To appear in *Human Cognition and Social Agent Technology*
Ed. Kerstin Dautenhahn, in the Series “Advances in Consciousness Research”,
John Benjamins Publishing

Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?)

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

Abstract:

This paper, an expanded version of a talk on love given to a literary society, attempts to analyse some of the architectural requirements for an agent which is capable of having primary, secondary and tertiary emotions, including being infatuated or in love. It elaborates on work done previously in the Birmingham Cognition and Affect group, describing our proposed three level architecture (with reactive, deliberative and meta-management layers), showing how different sorts of emotions relate to those layers.

Some of the relationships between emotional states involving partial loss of control of attention (e.g. emotional states involved in being in love) and other states which involve dispositions (e.g. attitudes such as loving) are discussed and related to the architecture.

The work of poets and playwrights can be shown to involve an implicit commitment to the hypothesis that minds are (at least) information processing engines. Besides loving, many other familiar states and processes such as seeing, deciding, wondering whether, hoping, regretting, enjoying, disliking, learning, planning and acting all involve various sorts of information processing.

By analysing the requirements for such processes to occur, and relating them to our evolutionary history and what is known about animal brains, and comparing this with what is being learnt from work on artificial minds in artificial intelligence, we can begin to formulate new and deeper theories about how minds work, including how we come to think about qualia, many forms of learning and development, and results of brain dmange or abnormality.

But there is much prejudice that gets in the way of such theorising, and also much misunderstanding because people construe notions of “information processing” too narrowly.

Architectural Requirements for Human-like Agents Both Natural and Artificial.

(What sorts of machines can love?)

Aaron Sloman
School of Computer Science
The University of Birmingham, UK

1 Can machines have emotions?

In February 1998 I was invited to a literary society to talk on whether machines can love. The presentation was a mixture of philosophy of mind, literary quotations on love, speculation about evolution, theoretical ideas from Artificial Intelligence, and conjectures about human minds. Later Kerstin Dautenhahn kindly invited me to convert my slides into a chapter for this book. The result is a collection of conjectures about information processing mechanisms underlying human emotions, moods, attitudes and other cognitive and affective states, like love and grief. I shall provide some sketchy evidence that both common sense and the work of poets and playwrights involve an implicit commitment to an information processing infrastructure. However, other things besides healthy adult human beings have minds, and different sorts of minds require different sorts of information processing architectures.

If we analyse familiar mental states and processes found in normal adult humans, and compare them with capabilities of infants, people with brain damage or disease, and other animals, we find evidence for a diverse array of architectures each supporting and explaining a specific combination of mental capabilities. This provides a broader and deeper explanatory theory than is normally found in philosophy or psychology. It also requires going beyond the majority of AI projects in considering both designs for *complete* agents and also *comparative* analysis of different sorts of designs as suggested in (Beaudoin & Sloman, 1993; Sloman, 1993; Mithen, 1996).

No amount of observation of the behaviour of any animal or machine can determine the underlying architecture, since in principle any lifelong set of behaviours can be produced by infinitely many different information processing architectures. We can attempt to constrain our theories by combining a number of considerations, such as: (a) trade-offs that can influence evolutionary developments, (b) what is known about our evolutionary history, (c) what is known about human and animal brains and the effects of brain damage, (d) what we have learnt in AI about the scope and limitations of various information processing architectures. I offer a brief and incomplete report on a theory based on such constraints. The main conjecture is that human information processing makes use of (at least) three different concurrently active architectural layers which evolved at different times, which we share with other animals to varying degrees, and which, along with various additional supporting modules, account for different cognitive and affective states, as well as offering the hope of explaining different kinds of learning and development, different possible effects of brain damage, and other abnormalities. Such an architecture could give robots human-like mental states and processes.

Prejudice about machines and information processing often gets in the way of understanding and evaluating such theories, so that people ignore some rich explanatory ideas developed in the last few decades (e.g. Herbert Simon's important ideas (Simon, 1967)). I shall therefore sketch and comment on the two main kinds of resistance to these ideas: doubting and fearing.

2 Doubters and fearers

Many people are sceptical about or disturbed by the idea that robots or software agents may one day have thoughts, feelings, hopes, ambitions and the like, or experience the world as we do. Some are influenced only by evidence, others by fear, or dislike.

2.1 Doubters: the perceived gap

Many are doubters because they see the limitations of existing computer-based machines and software systems and cannot imagine any ways of overcoming these limitations. They do not realise that we are still in the early stages of learning how to design information processing systems.

Existing AI systems do not yet have whatever it takes to enjoy or dislike doing something. They do not really *want* to do something or *care* about whether it succeeds or fails, even though they may be programmed to give the superficial appearance of wanting and caring. The attempts to replicate other animal abilities are also limited: for example, visual and motor capabilities of current artificial systems are nowhere near those of a squirrel or nest-building bird, as I have argued in (Sloman, 1989).

Because of the huge gap between machines developed so far and what animals can do, some people think the gap can never be bridged. That could turn out to be correct, if, for instance, the functioning of animal brains turned out to require some kind of mechanism that we have not yet dreamed of. The question is open.

It may be possible to convince some doubters by (a) enhancing their understanding of the real but unobvious possibilities of information processing machines, and (b) deepening their understanding of our ordinary concepts of 'feeling', 'thought', 'desire', 'love', etc., in order to reveal how our ordinary concepts of mind implicitly presuppose an information processing substratum.

Often defenders of AI do only (a). They try to remove doubts by demonstrating sophisticated things computers can already do, and pointing out that their capabilities will be enhanced by faster processors and bigger memories. That often fails to convince because it does not address the nature of mentality. Only by providing new insights into mental phenomena can we hope to convince real doubters that processes in computers may one day include feelings, experiences and thoughts. I shall sketch an attempt to bridge that gap below.

2.2 Fearers: the longed for gap

Some who reject the idea that robots and virtual agents can think and feel simply do not *like* the idea of machines (as they construe them) ever being so much like us. They may dislike it for many reasons, including fear of machines taking control (as in many science fiction novels) or more subtly because like Weizenbaum (1976) they fear that somehow human dignity is threatened if ‘mere machines’ turn out to be capable of all the interesting and important mental processes for which we value humans.

This kind of ontological neurosis (excessive concern about the place of humans in the overall scheme of things) lay behind at least some of the opposition in the past to the Copernican theory, which pushed us from the centre of the universe, and to the Darwinian theory of evolution, which blurred cherished boundaries between humans and other animals, a continuing concern of many researchers into animal capabilities.

In this paper I ignore the fearers who *dislike* the idea that robots will one day turn out to be like us. Dealing with such worries requires more than argument. Pointing out that intelligent machines could hardly do more horrible things to humans than humans do to one another is unlikely to help. I shall also not discuss theological objections, since I think they are based on false premisses.

2.3 Ask how, not whether

Whether machines can think, feel, care, hope, learn, have emotions, etc. is not in question, for humans are machines, though not artefacts. What sorts of machines?

3 Four kinds of machines

There are at least four kinds of machines known to science and engineering. They are not mutually exclusive: the same thing can be in two or more categories.

(a) Machines which manipulate force and energy.
These include many machines that people (and some animals) have made for centuries, including many kinds of tools.

(b) Machines which manipulate matter by reorganising it.
These include diggers, lawn-mowers, nut-crackers, looms, moulds, and also chemical and biological mechanisms which decompose and reorganise matter at the atomic or molecular level, for instance in production of solvents, detergents, drugs, etc. Every biological organism both transforms forces and energy and also uses matter-transforming

machines which take in nutrients and manufacture tissues, hormones, blood cells, sperm, and so on. Many physical machines are simultaneously of types (a) and (b).

(c) Machines which transform physical state.
These include ovens, forges, and many machines in chemical plants. At the molecular level they can be viewed as a special case of (b).

(d) Information manipulating machines.
These acquire, create, store, transform, manipulate, use and transmit information. Exactly what this means is a very subtle and complicated topic, discussed in (Sloman, 1996a; Sloman, 1996b). Information manipulating capabilities cannot exist without being implemented in a physical machine. In philosophers' jargon, information processing capabilities are *supervenient* on physical capabilities.

Organisms are not simply machines which manipulate forces and energy, and transform matter: they are also information processing machines. However, there is much diversity. They obtain information from the environment in different ways, store it, use it, transform it and communicate in different ways. They also deal with different kinds of information. An earthworm has neither the need nor the ability to know where Paris is, or how to multiply two numbers.

Humans depend on a mixture of mechanisms dealing with different sorts of information, processed in diverse ways, including sensing the environment, learning a language, absorbing a culture, generating new goals, making plans, evaluating and selecting plans, learning skills, learning generalisations, and many more. As Wiener noted, many of these processes are primarily concerned with *control*, e.g. control of attention.

Many people think of 'information processing' as restricted to computers manipulating bit-patterns in rigidly programmed ways, e.g. (Rose, 1993). This can lead to spurious arguments against information processing models of minds, or brains. We require a broader notion of 'information processing,' as used by many software engineers, biologists, and some brain scientists (e.g. Damasio, 1994).

3.1 Poets on love

I shall try to show how being an information processor is involved in many mental states, e.g. loving and fearing. For instance, Shakespeare was implicitly alluding to features of an information processing system when he wrote:

LOVE IS NOT LOVE
WHICH ALTERS WHEN IT ALTERATION FINDS

This implies that lovers can find alteration, i.e. perceive changes in loved ones. Finding alteration often diminishes or wipes out love and trust. Yet a common theme in literature is that true love is not so easily changed. It is resistant to common forms of information processing, including discovering disappointing facts about the beloved. Thus in love, some control states are unusually resistant to being changed by new information. There are many relevant entries on love in the Oxford Dictionary of

Quotations, including humorous poetry which alludes implicitly to information processing mechanisms, for instance when Sir John Suckling jokes about the oft claimed constancy of love:

OUT UPON IT. I HAVE LOVED,
THREE WHOLE DAYS TOGETHER
AND AM LIKE TO LOVE THREE MORE,
IF IT PROVE FAIR WEATHER

Of course, I am not claiming that such authors had clear ideas about information processing, though aspects of the chemical infrastructure of our information processing are often acknowledged, e.g. when Calverly wrote:

THE HEART WHICH GRIEF HATH CANKERED
HATH ONE UNFAILING REMEDY — THE TANKARD

I shall try to show that some of the information processing capabilities of most interest to us in our social life (including the ability to be in love) depend on aspects of our architecture which evolved recently and are probably not shared with most animals, except perhaps other primates (though I am not sure). We also have much older information-based control mechanisms which are shared with many other animals and which are easier for brain scientists to study (LeDoux, 1996). These explain some of our more primitive emotions, as explained below.

I shall not attempt to *prove* that we are information processing machines, but will merely try to explain in what sense we are, with illustrations of information processing capabilities. The ultimate test of the idea will be our ability to develop more detailed theories which are better able to explain the full range of human and animal mental capabilities than rival theories. That will take some time!

3.2 Cluster concepts cannot be defined precisely

Some readers may hope for definitions of terms like *information processing*, *mental process*, *consciousness*, *emotion*, *love*. However, each of these denotes a large and ill-defined collection of capabilities or features. There is no definite collection of necessary or sufficient conditions (nor any disjunction of conjunctions) that can be used to define such terms. The features and capabilities involved in mentality or consciousness or emotions can be present or absent in different combinations, in different animals, in people at different stages of development or after brain damage. (And, some of us claim, also in future robots.) Such concepts can be described as ‘cluster concepts.’ (Compare family resemblance concepts discussed in (Wittgenstein, 1953)).

If *emotion* is a cluster concept it is a mistake to ask how *it* evolved, what *its* function is, what *its* neural correlates are, etc., for there is no definite *it* to which the questions relate. (The same applies to *consciousness*.) I shall illustrate this below in relation to emotions by showing how different architectures support different subsets of the loosely defined cluster of features associated with our ordinary notion of *emotion*.

However, if the phenomena are all related to some general underlying principles, such as principles common to different information processing architectures, then it may be possible one day to define precise new technical concepts in terms of those principles. E.g. below I shall (partially) define 'tertiary emotion' in terms of a type of architecture. Such new theory-based concepts are often loosely related to pre-scientific cluster concepts which inspired the new theories. This has happened many times in the history of science, including refinement of our pre-scientific concepts of kinds of stuff, of kinds of animals, and kinds of chemical processes. It is only *after* we have deep explanatory theories that precise definitions can be given. Unfortunately we unwittingly deceive ourselves into thinking that we start with clear and precise concepts, e.g. of *experience, emotion*, etc. Likewise people thought they had a clear and precise concept of simultaneity, until Einstein exposed the problems.

4 The sorts of machines we are

Until recently the only significant information processing machines were organisms. However, since the middle of the 20th century our understanding of and ability to create new kinds of artificial information processing machines has accelerated rapidly, though the science is still in its infancy, and we have much to learn. In this and the next two sections I shall elaborate, first in very general terms, then in more detail, an information-processing model of mind. I assume that we are physical, chemical, biological and information processing machines:

- rooted in carbon, hydrogen, oxygen, nitrogen, iron and other physical stuff,
- evolved through millions of years of exploration,
- partly revealing our history in our design,
- grown in wombs, cots, playgrounds, and cultures,
- acquiring, storing, transforming and using energy,
- acquiring, storing, transforming and discarding matter,
- acquiring, storing, transforming and communicating, information,
- using information in many ways, including sensing, deciding, doing and feeling,
- writing poems, plays and newspaper reports,
- providing the stuff to write about,
- deceiving ourselves that we are unique,
- often wanting the truth to be THUS ..., rather than wanting to know what the truth is!

4.1 But we are not 'just' machines

We are machines but we are not *just* or *mere* machines, any more than computing systems are. Computing systems are certainly information processing machines, but they can also be personal assistants, factory controllers, tutors, translators, planners, network managers, automatic pilots, theorem provers. Beware the *nothing buttery* fallacy: the temptation to conclude that something is 'nothing but a ....'.

A system may be describable using a certain ontology without that being all there is to the system, even if the description is *complete* at that level. The people, buildings, transport mechanisms, etc. in a large city can (in principle) be described in very great detail using the language of physics and chemistry. But that does not mean there is nothing else, for there will probably also be crime, poverty, jobs, salaries and laws, obligations, contracts, and knowledge. Moreover these non-physical entities can have important causal powers, despite the causal completeness of the physical level.

In philosopher's jargon (explained at length in (Chalmers, 1996)), many non-physical entities, like crime and poverty, are 'supervenient' on the physical infrastructure. But that does not stop them being real and causally efficacious. Poverty *really* can cause crime and in doing so it can cause physical events, like TV sets moving through broken windows and knives or bullets through skin. Likewise events in an information processing *virtual machine* can cause events in a physical machine: for instance changes on a computer screen, movements of a robot, the safe landing of an airliner, and temperature control in a hospital intensive care ward.

If a physical description of a system is complete at its own level people sometimes infer that the system is 'nothing but' a collection of atoms, molecules, etc.. This is the 'nothing buttery' fallacy. An ocean wave might seem to be nothing but a large collection of molecules partaking of roughly vertical or circular motion at a fixed location, but that ignores the large scale horizontal motion and forces which can have such destructive effects when they hit the shore. Interactions between levels in information processors are more subtle: physical events cause virtual machine events and some virtual machine events cause physical events.

Some multi-level systems are much harder to understand because they are so subtle and complex, and because we have not yet learnt the concepts and techniques required for thinking about how they work. Trying to describe precisely the relations between minds and brains may be premature at present, if we do not yet have sufficiently rich and subtle concepts. It is much easier to start from systems we have designed. After developing conceptual tools for explaining how they work we may be able to extend those tools to deal with more complex cases.

We understand how a typical computing system contains several levels of virtual machines within which causal interactions occur between information structures. Processes such as reformatting a document, or finding a logical proof, or interpreting an image, are real and efficacious, even though at a lower level the computer is completely describable in terms of its digital electronics. At a still lower level quantum physicists use yet another set of concepts. Maybe physicists of the future will find something even deeper. But that need not affect the reality of the levels we now know.

To take any particular level and say: there is really nothing but *that* is to impose arbitrary constraints on what is real. We cannot ignore the existence and causal powers of poverty and crime. (Though some politicians may find it tempting.) Likewise a partially completed proof in a machine is real, and can cause both internal processing events and external physical events (on a screen) even though no physicist or electronic engineer could observe or measure the proof.

Lack of understanding of information processing virtual machines and how they can be implemented in physical systems has led many philosophers and theologians to assume that thoughts and feelings must inhere in some non-physical kind of mechanism, sometimes called a soul, or spirit, which can exist without any physical implementation. Gilbert Ryle scornfully labelled this 'The ghost in the machine'. His 1949 book had important ideas about internal, unobservable, information processing, e.g. in his chapter on imagination. But he lacked our conceptual tools, and as a result was wrongly interpreted as a behaviourist, denying the existence of the mental.

4.2 We are multi-level information processors

Although we have physical architectures we are not 'mere' physical machines. We also have information processing architectures, implemented in our physical architectures. Likewise we are not 'mere' information processors since we are also parents, teachers, criminals, lovers, scientists, etc. These features and relationships are implemented in our information processing capabilities in combination with a social context.

Information processing includes: sensing the environment, interpreting sensory information, modifying stored information (beliefs, desires, intentions, plans, skills) in the light of new information, generating goals by various means (some innate, some learnt, some unconscious, some conscious), inventing new options (things to do, to make, to look for...), considering and evaluating options, selecting among possible actions, wondering about consequences, reconsidering previous decisions, and much more. Analogous capabilities exist in sophisticated game playing machines.

Humans also enjoy some activities and dislike others, (sometimes) detect our own states (anger, puzzlement, hope, ...), evaluate our own thoughts and reasons (as selfish, unproductive, altruistic, creative, foolish, etc.), feel ashamed, guilty, fearful, excited, and become self-satisfied, infatuated, obsessed, ecstatic, ... Will AI systems ever have all these capabilities? I see no reason to doubt the possibility (or even to fear it). Some counter-arguments seem to be born of dislike of the idea (a 'longed for gap') rather than deep analysis of what it is to enjoy something or be infatuated.

Not all animals can do all those things. Not all humans can do them all: very young children, and people whose brains are either genetically flawed or damaged by accident or disease, may lack some of these abilities. We need to understand why.

4.3 Longing as a tertiary emotion

Consider an example: *Why can't a goldfish long for its mother?* Longing for one's mother involves at least: (i) knowing one has a mother, (ii) knowing she is not present, (iii) understanding the possibility of being with her, and (iv) finding her absence unpleasant. These all involve possessing and manipulating information, e.g. about motherhood, about one's own mother, about locations and change of location, and about the desirability of being close to one's mother.

Those conditions do not suffice. If someone in Timbuctu whose mother is in

Montreal satisfies conditions (i) to (iv) but hardly ever thinks about his mother, and simply gets on with his job, enjoys his social life, and always sleeps soundly, then that is not a case of longing. He may *regret* her absence (an attitude), but he does not *long* for her (an emotion). Longing for someone requires something more, namely (v) not easily being able to put thoughts of that someone out of one's mind. Thoughts of the longed for one will return willy nilly. (Though perhaps not in *mild* longing!) This is not just a matter of definition: it is a fact that some human mental states involve partial loss of control of attention.

You cannot lose what you have never had. So a requirement for being in such a states is having the *ability* sometimes to control what one is thinking of and also being able sometimes to *lose* that control. This presupposes an information processing mechanism some part of which can control which information is being processed, but which is not always in *total* control. Deep longing for one's mother involves partly losing control of thought processes, a *perturbant* state. An incomplete grasp of these ideas gives rise to a confused notion of *free will* which appears to some people to be inconsistent with causation (contrast Franklin, 1995).

I call these perturbant states 'tertiary emotions', as explained below. Other examples are anger (Sloman, 1982), grief (Wright, Sloman & Beaudoin, 1996), guilt, jealousy, excited anticipation, infatuation and many others. We shall see below that such perturbant states arise in an architecture where a high level control mechanism sometimes loses control. That capability is not one of its functions but is a side effect of other functions. Such states could also occur in some intelligent robots (Sloman & Croucher, 1981).

4.4 What sort of architecture could support being in love?

There are many other states which characteristically involve partial loss of control of attention. This is one of the differences between loving someone and being in love.

X is in love with Y IMPLIES *X's thoughts are constantly drawn to Y*

Love in general is an *attitude* and need not be emotional: You can love members of your family without constantly dwelling on them. You can also love your country, love the organisation you work for, love football, love the music of Mozart, without any of these constantly flooding your thoughts.

Loving your country does not involve thinking about it most of the time, but only when some relevant information or decision turns up. The rest of the time the love is just one among many *dormant dispositions* – but real all the same. Being *in love* is not so passive: thoughts of the beloved will return when there is no particular reason. In extreme cases (infatuation) it may be very difficult to think about anything else. Likewise in extreme grief. However, even extreme emotions can temporarily become dormant while some urgent and important task, or a gripping movie, holds one's attention. To explain all this we need to understand the underlying information processing architecture which makes attending and thinking possible at all, and which accounts for the possibility of redirection of attention. Explaining how we can *lose* control of our thoughts first involves explaining how we can *have* control.

5 Architectural layers

5.1 An architecture explains a collection of states and processes

A particular information processing architecture will support some states and processes, but not others. Some problem-solving processes require an architecture including a procedure invocation stack, providing an ordered 'memory' of unfinished procedures. A condition-action rule interpreter with no explicit stack makes it hard to implement strategies requiring deeply nested actions, although it provides good support for opportunistic information processing. An architecture which includes both a stack and mechanisms for explicitly inspecting or changing its contents, will be less restrictive than one without.

An architecture in which there is always only one process active makes it hard to implement self-monitoring and self-control, whereas an architecture supporting (physical or virtual) concurrent processes makes it easier for one process to inspect the state of another, and interrupt or modulate it, e.g. if looping is detected.

An information processing architecture explains a variety of states and processes somewhat as the atomic theory, a theory of the architecture of matter, generates and explains a variety of types of physical elements and chemical compounds.

Our knowledge of information processing architectures is still very primitive. Studying a wider range of architectures will extend our ability to explain how different collections of competences are possible. Each architecture provides a framework for generating a family of descriptive and explanatory concepts. We can expect to find different architectures in different sorts of humans (including infants and people with brain damage, etc.), different sorts of animals and different sorts of artificial agents.

5.2 A conjectured architecture for adult humans

Within an architecture we can distinguish perceptual subsystems, motor subsystems and more central mechanisms. I conjecture that in adult humans all of these consist of layers with different levels of sophistication which evolved at different times, and which are shared with different numbers of other animal species, and explain different aspects of human mentality, for instance different types of emotions. The layers act in parallel and both cooperate and compete with each other. The different 'layered' capabilities in sensory and motor subsystems evolved to work with the different central layers. All this has some similarities to many other theories, e.g. (Craik, 1943; Minsky, 1987; Damasio, 1994; Dennett, 1996; Mithen, 1996).

5.3 Three types of sub-architecture

The central layers are (1) a very old *reactive* layer, found in all animals, including insects), (2) a more recently evolved *deliberative* layer, found in varying forms in a

subset of other animals, (3) an even more recently evolved *meta-management* layer providing self-monitoring and self-control, perhaps found only in other primates, and probably not in very young human infants. I.e. the architecture of an adult human is not present at birth, but results from a boot-strapping process. Each layer is a collection of cooperative sub-mechanisms combining to perform a collection of internal or external functions. Additional modules support or modulate the three main layers:
(a) One or more global *alarm* systems able to detect patterns requiring rapid global reorganisation of internal and external behaviour. Compare the interrupt mechanisms discussed by Simon (1967) Oatley and Johnson-Laird (1987), and the role of the amygdala in the theories of LeDoux (1996), Damasio (1994).
(b) Associative content-addressable information stores, essential for 'what if...' deliberations, and for predicting what is likely to happen next (Craik, 1943).
(c) Mechanisms generating, comparing, selecting and prioritising motives (Simon, 1967; Sloman & Croucher, 1981; Beaudoin, 1994; Beaudoin & Sloman, 1993).
(d) Global quantitative and qualitative control subsystems which account for mood changes, differences between waking and sleep, types of arousal, etc. Some of these global controls in brains use chemical mechanisms, but similar functions might be implemented differently in artificial agents.

5.4 Layered perceptual and motor systems

Perceptual mechanisms need varying degrees of sophistication for their tasks. The contribution of vision to posture control uses relatively simple optical flow detection. Segmentation and recognition of objects in a scene requires more global and knowledge-based processing. Seeing a room full of people as a 'party' or a 'seminar' or as 'highly charged' requires far more abstract and sophisticated forms of processing. Learning to read text or sight-read music involves different collections of layered abilities (Sloman, 1989).

Some action mechanisms are old and relatively primitive, such as contracting muscles, raising or lowering a leg, clamping jaws shut. More abstract actions involve considerable sensori-motor coordination such as picking up a large, heavy, unwieldy and unfamiliar object. Some actions use tools as an extension of the body, such as parking a car or feeling the shape of a hole with a probe. There are also semantically very rich actions such as uttering a sentence, playing a musical phrase on the violin, or making courtly gestures. The 'layering' of sensory and action systems is obscured by thinking of such systems only as input and output transducers.

5.5 Evolution by copying and modifying

Some of the mechanisms might have evolved from simpler mechanisms via the common biological process of copying then modifying. For instance deliberative mechanisms used for planning require an associative store of information about actions possible in various situations and their consequences. This might have evolved by copying an older reactive association mechanism. The new mechanism instead of merely reacting to its input by producing *control* signals (‘do this next’) might have answered ‘what if’ questions about past actions, as part of a learning and debugging mechanism. Later it could answer ‘what if’ questions about future actions. (Or perhaps planning came first?)

All this requires an ability to synthesise hypothetical context descriptions to feed into the memory, instead of using only current sensory information to drive it. That might have resulted from earlier developments producing high level perceptual mechanisms able to create abbreviated abstract descriptions of external objects or situations.

Such developments could involve multiple evolutionary stages, not yet understood. Perhaps a self-monitoring meta-management mechanism was evolved by copying the global alarm system and then changing its activities, including making it more amenable to rule-based control, and allowing it to use deliberative strategies. This would be much slower and more flexible than an alarm system. Moreover, with rule-learning, it can improve itself, and be influenced by a culture.

5.6 Different layers explain different sorts of emotional processes

We can distinguish at least three major categories of emotions, explained by the three sorts of processing layers.

(1) *Primary emotions* (Damasio, 1994; Picard, 1997). These are primitive emotional states (like being startled, terrified, sexually stimulated) based on the old reactive layer and global alarm system shared with many other animals. Patterns in sensory inputs are detected by the global alarm system which rapidly sends out a wide range of control signals, some causing physiological changes producing or preparing for action. Compare robots programmed to ‘freeze’ as soon as a human gets dangerously close.

(2) *Secondary emotions*. These states (like being anxious, apprehensive, relieved, pleasantly surprised) are generated in the deliberative layer, in which plans can be created and executed, risks noticed in advance, progress assessed, success detected, etc. They depend on the ‘what if’ representational capabilities provided by a deliberative mechanism. Some emotions (like relief that an accident was avoided) require counterfactual information about the past (what *might have* happened). An alarm system detecting a feature or pattern in the contents of current thoughts or problems can be triggered to produce a rapid global reaction, or change of state (e.g. producing nervousness and thereby more attention to detail, raised interrupt thresholds, more cautious movements, etc.). Detection of success, or receding danger could trigger reversion to a more normal state, as in relief.

Damasio calls emotions triggered by such cognitive processes ‘secondary emotions’. In chapter seven (especially page 137) Damasio (1994) suggests that secondary emotions *always* activate the same physiological mechanisms as primary emotions (Picard uses the phrase ‘sentic modulation’). However ‘always’ is an over-generalisation. There are considerable individual differences regarding whether secondary emotions triggered by cognitive processes produce physiological changes. The more important feature of secondary emotions is the ability of something like the alarm mechanism to redirect *mental* processes, so that, for instance, dangers and opportunities are noticed and appropriate actions considered. Moreover emotional maturity sometimes involves suppressing normal physical reactions and dealing with emotion-generating phenomena entirely mentally. (Not everyone can do this.) Purely mental processes need not be *cold* and *unemotional*. On the contrary, they can be rich in evaluative content and powerful in their effects on other mental processes, as when horrific news grips our attention despite our best efforts to think of something else.

We can therefore distinguish *purely central* secondary emotions from *peripheral* secondary emotions which invoke the bodily mechanisms used by primary emotions. Steve Allen alerted me to Damasio's chapter eight (page 184), where he explicitly adopts a similar position, suggesting that in some cases there are "as if" mechanisms which bypass the route through the body, so that "the prefrontal cortices and amygdala merely tell the somatosensory cortex to organise itself..." in a pattern that it would have assumed if signals had come through the body. (Damasio's ideas in that chapter have much in common with the ideas presented here, including his speculations about the role of ancient insect-like mechanisms in human brains.)

(3) *Tertiary emotions*. These are typically human emotional states involving partial loss of control of thought processes (perturbance), e.g. states of feeling humiliated, infatuated, guilty, or full of excited anticipation, where attempts to focus attention on urgent or important tasks can be difficult or impossible, because attention is drawn back to the focus of the humiliation or infatuation, etc. This can happen despite a meta-management decision to attend to something else.

These fit the definition of secondary emotions, but involve something more, namely partial loss of control of attention. This is possible only if there is something which normally provides that control. Only then does the notion of 'losing control' become relevant. Without meta-management you cannot explicitly evaluate the possibility of attending to A and to B, and then decide to attend to B because you judge that better. Without meta-management, an alarm mechanism or other mechanism cannot undermine a decision based on an explicit judgement that it would be better to attend to or think about B. *Only if there is a control mechanism can control be lost.* Thus tertiary emotions require a particularly sophisticated information processing architecture.

Not *only* meta-management can redirect attention. Normal processes of deliberation involve shifting attention, e.g. switching attention to new goals, and switching from thinking about ends to thinking about means. Reactive mechanisms, e.g. detecting something bright or something moving, or detecting utterances of one's own name, can redirect perceptual resources. A deliberative system may have many subprocesses competing for attention, perhaps using a network combining connection weights and activation levels. A meta-management mechanism explicitly evaluates and selects alternative foci of attention, recording that this has been done. Put another way: *just as external behaviour can be either reactive or deliberative, so can internal behaviour*. And in both cases the reactive mechanisms can sometimes defeat the deliberative mechanisms.

Further subdivisions are possible, of course. For instance, not all perturbances are emotional: emotions involve strong *evaluations* as well, and there are different sorts of evaluations: selfish, ethical, etc. I have no idea to what extent non-human animals (e.g. bonobos) have such self-monitoring and self-evaluating capabilities. They may have simpler versions supporting a different range of emotions involving some self awareness and self evaluation. This is a topic requiring empirical research and theoretical analysis of new architectures.

By using an architecture-based framework for defining different classes of emotions, and related notions like ‘mood’, ‘desire’, ‘enjoyment’ and ‘pain’, not discussed here, we avoid much argumentation at cross-purposes about emotions, because the theory supports a range of types. Dozens of different definitions of ‘emotion’ have been proposed. We can now see that there is no point arguing about which of N definitions is correct if there are N types of phenomena all of which need to be studied and explained. It does not matter whether we call them ‘emotions’ or not. The phenomena are important, not their names.

In various previous publications, e.g. (Sloman, 1992), I focused mainly on tertiary emotions, because I thought they were of most interest and importance in human interactions. I mentioned other kinds e.g. reflexes and startles and also the types discussed by Oatley and Johnson-Laird (1987), but did not have a clear view of how the different types could fit into a common architecture.

6 Different architectural layers and evolution

The three layers can now be described in a little more detail. Diagrams are used very impressionistically to indicate some of the features of the different mechanisms.

6.1 Reactive mechanisms (Figure 1)

Reactive mechanisms evolved very early and are widespread in plants, insects and all other animals. The more recently evolved deliberative mechanisms have not *replaced* the older mechanisms, but function alongside them, but not necessarily always dominating them.

The main feature of reactive mechanisms is their *inability* to contemplate, evaluate, and select possible future courses of action or other hypothetical possibilities. They can merely react to actually detected internal or external situations. In other words they cannot consider novel options before selecting them, or create new plans. They merely act, though some of the actions are internal and some external. Figure 1 crudely depicts a fairly sophisticated reactive architecture, including alarm mechanisms, and the layered sensory and motor systems described below.

Reactive systems are very varied. They may be composed of purely analog

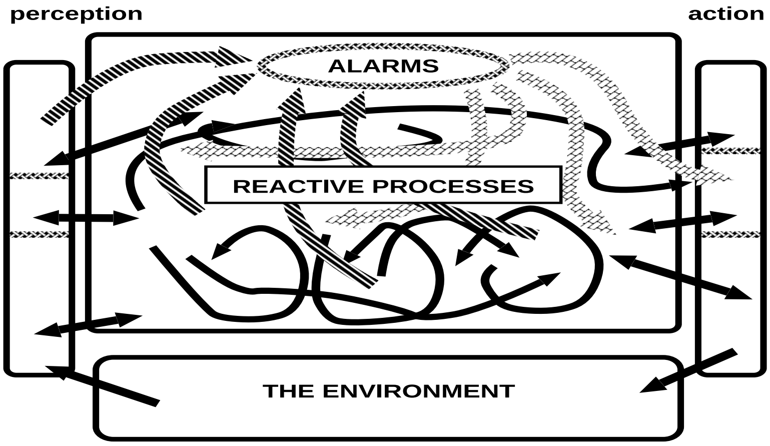

Figure 1: *Reactive systems with alarms: emotional ants?*
Internal reactive mechanisms may be chained together and may involve feedback loops, and changeable 'state variables'. Horizontal bars separate layers of abstraction in sensory and motor systems. Two-headed arrows indicate two-way flow, or feedback. Textured arrows represent fast links to and from the global alarm system.

(continuous), purely digital, or a mixture of analog and digital mechanisms. They may be totally *environment driven* so that the same environment always produces the same response, or partly *state driven* so that changeable internal states help to select actions. Detected needs can change internal states so as to modify subsequent selection among actions. (These are sometimes called *drives.*) This constitutes a primitive form of *goal-directed* though purely reactive behaviour. Compare Nilsson's Teleo-Reactive programs (1994).

Plants have many uncoordinated reactive mechanisms. Animals with a central brain have more coordination and global control. A colony of such animals can often be thought of as a higher level reactive organism with totally distributed control.

Within the brain of a reactive animal, the internal routes between sensors and effectors may be more or less indirect. Different indirect routes may operate concurrently, processing information at different levels of abstraction for different purposes (e.g. controlling posture and detecting food). The processing may be *uni-directional* or may use internal feedback loops. Some loops may be chained: e.g. move randomly until food visible, then go towards food until it is graspable, then eat it until satiated, etc.

Reactive systems may be very fast because they use highly parallel implementations. This may lead to simultaneous activation of different actions. Sometimes different actions can be performed in parallel, or combined, using something like vector addition (e.g. increasing speed and increasing angle of turn). If they are inconsistent a *selection* mechanism is required, e.g. symbolic priority rules, winnertakes-all neural nets, or a simple voting mechanism.

Layered sensory and action subsystems.
Short horizontal lines in the diagram indicate divisions between sensory and motor processes operating at different levels of abstraction. For instance, some reactions depend on relatively simple measures of optical flow or contact pressure, while others use more sophisticated and global percepts produced by complex interpretive procedures classifying entities in the environment relevant to higher level decisions (e.g. recognising something as a shelter, or as dangerous). Likewise actions may be simple internal or external changes (e.g. contraction of a particular muscle) or more sophisticated hierarchically controlled actions. Feeding often requires high level coordination of limbs (grasping food) and jaws. Many forms of running, jumping, flying, nest-building and mating require very complex coordination of complex collections of muscles.

More flexible reactive systems.
Some chained reactions involve innately determined sequences of internal states implementing plans selected by an evolutionary mechanism and encoded in genes. Others may be a result of new links produced by learning. Although purely reactive systems are rigid in that they cannot ‘think ahead’ creating and evaluating new plans, they may use learning mechanisms to alter weights linking conditions and actions. Where some actions change internal states forming conditions for subsequent actions, learning can be used to chain sequences of responses to produce the effect of learnt plans, provided that the architecture already has links which learning can strengthen.

Further flexibility can be achieved by allocating internal storage for different contexts which can be turned on and off (or varied continuously), to modulate reactive behaviours. Yet more flexibility can be achieved by allowing internal reactions which create simple temporary structures, for instance representing goals (e.g. ‘catch that animal’, ‘find a hiding place’).

Global alarm *mechanisms. (Figure 1)*
If chains of internal reactions intervene between sensory input and corresponding output, this may sometimes cause fatal delays or missed opportunities. Some sort of global ‘override’ mechanism could deal with this: an ‘alarm’ mechanism which allows rapid redirection of the whole system in response to detected patterns indicating opportunities or dangers. The alarm mechanism, which might be either entirely innate or partly trainable, could be simply another reactive sub-system with inputs from all parts of the organism driving a fast trainable pattern-recogniser able to trigger outputs to all parts of the system. Normally it would do nothing, but when turned on by appropriate conditions it could rapidly redirect the rest of the organism to produce freezing, attacking, feeding, fleeing, mating, attending (sudden high alertness), more general arousal, or more specific innate and learnt responses.

It appears that such systems first evolved a long time ago: many animals have one or more global alarm mechanisms. The brain stem and the amygdala both seem to implement alarm systems which evolved at different times. Different global alarm mechanisms could specialise in particular types of activation patterns and response patterns. Primary emotions in vertebrates appear to be implemented in such systems (LeDoux, 1996; Goleman, 1996).

Robots and software agents do not yet have all the characteristics of the reactive architectures described here. However, there has been a lot of work in robotics labs on reactive systems (much of it inspired by Rodney Brooks at MIT), and it is very likely that more and more sophisticated insect-like, or lobster-like creatures will emerge from such laboratories in the next few years, and also software systems controlling chemical plants, power stations, etc., all with the capability to have the sorts of primitive emotions sketched here.

Whether they will *know* they have them, and whether ants fleeing ‘in terror’ know they are terrified is another question. The third architectural layer sketched below can explain self-awareness. We may have to get used to the idea that without it reactions of terror and other primitive emotions may occur without being *experienced* as such by the organism. This could be equally true of new-born human infants, if they lack the third layer described below. The suggestion may seem repugnant, but that does not make it false.

6.2 Architectures with deliberative mechanisms (Figure 2)

A deliberative mechanism provides capabilities missing from reactive mechanisms, especially the ability to achieve an objective, in a *new* situation, by chaining together a novel sequence of actions. A reactive system may be able to invoke an *existing* plan, e.g. if a need is detected and allowed to trigger a sequence of context-driven reactions. But that presupposes a pre-existing implicit or explicit plan, produced by evolution or previously learned chained responses.

Novel complex actions may be discovered by a reactive explorer using trial and error with reinforcement learning, but this can be dangerous and time consuming. If a system has the ability to do hypothetical reasoning it can search a space of *possible* action sequences until it finds a suitable plan, as Craik pointed out (Craik, 1943). This requires a content-addressable associative memory store which can answer questions like: ‘What actions are possible in situation X?’ and ‘What effects would follow if action A were performed in situation X?’ ‘Which actions are relevant to a goal of type G?’

A system able to create potential new plans to evaluate requires a re-usable memory in which to build partial plans before selecting them. The tree-like structures in Figure 2 indicate partially constructed possibly hierarchical solutions to problems. These are in the re-usable work space. This re-use will make the process of exploration serial. There are other reasons why deliberative mechanisms must be sequential (digital) and

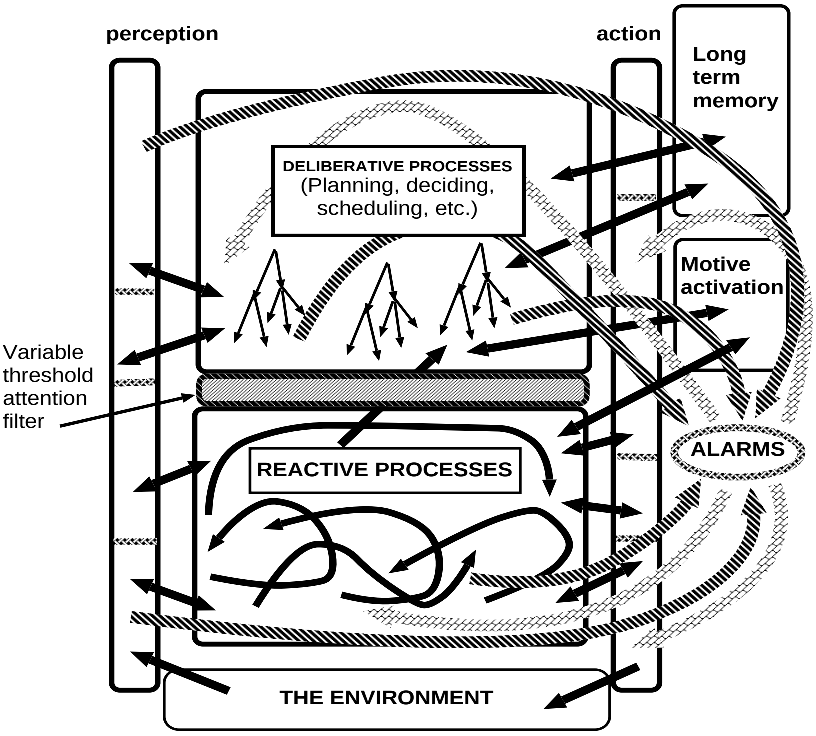

Figure 2: *Hybrid reactive and deliberative architecture, with global alarms. Now the deliberative layer also has links to and from the alarm system. A filter with dynamically varying interrupt threshold protects the resource-limited deliberative layer when dealing with tasks that are important, urgent and resource consuming.*

discrete, and relatively slow. For instance, even if the association store operates using a highly parallel and distributed neural implementation, it could still be restricted to answering one question at a time.

Extending the alarm mechanism.
As before, alarm mechanisms may be useful for rapidly redirecting a deliberative system when dangers and opportunities are detected. States within the deliberative layer could also feed into the alarm system, alongside signals from sensors and reactive mechanisms. Similarly the alarm system could send interrupts and redirection signals to the deliberative mechanisms, re-directing attention or changing the mode of processing. This is indicated crudely in Figure 2.

An attention filter.
A fast-changing environment can cause too many interrupts and frequent re-direction of attention, with more time spent switching between deliberative tasks than actually solving the problems (like a thrashing operating system). It may be important to prevent interruptions and diversions (e.g. by new goals) when the current goal is

very important, urgent and cognitively demanding. A partial solution could be a variable-threshold interrupt filter, depicted in Figure 2. This might also suppress global alarm signals under some circumstances (e.g. soldiers in battle not noticing injuries). However, as argued in (Sloman & Croucher, 1981; Wright, Sloman & Beaudoin, 1996) the priority and filtering mechanisms must be *fast* which means using unintelligent processes, sometimes leading to undesirable interruptions and emotional states.

Saving new plans for reuse.
Useful new plans generated by deliberative mechanisms can be transferred to the reactive system (the cerebellum?), perhaps as a result of repetitive operation. Storing them in the reactive mechanism may support much faster though less flexible execution.

6.3 The need for self-monitoring i.e. meta-management (Figure 3)

A deliberative mechanism needs strategies for deliberating. Those produced by evolution may be too rigid for changing physical and social environments. A meta-management layer allows deliberation processes to be monitored and improved e.g. learning to raise interrupt thresholds during 'busy' states, or noticing that certain planning methods fail in certain conditions. Such learning may reduce failure in deliberative tasks, reduce interference between goals, detect time wasted on unsolvable problems, etc. Flexibility is even greater if meta-management can use rules, categories and values absorbed from the surrounding culture.

The ability to attend to and categorise internal states has subtle consequences, which may have influenced evolution of self-monitoring capabilities. Parents can diagnose a child's problems more easily if the child can attend to and describe internal symptoms. Compare describing visual experiences to an optician, or telling a dentist which tooth hurts. Attending to intermediate visual data-structures is required for drawing accurately: noticing how things *look* (e.g. elliptical) as opposed to seeing how they *are* (e.g. circular). (This could explain the existence of *qualia*.)

Further extension of the alarm mechanism
The alarm mechanism described previously could be extended with inputs from and outputs to meta-management processes, allowing alarm reactions to be triggered by and to modify meta-management. Alarm systems require rapid reactions, so they must depend on fast, and therefore shallow, pattern recognition rather than deep analysis. Consequently, alarm processes will not always be optimal and some of the interruptions and redirections will be undesirable. Tertiary emotions include such cases. Perhaps some addictions, obsessions, and some attentional disorders depend on transformations in the alarm mechanism.

Limitations of meta-management
Self-monitoring, self-evaluation and self-control are all fallible. No system can have full access to all its internal states and processes, on pain of infinite regress. Prefer-

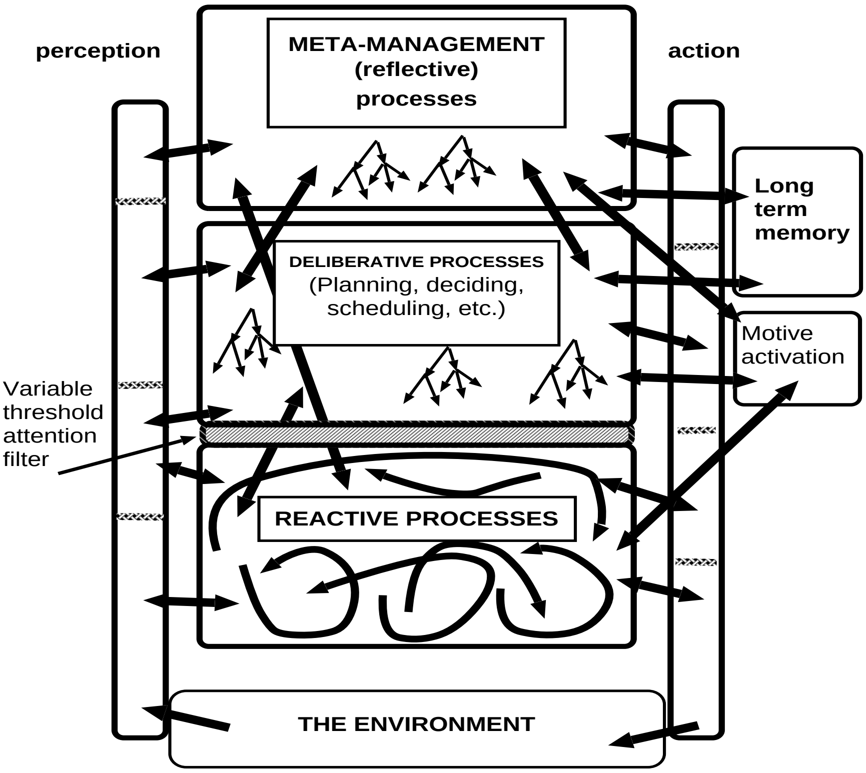

Figure 3: *Towards a human-like architecture, with reactive, deliberative and meta-management mechanisms. (Alarm mechanisms not shown)*

ences used in selection and self-evaluation may be erroneous or ill-judged (e.g. self-evaluation based on religious indoctrination). Control over deliberative processes may be partial, e.g. because the global alarm mechanisms cannot safely be suppressed completely, or because of loud noises, addictions, etc.

Figure 3 gives a crude indication of the sort of three layered architecture we are discussing, including showing (very inadequately) that perceptual and motor systems are also multi-layered. The alarm system is not shown because that would make the figure too cluttered (compare Figure 2). To envisage the addition of an alarm system in Figure 3 imagine an octopus on one side with tentacles extending into all the other sub-mechanisms, getting information and sending out global control signals.

6.4 Non-semantic control

The sorts of meta-management control sketched above involve precise direction of attention, or invocation of a strategy or evaluation of some state. These processes involve semantic content, e.g. reference to objects of attention or actions to perform. There is another type of control which produces global quantitative changes, for instance changing speed of operations, or degree of perseverance, or thresholds for attention diversion, or likelihood of adopting risky strategies. Some of these can be described as changes of *mood*. In animals a very complex chemical infrastructure is involved in some of these general control changes, as indicated by the effects of hormones and drugs, including, for instance, producing or alleviating depression, producing euphoria or hallucinations, reducing precise control of thoughts or actions. The chemical infrastructure can be influenced by alcohol and other drugs, smoking, disease, as well as by mental processes and natural bodily cycles.

Some people accept that architectural features outlined earlier could be implemented in computer-based systems, but doubt that functions based on chemical processes can be simulated computationally. That is an empirical question whose answer will depend on the precise nature of these functions. It may turn out that equivalent non-semantic control functions could use alternative mechanisms, for instance electronic analog devices or even software control mechanisms. It is easy to use a global real variable to have a global effect analogous to concentration of a chemical. Replicating spatially varying concentrations requires a little more ingenuity.

7 Some qualifications and implications

7.1 Variability in meta-management

Meta-management need not use a rigidly fixed collection of strategies. It should be modifiable by learning, so that one can detect new aspects of one's mental processing and evaluate them or control them in new ways. Examples would be learning to detect that one's grasp of a topic is confused, or that one is deliberating in a selfish way; and learning to disapprove of that sort of deliberation (which does not come naturally).

Instead of being rigid and monolithic, meta-management strategies may be different in different contexts. So the system may be thought of as 'occupied' by different 'control regimes' at different times, for instance: being a gentle parent at home, then driving a car aggressively, and becoming a cold and ruthless manager at the office.

Perhaps this is relevant to multiple personality disorders and other sorts of problems which lead people to seek therapy? This suggests many empirical questions. What are the 'role-switching' mechanisms? How can they go wrong? Can abuse in infancy produce long term damage in the architecture, and if so how?

Some of the states and processes described here, especially some of the high level emotional states in which there is a partial (and sometimes undesirable) loss of control, are not produced by mechanisms which evolved to produce them. They are side-effects, or emergent features of interactions between several mechanisms with other functions. Thus it is pointless asking what the functions of such states are.

In particular, the more sophisticated secondary and tertiary emotions are not directly implemented in an emotional mechanism, even if the simpler primary emotions are directly implemented in a global alarm system.

More generally, not everything supported by a mechanism is part of its function: multi-processing computer operating systems support thrashing, but do not have a thrashing mechanism! In some heavy load conditions they simply do far too much paging and swapping instead of doing useful work. So some functional mechanisms have dysfunctional consequences. In some cases additional mechanisms can detect those consequences and take corrective action, as in an operating system which detects that it is thrashing and prevents any new processes from starting up.

Sometimes a side-effect turns out to have beneficial consequences, which are then exploited for their effects. A person who finds that certain kinds of distress generate sympathy and support in others, may ‘learn’ to be distressed more often and in an exaggerated way. Likewise, a teacher may discover that real anger can be used to control a classroom, and learn to become angry. (Second-order functionality.)

7.2 Forms of learning and development

In such a complex architecture there are many different forms of development or learning that can occur, including: adding new capabilities to existing modules, creating new modules to extend the architecture, adding new links between modules, extending the formalisms used within a module (e.g. learning a new language, or a new notation for mathematical reasoning or music), storing new facts and associations in the long term factual memory, copying a new plan or strategy developed by the deliberative mechanism into the reactive mechanisms, thereby creating new reactive skills. In humans this kind of copying uses repetition of actions, with the deliberative system apparently supervising the training (or re-training) of the reactive system.

Different types of learning can be expected in different parts of the architecture. For example, perceptual mechanisms may learn to use new high level categories in classifying or interpreting perceived events. Examples are learning to read text or music fluently. In fluent reading the new percepts trigger *internal* actions. Action subsystems may learn to produce new complex orchestrated behaviours in response to more abstract “instructions” from the central mechanism.

Some new meta-management strategies, e.g. attention control strategies, seem to be produced by certain kinds of training, e.g. in meditation. The influence of a culture operating on the meta-management level can lead an individual to develop new ways of categorising and evaluating internal states, new forms of motivation, new motive generators, and new motive comparators, though much work needs to be done to explain how this works in detail. Cultural learning can vastly speed up learning by individuals. Forms of loving are also influenced by the culture, including disappointments caused by unrealistic culturally generated expectations.

Some subtle kinds of learning seem to involve the suppression of innate or previously learnt reactions. For instance emotional maturity includes learning to suppress or ignore some of the reactions of the global alarm system. This may include learning new strategies for adjusting the attention filter.

When a system can be changed in so many different ways, it is to be expected that in addition there are also many ways in which damage, disease, or genetic disability can change the system so as to interfere with its functioning. I suspect most professionals concerned with the identification and treatment of such problems, whether in education, counselling, psychiatry, etc. are aware of at most a tiny subset of the things that can happen. Perhaps these ideas will lead to helpful expansion of therapies.

7.3 There is no unique architecture

Many of the ideas sketched here are speculative. One problem is the difficulty of inferring architecture from known capabilities, since alternative architectures can in principle produce the same performance over a life time, as already remarked. Yet by analysing the trade-offs we may be able to rule out theoretically possible cases.

For instance, a purely reactive system could in principle do everything that can be done by a system with deliberative capabilities. However, the time required to evolve a collection of reactive behaviours large enough to cover the actions that a particular planning system could generate may be too long for the history of the universe. Moreover, storing them might require a brain too large to fit on the planet, and DNA molecules might be too small to encode them all. (Even the game tree for chess could not be fully encoded in any physical system.)

It may be that evolution 'discovered,' as AI designers have, that a good way to overcome these obstacles is to produce systems which are modular in the manner sketched above, and capable of explicit deliberation and planning. Or it may have found some alternative method which we have not yet thought of.

Unlike behaviourist psychologists and some AI researchers who reject explicit deliberation, I am inclined to regard human introspection, and everyday observation as providing at least *prima-facie* evidence for some of our capabilities. This gives me good reason to believe that people I know can plan many facets of a trip to an AI conference well in advance of taking a taxi to the airport. Likewise I know that people can memorise and (sometimes) reliably reproduce or use poems, stories, jokes, algebraic formulae, rules of many games, piano sonatas and moves in a dance. They can also do calculations and problem-solving in their heads and report many of the steps. All this is evidence for the existence of some sort of symbol manipulating virtual machine, no matter how it may be implemented in brain mechanisms, and no matter what other mechanisms interact with it.

Even a besotted lover can dream about what he might have said during the last encounter, plan what he should do and say at the next one, and speculate about the thoughts and feelings of the object of his attention. 'What if' deliberative capabilities enriched by human language seem to be central to all aspects of human life, even if few other animals share them.

8 Conclusion

I have presented a collection of ideas some of which are very speculative while others are largely based on evidence gleaned over many years, including observation of a wide variety of humans of all ages, and what I have learnt from interactions with researchers from a range of disciplines including philosophy, computer science, biology, psychology and brain science, and reading their work. I have not tried to present all the evidence that inspired this work, since that would make the paper far too long, and in many cases I have not kept records. Many of the ideas are not original: much current research in AI involves investigating mechanisms of the kinds proposed here.

With colleagues and research students I am exploring some of the ideas (still in a very simplified form) in computational experiments using the Sim_agent toolkit, which runs under Poplog, and was specifically designed for such explorations. Code and documentation can be found at: ftp://ftp.cs.bham.ac.uk/pub/dist/poplog/

More detailed conceptual analysis, for which there is insufficient space here, would show that familiar mental states and processes such as seeing, deciding, wondering whether, hoping, regretting, enjoying, disliking, learning, planning and acting all involve various subtle and implicit sorts of information processing. (Many relevant ideas are in (Ortony, Clore & Collins, 1988)).

Robots with meta-management capabilities allowing them to attend to internal virtual machine states, including intermediate sensory databases, might discover that they have qualia, and might wonder whether humans are zombies, since they are built quite differently. Work in progress explains this in more detail and argues that when we understand the full nature of that information processing we shall see that it *suffices* to produce what we ordinarily understand by experience, consciousness, etc. (A draft is accessible via my web site, along with other papers elaborating on these ideas.)

Ultimately the ideas will need to be tested not on the basis of the evidence that suggested them, but on the basis of their explanatory power and ability to generate productive research. It takes time to distinguish what Lakatos referred to as ‘progressive’ and ‘degenerative’ research programmes, and there are no simple criteria of adequacy, for reasons I explained in chapter 2 of (Sloman, 1978). There are still many unanswered questions, especially questions about the variety of information processing architectures, what their properties are, which ones could evolve naturally and which can only be produced by explicit engineering design. Investigating these questions requires collaboration between AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics, Philosophy, etc. Such work should not only be of scientific and philosophical interest, but may also lead to new developments in education, therapy and counselling. People often need professional help, but the professionals do not always understand normal functioning of the information processing architectures with which they are dealing, and therefore cannot account for failures and deviations from normality, nor provide help reliably except in a small subset of cases. A deeper understanding of information processing architectures and ways in which they can develop or go wrong could have profound practical significance.

Artificial agents may also need therapy and counselling, for the same reasons as humans. And existing human therapies may fail on them too!

Acknowledgements and Notes

This work has been supported by the UK Joint Council Initiative, The Renaissance Trust, and DERA Malvern. Many have helped with ideas, criticisms, software, and in some cases close collaboration, including: Steve Allen, Luc Beaudoin, Margaret Boden, Chris Complin, Monica Croucher, Kerstin Dautenhahn, Darryl Davis, Stan Franklin, Glyn Humphreys, Brian Logan, David Lodge, Roz Picard, Riccardo Poli, Louise Pryor, David Pym, Tim Read, Ed Shing, and Ian Wright. Anonymous referees were very helpful. Papers by various authors developing these ideas are in the Cognition and Affect Project ftp directory:

ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/0-INDEX.html

References

Beaudoin, L. (1994). *Goal processing in autonomous agents.* PhD thesis, School of Computer Science, The University of Birmingham.

Beaudoin, L. & Sloman, A. (1993). A study of motive processing and attention. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial Intelligence*, pages 229–238. Amsterdam: IOS Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* New York, Oxford: Oxford University Press.

Craik, K. (1943). *The Nature of Explanation.* London, New York: Cambridge University Press.

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

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

Franklin, S. (1995). *Artificial Minds.* Cambridge, MA: Bradford Books, MIT Press.

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

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

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

Mithen, S. (1996). *The Prehistory of the Mind.* London: Thames & Hudson.

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

Oatley, K. & Johnson-Laird, P. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion*, 1:29–50.

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

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

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

Rose, S. (1993). *The Making of Memory.* Toronto, London, New York: Bantam Books.

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

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.* Hassocks, Sussex: Harvester Press (and Humanities Press).

Sloman, A. (1982). Towards a grammar of emotions. *New Universities Quarterly*, 36(3):230–238.

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

Sloman, A. (1993). Prospects for AI as the general science of intelligence. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial Intelligence*, pages 1–10. Amsterdam: IOS Press.

Sloman, A. (1996a). Beyond turing equivalence. In Millican, P. & Clark, A. (Eds.), *Machines and Thought: The Legacy of Alan Turing (vol I),* pages 179–219. Oxford: The Clarendon Press. (Presented at Turing90 Colloquium, Sussex University, April 1990. Also Cognitive Science technical report: CSRP-95-7).

Sloman, A. (1996b). *Towards a general theory of representations, in (Peterson, 1996)*, pages 118–140.

Sloman, A. (1997). What sort of control system is able to have a personality. In Trappl, R. & Petta, P. (Eds.), *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Berlin: Springer (Lecture Notes in AI).

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

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgement to Calculation.* W.H.Freeman.

Wiener, N. (1961). *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge, Mass: The MIT Press. 2nd ed.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell. (2nd edition 1958).

Wright, I., Sloman, A., & Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126.