof of Goldbach's conjecture was published might itself be sufficient to instantiate the state of being afraid that there the proof has a hidden flaw. One might even argue that this state can never be instantiated by perceptions, but rather only by virtue of deliberative processes (which may or may not include representations of perceptions).

Furthermore, the above states all 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).[29]

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 – note that state 2 then is a dispositional state in that the organism is always in it by virtue of the associative connection between large vehicles and state 3, which will be instantiated upon perceiving a large vehicle alone, 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 states, 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

[29]How this kind of mechanism is realized in neural circuitry has been extensively researched by LeDoux and colleagues (LeDoux 1996)

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, fear, grief, pride, jealousy, excited anticipation, infatuation, relief, various kinds of joy, schadenfreude, spite, shame, embarrassment, guilt, regret, delight, enjoyment (of a state or activity) etc.[30] Note that the same kind of analysis can also be applied for other non-emotional, affective states such as attitudes, moods, states like surprise, expectation, and the like.

7 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

[30] So far, detailed architecture-based analyses have been developed for *grief* (Wright et al. 1996) and *anger* (Sloman 1982).

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.

7.1 Do robots need emotions and why?

One of the questions researchers interested in designing artificial emotional agents or robots have to address is whether there is any principled reason why their robots would need "emotions" to perform a given task, even if it were possible to clearly and unambiguously defined the notion of emotion.[31] However, before we can consider in what sort of circumstance artificial agents should have emotions, we have to address the more general question whether there is any task that cannot be performed by a system that is not capable of instantiating emotional states (e.g., as defined above) needs to be addressed.

The answer to this question is certainly non-trivial in the general case. For simple emotional control systems, one could argue that it is always possible to define a finite state machine, which has the exact input-output behavior of the emotional system, but does not instantiate any emotion. I.e., the architecture of the finite state machine is not capable of defining states that satisfy the requirements for being emotional according to our definition. Most 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 in finite state machines state transitions are immediate, they may take several steps in the complex case. This difference is crucial for emotions as 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

[31] Even Commander Data in the Star Trek Saga seems to work fine without them!

complex agents that need to react quickly in real-time to unforeseen events. Such systems, then, might without their designers intending it, instantiate emotional states as defined above (e.g., an operating system with a mechanism that terminates processes, limits and reallocates resources, etc. in response to an overload in order to prevent the system from thrashing might be construed as being in an emotional state).

Coming back 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. It will define a sort of "niche", i.e., a set of requirements to be satisfied that, in turn, will determine a range of architectures able to satisfy the requirements relatively well. 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.

7.2 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 & Croucher 1981, Sloman 2001*a*)) 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".

Along the same lines, the question arises whether "emotions" are tied to visceral processes, as in biological systems notions like "emotion", "affect", "mood", or their more refined theory-based counterparts are often construed as characterising physical entities (animal bodies, including brains, muscles, skin, circulatory system, hormonal systems, etc.). This is of crucial consequence to designers of systems with artificial emotions, for 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.

We believe that the above alternatives might not necessarily be mutually exclusive. It seems possible and plausible to us that there are

> (1) deep, implementation-neutral, architecture-based concepts of emotion, which are definable in terms of virtual machine architectures without making reference to implementation-dependent properties of the physical substratum of possible implementations of the architectures

and that these deep concepts have

> (2) 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) and (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 to be the case), then any of their agent-based emotions, i.e., emotions elicited by agents, cannot be instantiated in architectures that do not allow for the ontological distinction between objects and agents. Such systems could consequently never be jealous (as being jealous involves other agents).

7.3 Comparison with other work

There is now so much work on emotions in so many different 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 is strongly opposed to some very well known views of emotions, in particular the class of views associated with William James, e.g. (James 1890), 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. Such emotional states also appear to be inconsistent with the sort of theory propounded by Damasio (1994).

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). We also believe that our emphasis on the link between the concept of emotion and mechanisms that produce strong dispositions to disrupt and redirect other processing fits much folk psychology and also many of the features of emotions that make them the subject of many 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. which are all of great interest. 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, though the focus of most current research on such "peripheral" phenomena as facial expression, posture and other easily measurable physiological states seems to ignore this.

When a robot first tells us in detail why it is upset by the opinions we have expressed about the poems it has written, many people will be far more likely to believe it has emotions than if it 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.

7.4 The next steps

Emotions, in the sense defined in Section 6.1, 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).

However, as in the case of humans, we can expect some emotional states, to be undesirable in machines that we design, but hard to avoid in certain contexts, if the machines have affective control mechanism that interact in complex ways. This is like thrashing in operating systems which can in certain situations of heavy load arise out of interactions between useful mechanisms.

The useful affective control mechanisms are likely to evolve in the sense that there are many evolutionary trajectories that, given a set of well specified initial conditions and fitness functions, will lead to those control systems (e.g., (Scheutz 2001*b*, Scheutz & Schermerhorn 2002).

A subset of those will be control mechanisms that can produce emotional states suited to dealing 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. It seems clear from the example of human beings that machines with architectures containing useful mechanisms will be capable of getting into highly dysfunctional states through the interactions of those mechanisms.

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. Condition-action rules, for example, may relate affective states recognized in others and behavioral expressions of affective states in a way that does not implement the kinds of control mechanisms which we called "emotional". Hence, such systems may appear to be emotional without actually being 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 many 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. Architectures for such systems will necessarily have to involve mechanisms that can preserve integrity of systems dealing with unexpected situations (again, compare this to (Simon 1967)). 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.

Currently, there is a good deal of work in computer science and robotics that – without explicitly acknowledging or even being aware of the similarity – 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)).[32]

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, 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

[32] Recent proposals for the development of "autonomic systems" are also headed in this direction.

(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* J. O. Urmson & G. J. Warnock, eds, 'Philosophical Papers', Oxford University Press, Oxford, pp. 175–204.

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/).

Boden, M. A. (2000), 'Autopoesis and life', *Cognitive Science Quarterly* **1**(1), 115–143.

Breazeal, C. (2002), *Designing Sociable Robots*, MIT Press, Cambridge, Mass.

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. & 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. & 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.

Maturana, H. & Varela, F. (1980), *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel, Dordecht (Holland).

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. & Jenkins, J. (1996), *Understanding Emotions*, Blackwell, Oxford.

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

Ortony, A., Clore, G. & 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. (1999), 'When physical systems realize functions...', *Minds and Machines* **9**, 161–196. 2.

Scheutz, M. (2001*a*), 'Causal vs. computational complexity?', *Minds and Machines* **11**((4)), 534–566.

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

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

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

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

Sloman, A. (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* C. Hookway & D. Peterson, eds, 'Philosophy and the Cognitive Sciences', Cambridge University Press, Cambridge, UK, pp. 69–110.

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

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

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

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

Sloman, A. (2000*c*), Models of models of mind, *in* M. Lee, ed., 'Proceedings of Symposium on How to Design a Functioning Mind, AISB'00', AISB, Birmingham, pp. 1–9.

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

Sloman, A. (2001*b*), Evolvable biologically plausible visual architectures, *in* T. Cootes & C. Taylor, eds, 'Proceedings of British Machine Vision Conference', BMVA, Manchester, pp. 313–322.

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

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

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

Sloman, A. & Croucher, M. (1981), Why robots will have emotions, *in* 'Proc 7th Int. Joint Conference on AI', Vancouver, pp. 197–202.

Sloman, A. & Logan, B. (2000), Evolvable architectures for human-like minds, *in* G. Hatano, N. Okada & H. Tanabe, eds, 'Affective Minds', Elsevier, Amsterdam, pp. 169–181.

Sloman, A. & Scheutz, M. (2001), Tutorial on philosophical foundations: Some key questions, *in* 'Proceedings IJCAI-01', AAAI, Menlo Park, California, pp. 1–133. http://www.cs.bham.ac.uk/~axs/ijcai01.

Turner, T. & Ortony, A. (1992), 'Basic Emotions: Can Conflicting Criteria Converge?', *Psychological Review* **99**, 566–571. 3.

Wright, I., Sloman, A. & 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.

Young, R. (1994), 'The mentality of robots', *Proceedings Aristotelian Society*.