To appear (with commentaries) in **Philosophy Psychiatry and Psychology**

# Towards a Design-Based Analysis of Emotional Episodes

Ian Wright, Aaron Sloman & Luc Beaudoin[1]
Cognitive Science Research Centre
School of Computer Science
University of Birmingham
September 1995

**Abstract**

The design-based approach is a methodology for investigating mechanisms capable of generating mental phenomena, whether introspectively or externally observed, and whether they occur in humans, other animals or robots. The study of designs satisfying requirements for autonomous agency can provide new deep theoretical insights at the information processing level of description of mental mechanisms. Designs for working systems (whether on paper or implemented on computers) can systematically explicate old explanatory concepts and generate new concepts that allow new and richer interpretations of human phenomena. To illustrate this, some aspects of human grief are analysed in terms of a particular *information processing architecture* being explored in our research group.

We do not claim that *this* architecture is part of the causal structure of the human mind; rather, it represents an early stage in the iterative search for a deeper and more general architecture, capable of explaining more phenomena. However even the current early design provides an interpretative ground for some familiar phenomena, including characteristic features of certain emotional episodes, particularly the phenomenon of *perturbance* (a partial or total loss of control of attention).

The paper attempts to expound and illustrate the design-based approach to cognitive science and philosophy, to demonstrate the potential effectiveness of the approach in generating interpretative possibilities, and to provide first steps towards an information processing account of 'perturbant', emotional episodes.

[1]The first two authors are at the University of Birmingham. The third author is now at Newbridge Microsystems, Canada. This paper was written by the first two authors, making considerable use of ideas developed with the third author and reported in his PhD Thesis (Beaudoin 94)

# 1 Introduction

The human mind, and its underlying engine, the brain, are incredibly complex collections of mechanisms of many kinds, produced over millions of years of evolution. As a result of its origins there are several levels of control, of varying degrees of sophistication. Some, like reflex arcs, are shared with many other organisms. Some, like the mechanisms involved in arousal of various kinds, e.g. those involving the limbic system, seem to be shared with many other mammals. Some, like cortical mechanisms involved in the ability to long for recognition, the ability to enjoy the admiration and respect of others, the ability to be thrilled by a mathematical discovery, and the ability to grieve at the death of a friend, require sophisticated cognitive capabilities, which may be unique to humans. Because many researchers into emotions do not clearly distinguish the different types of phenomena, there is much confusion about what is being studied and what is explained by various theories. Our concern is primarily with mental processes that are typically to be found in human beings, which involve high level cognitive functions and which often have social consequences. They may also, in fact, involve older more primitive mechanisms, though those are not the concern of this paper. (A *complete* theory of the human mind would have to include them.)

This paper has two main goals: (a) to illustrate the design-based approach to the study of some 'higher level' human mental processes; and (b) to make theoretical progress towards a design-based account of certain emotional episodes, namely those that involve a partial or total loss of control of thought processes. Our work derives ultimately from suggestions in (Simon 67), though we have extended and generalised Simon's ideas.

We try to show how a certain sort of information processing architecture, extending ideas in Artificial Intelligence, can serve as a new explanatory ground for some well-known emotional phenomena. Whether the proposed outline architecture is correct, how it might be implemented in neural mechanisms and what the implications of further refinements will be, remain questions for future investigation. The architecture certainly does not yet account for all aspects of grief, and it also leaves unexplained other important mental phenomena which points to the need for extensions to the architecture, which we shall continue to explore.

Section 2, which follows this introduction, is primarily theoretical: subsection 2.1 introduces key ideas of the design-based approach to the study of mind, including the idea of a 'broad but shallow' architecture. Subsections 2.2 and 2.3 sketch our high level design for autonomous agents, including the distinction between highly parallel 'automatic' attention-free processing and resource-bound 'management' and 'meta-management' processes (Beaudoin, 94). Allocation of 'management' resources is one aspect of attention (the management of thought).

Section 3 applies our theory to a concrete example: subsection 3.1 presents an example of a first-hand account of human grieving. We identify certain characteristic features of grieving to be explained (along with other phenomena which we ignore) by any complete theory of human emotions. Subsection 3.2 introduces the notion of self-control and limits to self-control. Subsection 3.3 introduces the type of personal attachment which plays a decisive role in states of grieving. Subsection 3.4 explains in outline how the theory provides architecturally grounded explanations for the phenomena of grief.

Section 4 provides concluding comments and reservations.

# 2 The design-based approach

This section outlines the design-based approach, sketches an architecture for autonomous resource-bounded agency, and introduces the notion of a 'hierarchy' of dispositional control states. Some known gaps in the theory are discussed later.

## 2.1 Ontology and the design-based approach

We assume (a) that information-processing architectures exist, are implemented on human brains, and mediate both internal and external behaviour; and (b) that our methodology allows a systematic approach towards high level functional congruity between artificial, explicitly designed architectures and certain important aspects of evolved, naturally occurring architectures, despite differences in low level implementation details.

Claim (a) underlies much contemporary Cognitive Science and has been argued for or presupposed by many theorists (e.g., see (Miller et al., 70), (Johnson-Laird, 88), (Simon, 67, 69, 95); and (Palmer & Kimchi, 84) for different sub-theses).

Claim (b) is more contentious and depends on finding appropriate levels of abstraction. There are other examples of congruity at high levels despite low level differences. Two physically quite different computing systems may both implement the same virtual machine architecture (e.g., both may be Prolog systems, or both may implement internet utilities, including mail, news, telnet and the World Wide Web). Similarly, it is often taken for granted that general principles of feedback control apply both to natural and artificial systems. What we need to add to this, following much work in Artificial Intelligence, and the ideas in Simon (67), is a level of explanation that involves richer and more profound forms of control of both external and internal behaviour using richer semantic structures and new sorts of control architectures to support various kinds of motivational processes (e.g., see Simon 67, Sloman & Croucher 81, Sloman 87, Beaudoin & Sloman 93, Sloman 93a, 93b, Beaudoin 94, Sloman 94, Sloman, Beaudoin & Wright 94.)

Brains appear to support several rich ontologies at different levels of abstraction. In computing systems, ontologies are often 'stacked' in layers of implementation. For instance, a word processor package that manipulates pages, paragraphs, sentences, words, letters, etc. may be implemented in a 'virtual machine' corresponding to a high level programming language, which, in turn, is implemented in a lower level machine language, and ultimately by quantum physical states of electronic components, with several machine levels in between.

The abstract machines at all levels are compound objects, composed of many different kinds of entities, relations and processes.

Moreover, causal and functional relations may hold between the high level abstract machine structures. (Changes in an abstract data-structure, such as a database of information about employees, can cause changes in what gets printed on pay slips.) These data-structures may have *semantics* in that they refer to individuals and their salaries, etc. We have argued elsewhere that this can include semantics *for the machine* (e.g. Sloman 94).

In the case of human brains we do not know what the layers are. Yet causal relations between abstract structures clearly occur when a person's seeing something causes him to get angry, which in turn may cause him to strike out. The fact that ultimately people, like computers, are implemented in (ill-understood) physical mechanisms is not inconsistent with this. Even physical phenomena are normally explained well above the level of fundamental physics: most people who learn how a car engine works are not taught about quantum physics, but about carburettors, chokes, pistons, etc.

Though a designer often knows a great deal about how a complex system works, it may be impossible for others who merely observe the system, to infer the internal processing. (Sometimes even the designer does not understand all the internal interactions.) This means that any philosophy of science that assumes that theories must be directly or easily testable is ill-conceived: it will fail for complex information processing systems, most of whose behaviour is internal and unobservable. Moreover, even knowing how the system works may not provide a basis for predicting particular behaviours if the behaviour depends not only on the design and current circumstances but also on fine details of enduring changes produced by a long previous history.

When studying systems we have not designed we can, at best, hope for a succession of theories accounting for more and more phenomena, using increasingly powerful explanatory principles, tested in part by implementing the theories in working designs and in part by relating them to the ever growing body of knowledge in neuroscience. There may never be a total ordering of merit among such theories, and the ordering may change over time as new phenomena are discovered. Objections to our approach are often based on a naive philosophy of science, or misplaced 'physics envy'. (For a broader view see Lakatos 70, chapter 2 of Sloman 78, and Bhaskar 78, 94).

The design-based approach draws its inspiration from software engineering and conceptual analysis in philosophy (see chapter 4 of Sloman, 78). It construes AI as a methodology for exploring an abstract space of possible requirements for functioning agents (*niche space*) and the space of possible designs for such agents (*design space*) and the mappings between them (Sloman 94, Sloman 95). Research strategies vary: they may be top-down, bottom-up or middle-out. All are potentially useful. This paper is largely top-down, but we do not exclude other options, e.g. the use of genetic algorithms to create designs by simulating evolutionary processes.

Although it is often assumed that AI is concerned only with algorithms (e.g., Searle

80, Penrose 89), *architectures* are more important. We need to understand global designs for *complete* systems, including their functional decomposition into coexisting interacting subsystems. Early work, still exploring general principles, need not make any commitment to the implementation details of mechanisms; for example, we take a neutral stance towards symbolic or connectionist engines. We start with 'broad but shallow' (Bates et al. 91) architectures that combine many sorts of capabilities (such as perception, planning, goal management, and action). Each capability is initially implemented in a simplified fashion. Subsequent work gradually refines and deepens the implementations.

We claim that *architecture dominates mechanism* (Sloman, 93b), i.e. global design normally determines global capabilities to a greater extent than implementation details. Of course, we must ultimately link designs to neural details and will profit from the 'bottom up' studies of such details, which impose constraints on high level designs. Most of the constraints seem to be quite weak. Exceptions are the high level effects of drugs, which we have not taken into account.

We do not assume congruity between the design decisions 'taken' by evolution under environmental and competitive pressures and those taken by a designer when moving from initial requirements (what the system should do) to prototype design (how the system will do it). Rather we merely claim that the design-based methodology is a source of potential explanatory theories. Such theories will be improved under pressure of criticism, either because of things they fail to explain, or because they explain too much (e.g., capabilities people don't have), or because the designs could not have evolved naturally, or could not be implemented in brains.

Even an oversimplified or incorrect theory that yields a workable design can help our exploration of design space. Comparing it with other more 'realistic' theories aids our understanding of the latter, for we don't really understand any system if we don't know how changing it would produce different capabilities.

Designs satisfying the same information processing and control requirements may possess common design features, whether produced by natural selection or human engineering, just as birds and aeroplanes are both constrained by principles of aerodynamics. Over time the design-based approach may gradually approximate natural 'designs'. This could happen by increasingly taking account of empirical constraints and iterating the development cycle to deepen requirements and extend designs. Such designs can also be tested empirically and compared in more and more detail with their natural counterparts. The total research community is effectively engaged in a parallel cooperative search.[2]

[2]Disparate research efforts are already developing architectures that share a subset of design features given similar requirements. The design of autonomous agent architectures grows apace. For example, Firby's RAP (Firby, 87; Hanks & Firby, 90), the Oz Project (Loyall & Bates, 91; Reilly, 93), Hayes-Roth's intelligent control systems (Hayes-Roth, 91a, 91b, 93a & 93b), the Heuristic Control Virtual Machine (Fehling et al., 89), Georgeff's PRS and Beaudoin's NML1 all share significant design features (for a review see (Wright, 94) and (Wooldridge & Jennings, 95)). However, this phase of agent design may be overturned by revolutionary developments of ideas and techniques. For example, (Brooks & Stein, 93) claim that increased parallelism and the building of situated agents bottom-up will generate such a change.

To summarise: (a) An architecture has causal powers that determine the capabilities of an agent and explain its ability to 'fit' into a part of 'niche' space; and (b) the design-based approach generates candidate architectures that may correspond to naturally occurring high level causal structures implemented upon neural substrates. These candidates can guide empirical investigations to check such claims.

## 2.2 A motive-processing architecture

We now sketch an architecture[3] that is partly similar to Georgeff's Procedural Reasoning System (Georgeff & Ingrand, 89; Rao & Georgeff, 91; Rao & Georgeff, 92), but allows a richer mental ontology, including asynchronous goal generation, more coexisting concurrent sub-mechanisms, a richer set of representations relating to motivators, an attention filtering mechanism and meta-management processes. (Our ideas on all this are still evolving.)

A full account of the proposed architecture would be too long for this paper, so we focus mainly on the processing of motivators. 'Motivator' is used to refer to a subclass of information structures with dispositional powers to determine action (both internal and external). This subsumes desires, goals, intentions and wishes. The precise definitions of these structures and their powers can be given only in terms of the architecture, which is roughly sketched in Figure 1: an impressionistic diagram. For more details see (Beaudoin 94). Within this architecture, motivators can be generated or re-activated asynchronously as a result of internal or external events, and can generate processes of varying complexity, including evaluation, prioritisation, selection, planning, plan execution, plan suspension, and many more. Some of the processes that emerge from such interactions we call 'perturbances', and are described below.

The large shaded area represents 'automatic' processes (associative memory, low level sensory analysis, low level motor control processes, innate and trained reflexes) all implemented in highly parallel dedicated (but trainable) 'hardware'. We assume that such processes include mechanisms shared with many other animals. The larger unshaded area above that represent 'management' processes involved in (among other things) deciding whether new motivators should be adopted or not, assessing their relative importance and urgency, deciding how to achieve them, working out whether they are in conflict, deciding whether to abandon them, reasoning about new information, formulating questions about puzzling information, and so on. (All these processes can be implemented at least

[3]There are are at least two uses of the word 'architecture'; one referring to an abstraction or design that is common to many instances of the architecture; and the other to concrete instances of such designs. We use the former sense, in which an architecture is a collection of features common to a class of entities. Each instance of an architecture is composed of coexisting, interacting substructures with various capabilities and functional roles. A substructure may also have an architecture. The architecture of a complex system can explain how its capabilities and behaviour arise out of the capabilities, behaviour, relationships and interactions of the components. An architecture can be specified at different levels of detail, e.g. at a high level of abstraction the architecture of a house will not include the occurrence of particular bricks, whereas a more detailed architectural specification would.

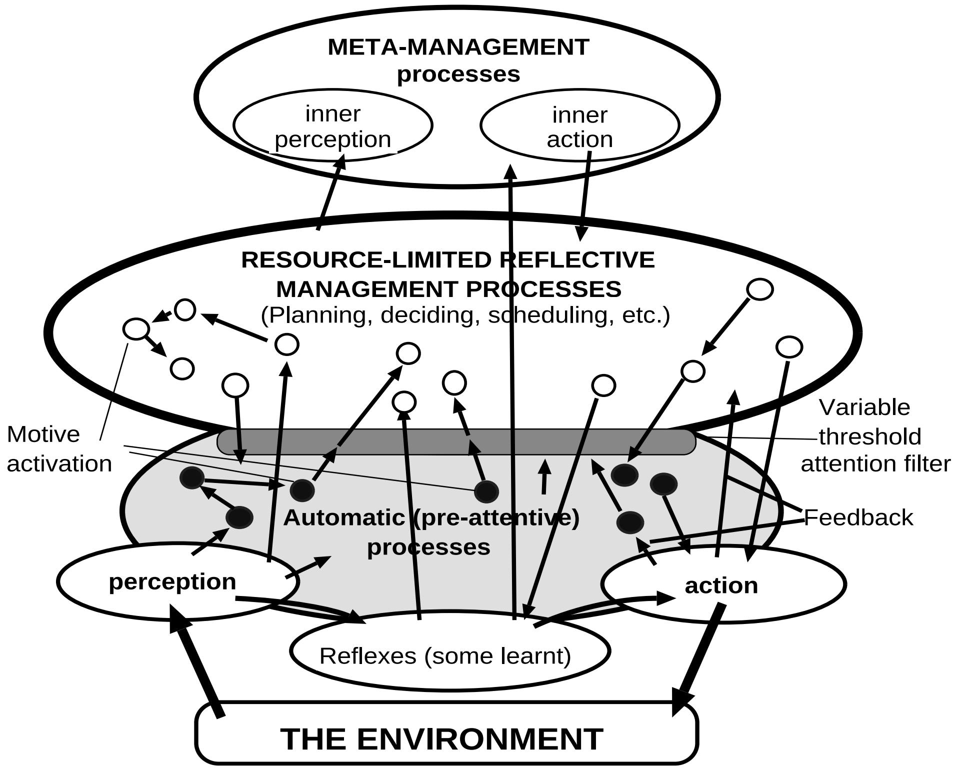


Figure 1: Towards an Intelligent Agent Architecture

approximately using AI techniques.)

The management processes, but not the low level processes, can create, consider and evaluate explicit (representations of) options before selecting between them, for deliberation, planning and problem-solving. This requires the ability to create temporary complex information structures, often with complex syntactic forms, often representing things that are not present to the senses (e.g. future possible actions) and to switch attention between the structures. We suspect that most other animals do not share this capability with humans: their architectures are not sufficiently rich. There are both empirical and design arguments supporting the claim that despite parallelism in the brain such high level cognitive processes are resource-limited: not all potentially useful management tasks can be performed concurrently. So not all new motives, thoughts, problems, can be considered simultaneously. This 'processing-limit' leads to a requirement for some type of filtering of 'attention-distractors', as explained below.

The architecture has the following components (among others), which coexist and operate concurrently. Many of these components require a depth and complexity that we cannot describe here.

- **Perceptual mechanisms.** These are extremely complex systems, which detect potentially relevant sensory episodes and analyse and interpret them in terms of states and processes in the environment, creating or modifying internal representations. A perceptual control mechanism directs sensing operations. (Figure 1 oversimplifies drastically.)
- **Database of 'beliefs'.** Information derived from perceptual representations, internal monitoring and reasoning processes is stored in a 'world model', which acts as a store of information. This will include both specific and general information, including a generic ontology for objects, processes, actions, etc. It need not look anything like current computer databases, for instance if implemented in a neural net. Planning may use temporary 'what-if' extensions to this store.
- **A changing collection of motivators.** A motivator is a semantically rich information structure that tends to produce, modify or select between actions. It typically expresses a motivational attitude ('make false', 'keep true', etc.) towards a possible state of affairs ('short of food', 'warm', 'in danger', etc.), which may be expressed in propositional or non-propositional from. Motivators have various associated information items, including urgency, importance and an insistence value (Sloman, 87).[4]

A new motivator's *insistence level* determines its ability to penetrate a (variable threshold) filter in order to be considered by management processes. *Importance* helps to determine whether it is adopted as something to be achieved, if it is considered. In order not to divert scarce resources, mechanisms assigning insistence values must work using simple 'heuristic' measures of importance and urgency. Computing accurate measures of urgency and importance could be too slow and computationally expensive, possibly diverting management processes which the insistence mechanism is 'designed' to protect. However, insistence measures based on fallible heuristics can sometimes cause 'bad' decisions about what should

[4] There are other motivator components not detailed here; see (Beaudoin 94, Sloman & Poli 95).

and should not divert attention.

• **Pre-attentive and attentive motive generactivators** (Beaudoin 94). These express agent 'concerns' (Frijda 86; Moffat & Frijda, 95). They operate asynchronously in parallel, triggered by internal and external events. They generate and activate or reactivate motivators, and set or reset their insistence level. E.g. the concern to maintain fluid levels may activate a drink seeking motivator. A simple type of generactivator could scour the 'world model' for its 'firing' conditions, and when they are met a motivator is constructed and its insistence level set. Other generactivators may be built into the physiological control system or into perceptual mechanisms. Some may be very abstract and general, e.g. reacting to states where another agent is in difficulty and creating a desire to help. Some are innate, many are learnt and culturally determined.

• **Variable threshold attention filter.** An attention filter protects management processes by allowing only items with insistence values 'higher' than the current filter threshold to divert management processes. When first learning to drive, many find it difficult to hold a conversation simultaneously, as all high level processes are required for driving and the filter threshold is set high. However, attention can still be diverted by, say, the passenger screaming (loud noises can trigger high insistence levels). Later, when the driver is an expert, lower level mechanisms derived from earlier management processes control most of the driving. This reduces the management load, allowing a lower filter threshold and easier diversion of attention. Filtering is content sensitive, with different thresholds for different contents (Beaudoin, 94): A baby's cry can divert attention even when other loud noises do not. Motivators that fail to surface may remain available, so as to take advantage later of a lower filter threshold, or they may die unless continually reactivated by the relevant generactivators. Motivators survive until they are satisfied or decay.

• **Motive management.** Once a motivator has 'surfaced' it can cause many diverse and complex processes, including: assessing (evaluating its importance, urgency, costs and benefits etc.), deciding (whether to adopt the motive, i.e. form an intention), scheduling (when and under which conditions to act), expansion (how to do it, i.e. planning), prediction (projecting the effects of hypothetical decisions), detecting motive conflicts, detecting opportunities[5], abandoning a motive and changing filter thresholds. Information about current motivators (including intention structures discussed below) is stored in a 'motivator database'. Only a limited amount of parallelism is available for management processes.

• **Plans and other 'databases'.** The system will have many short term and long term memories containing information about current, future, or possible activities. In particular, certain goals require plans, and some re-usable plans will be stored as well as information about how to create new plans. Other information stores will include collections of particular skills, e.g. linguistic skills, mathematical skills, skills relating to social activities, games, one's job, etc. Some of this will be stored in an opaque form among the pre-attentive mechanisms. Some will be accessible to management processes. Some will be

[5]See (Pryor & Collins, 92).

innate, others learnt, some fixed and others modifiable.

• **Meta-management.** A meta-management process is any goal-directed process whose goal refers either to a management or to a meta-management process. Deciding whether to decide whether to adopt a goal, deciding which management process to run now, and deciding if too much motive-swapping is occurring, are all examples of meta-management processes, where motive management itself is the object of control. Meta-management requires some degree of ability to monitor and change the management processes. Conflicts can lead to 'loss of self control'. Again, resources are limited.

• **Plan execution and effectors.** Selected motivators are *intentions*. Executing them may occur with or without planning, with or without high level management, with or without monitoring. External actions are dispatched to an effector driver that controls agent actions within the environment. Internal actions may produce a succession of management states.

• **Self-monitoring mechanisms.** A variety of types of self-monitoring are required. Meta-management monitors include those that can detect global states of the management processes, such as noticing that new motivators are surfacing faster than they can be processed. This could lead to a raised filter threshold. Other states worth detecting include repeatedly considering the same goals or problems without making progress in solving them (described as 'maundering' in Beaudoin 94). A concept-learning mechanism would be required for discovering useful new ways of describing internal states. Humans can detect inconsistencies between internal management tendencies and personal ideals.

• **Pleasure and pain mechanisms.** Pleasure and pain have not been analysed adequately yet. They seem to involve phylogenetically old structures in the nervous system that are deeply implicated in mechanisms for both learning and online control: i.e. whether current activities should be maintained or terminated. Some people doubt that an information-processing architecture can fully account for such subjective states. For now we assume that there are mechanisms producing some motivational control states that are pleasurable in the sense of involving a disposition to preserve or extend some current state or activity, and some that are painful, involving dispositions to terminate or reduce some state or process. Some of the control states in the architecture may have a positive or negative valency that arises out of genetically programmed drives (e.g., seeking after novelty, urge to procreate and form attachments) and states of body arousal. Others involve learnt evaluations. For pain and pleasure to be *experienced* requires self-monitoring mechanisms that detect these 'preserve/terminate' control states. The mechanisms may be linked in that termination or reduction of pleasure inducing states functions as a pain inducing state and *vice versa*. In some cases these links will need to be built up through learnt associations. In others they may be 'hard-wired' in the architecture.

• **Global control mechanisms.** An agent sometimes needs to modulate global features of its processing, for instance speeding everything up in times of extreme danger, slowing things down when losing energy too fast, generally being cautious when the environment is unfriendly, not wasting time on caution when the environment is generally friendly.

Other more subtle global changes may be required when the social context changes, for instance when an individual of higher status is present. These global changes of state seem to be related to the colloquial concept of 'mood'. In humans it seems that some of these global mood states are implemented in part at the chemical level (as shown by effects of drugs and hormones). Some control mechanisms may be 'global' only relative to a subset of a system: e.g. global changes in the ease with which unsuccessful attempts are abandoned.

We acknowledge that this design sketch is speculative, vague and subject to revision in the light of implementation problems or empirical evidence. Detailed designs may vary, e.g. some using rule-based systems, others neural nets. The components may exist within a virtual (abstract) machine whose components and processes do not correlate in any simple way with physical structures and processes.

A small subset of the architecture has been implemented (Wright, 94) and tested in a simulated domain, using very simple management processes. A more elaborate implementation is in progress, using a toolkit developed at Birmingham (Sloman & Poli 95).

The architecture is 'broad' in that it combines many different capabilities. It is 'shallow' in that the components are not themselves specified in any great depth. For testing purposes simple implementations are used at first. Much of the specification is still provisional and speculative, in that work on implementation may reveal serious problems. It is also speculative in that empirical checking has not yet been attempted (and may be very difficult.) Nevertheless we think the architecture already has explanatory power.

## 2.3 Varieties of control states

A mind can be viewed as a control system with a rich collection of control states (Sloman 93b), involving dispositions to respond to internal or external conditions with internal or external actions. Some control states are implemented in high level abstract machines, others in various sorts of neural states, or in states of chemicals in the blood or nervous system. The described architecture provides a framework within which many control states can coexist and interact.

Examples of control states referred to in folk psychology are beliefs, images, suppositions, desires, preferences, intentions, moods, learnt associations, innate or trained reflexes, personality traits, and emotional states. The dispositions may be of a high order (i.e. dispositions to invoke dispositions to invoke dispositions. . . (Ryle 49).)

Every architecture supports a family of concepts describing its states and processes. Exactly which control states are possible depends on the architecture. A leaf's movements may speed up or slow down, but it cannot be described as 'braking' because it lacks the architecture required for self-modification of speed, namely one sub-component which acts on others to slow them down.

The architectural presuppositions of ordinary mental terms are far more deep and difficult to analyse. By defining states in terms of an architecture that supports them we can offer a rational reconstruction of many ordinary mental concepts, just as theories of the architecture of matter rationalised concepts of physical and chemical stuff (e.g., 'iron', 'carbon', 'water', etc.). Folk psychology concepts may implicitly (pre-theoretically) refer to mechanisms of the general sort we describe, even though they contain contradictions and confusions. We use them only as first approximations. Later we hope to offer something more systematic and analogous to the periodic table defining chemical elements.

Every control state has structure, aetiology, powers, transformation capabilities, liabilities, and in some cases semantics. These ideas include analogues of the linguistic notions of syntax, semantics, pragmatics, and inference (Sloman, 94). For example, a motivator may have a complex internal structure (syntax), a content (based on that structure) referring to certain states of affairs (semantics), a functional role, i.e. dispositional powers to determine internal and external actions (pragmatics). It may also enter into processes that derive new motivators or plans (inference), it may be brought about or triggered in various ways (aetiology), and may be modified, suppressed or terminated by other processes (liabilities). Some control states are short-lived (e.g., a motivator which is immediately rejected, or whose goal is quickly achieved.) Others endure.

A control state may exist but not express its causal powers, or may express its causal powers but not determine action. Observed behaviour will generally fail to indicate the full richness of the underlying control states, especially the dormant states.

Certain high level, long-term (motivational) states (such as selfishness and other personality traits) are difficult to change, are modified only very slowly by learning and possess pervasive causal powers corresponding to their abstractness and general control applicability. Low-level, short-term states, such as a desire to scratch an itch or to repay a favour, are easier to change and have more direct causal powers corresponding to their specificity and narrower control applicability.

Some control states will gradually change their status over time, via a process we call *circulation* in which control states move around the system. Useful control states 'percolate' up the hierarchy, gaining in abstractness and resistance to change, and increasing their field of influence. Defunct long-term control states may gradually lose influence through lack of use, leaving only a few relatively specific active instances. An example may be standards and principles from a culture one has left.

Control states may be qualitatively transformed during circulation, for instance acquiring more general conditions of applicability. Higher level general attitudes such as generosity of spirit, may also spawn derivative specialised control states such as favouring a certain political party — another aspect of circulation. Internal connections between control states will set up suppressive or supportive relationships, dependencies, mutual dependencies and, occasionally, dead-locks.

The net effect of all this is a process of 'diffusion' in which the effects of a major motivator are gradually distributed in myriad enduring control sub-states, i.e. in motive generators, plan schemata, preferences, and predictive strategies. In some cases the effects will be *irreversibly* embedded in a host of reflexes and automatic responses. (This is loosely analogous to the process of compiling a high level language.)

The totality of control states is a dynamic structure with complex internal relations and many levels of control. High level control states amenable to some degree of self-monitoring are among the requirements for any architecture underlying human-like capabilities. Where there is appropriate self-monitoring (described below) the system may also acquire self-knowledge that triggers negative evaluations and new high-level motivators attempting to change some aspects of the system itself, whether previously learnt cognitive reflexes or high level attitudes.

A full specification of the architecture is work for the future. Nevertheless, we can begin to use it schematically to explain certain aspects of human mental functioning by mapping familiar experiences and processes into control states and processes that could occur in this sort of architecture; for instance, affectional bonds and 'perturbant' emotional states.

# 3 Architecture as explanatory ground

In this section we apply the architecture to a concrete emotional phenomenon: grief at the loss of a loved one. This is not a comprehensive 'theory of grief' but an application of a 'broad but shallow' architecture to the understanding of a complex cognitive phenomenon. It will be seen that a *complete* architecture can account better for the diversity of internal and external behaviour that can occur during grief than an explanation focusing only on some emotional mechanism.

## 3.1 Personal reports on grief

The following quotations (a) to (p) are based on extracts from a first-hand account of grief, edited to disguise the origin. After each item we make some preliminary comments in italics, relating it to the architecture. Later we expand on these comments.

**(a)** When the person finally passes on, the hurt is like no other. ... Now the memories of the good times seem to be like a film playing in my head.

*This is one of many examples of partial loss of control of thought processes in emotional states: thoughts, motives and memories relating to the object of the emotion intrude and may even dominate thought processes, making it hard to attend to other matters, even those judged to be urgent and important.*

**(b)** XXX called us the day after his friend died suddenly. He was broke, alone, grieving and gravely ill, staying at the hospice where he'd spent the last month by his friend's bedside. I flew to YYY and brought him home, where he spent the next ten weeks in hospital. He was *so* ill that it was hard even for me – a medically sophisticated person – to comprehend and accept.

*New facts may not easily fit into one's belief system. This is linked to the pain they cause: not physical pain but mental pain, a major phenomenon to be explained.*

**(c)** I started grieving the day I received the phone call, continued grieving the day I met him in the hospice in YYY, and all the time I sat by his bedside back in ZZZ. It tore me apart to see what this disease had done to him. . .

*Grief can be both extended and highly disruptive. Notice that 'tore me apart' does not describe physical or physiological processes, but in a highly metaphorical way describes mental processes. Unpacking that sort of metaphor is part of the job of a deep theory of emotions.*

**(d)** At least, that's all I wished for silently over and over again as I sat next to him in the hospital.

*Powerful new motivators sometimes cause futile behaviour.*

**(e)** When he died, I was initially in shock.

*Shock can be physical or mental. We can't tell which was referred to here but it could easily have been both. One of the things to be explained is how mental events, e.g. learning about the death of a loved one, can produce profound physical effects. But we are primarily concerned with mental effects.*

**(f)** And then there are details like writing obits, funeral arrangements, meeting his family, etc., which keep you busy. And then the grief resumes.

*The ability of grief to control processing is a disposition in the sense of Ryle (49). (Dispositions and capabilities can exist for a long time without being manifested because their 'triggering' conditions do not occur, like the fragility of a wine glass, or countervailing processes prevent their expression, like a dam wall preventing a potential flood.) Grief involves dispositions that may be temporarily ineffective because urgent and important tasks manage to hold attention. But during that time the disposition remains and when there's a chance it regains control. From this point of view it is misleading to say 'the grief resumes'. It was there all along, though its manifestations were temporarily absent.*

**(g)** It has been two-phased for me. This summer I kept confronting the *unreality* of it all. XXX, dead? I was wracked with insomnia. Every night as I'd try to go to bed, indeed, during much of the day too, I would replay the details of the previous three months over and over and over again.

**(h)** Each recollection was full of almost indescribable pain, but it was also fresh with his presence, something I cherished and which I'd missed daily for almost four years before. I wish I'd had some 'good times' to replay too, but they were from an earlier time, ... and they were crowded out by the intensity of the recent months.

*This helps to bring out some of the complexity of human emotional states: mixtures of different states are not uncommon. Several different dispositions can be 'fighting' for control and the balance may shift from time to time. Not long ago, in a BBC radio* *interview, the captain of a women's yacht in a round the world race described her emotional state on arriving at the final port. It was an enormously complex mixture of: pride in achievement, sadness that it was all over, joy at the prospect of seeing loved ones again, delight at the prospect of eating (rations had been exhausted two days earlier), regret that they had not won the race, happy memories of teamwork and obstacles overcome, sadness that the team would now be parting, and so on.*

**(i)** I managed to keep up a bit of a social life during this summer.

*In a state of grief, interactions that would normally be taken for granted and enjoyed may be difficult.*

**(j)** My overall mood was somber, but I was at least willing to be distracted by friends...

*Sometimes when self control of thoughts is hard, external help is effective. A mood is a type of global state, which need not have a semantic focus, unlike emotions. Moods also need to be explained by a theory of the architecture of a mind.*

**(k)** I became more and more of a hermit. Things lost their savour for me, and I withdrew.

*Some of the high level control states, which we may think of as forming the personality, can be profoundly changed by grief. This may affect a wide range of preferences, choices, strategies, plans, and behaviour.*

**(l)** Grief isn't something you can show for too long. People are uncomfortable with it. They will indulge you only for so long, so you just hold it inside, slog along and try to get on with life.

*Besides the control problems which relate to tasks and goals that form part of normal life, the observation that grief has undesirable social consequences can generate a new second order control problem – namely, not allowing the grief to be shown or to intrude in social interactions. Many emotions generate second-order motivations relating to the control of those emotional states themselves.*

**(m)** The holidays and the new year are a natural marker, and they've kept me busy and distracted me.

*Another example of external help with the control problem.*

**(n)** I try to reintegrate myself with my social circle, and start seeing friends again more regularly. But I'm worried; I don't want to let go of the grief. Sometimes I think it's all I have left of XXX.

*Motivation in relation to the control of the emotional state may be very mixed. The griever who manages to control the grief and get on with life may also suffer feelings of guilt or regret because another motivator involving one's duty to the deceased seems to be violated, or because of the feeling that overcoming the grief would itself be a sort of loss of contact with the deceased.*

**(o)** I don't want to enter his bedroom one day with indifference, and wonder how we might use the space. I don't want to stop crying every day when reminded of another piece of our time together. Well, I *do*, eventually, and I know I will, but I am not comfortable with this yet.

*Another example of mixed and conflicting second order motivation relating to the state of grief.*

**(p)** This was a gruelling, soul-grinding and exhausting year. It has been the worst year of my life.

*Emotional states normally involve evaluations. Powerful emotions often arise out of 'intense' evaluations. The emotional states themselves can be evaluated and contribute to the judgement that something bad, or something good, is happening.*

Those who have grieved or had close contact with a griever may empathise with these extracts. Others will also understand, for these are pervasive phenomena and play a major role in many works of art. But we need to get beyond empathy and folk psychology to understand the architecture underlying grieving and other affective states. Our proposed architecture provides a provisional first draft explanation of certain characteristic cognitive features of mourning in terms of causal relations between underlying information processing mechanisms.

The following are among the surface phenomena of grieving, illustrated by the above extracts:

1. The continual and repeated interruption of attention by memories or thoughts relating to the friend's illness and death; i.e. loss of 'normal' control of thought processes (explained below). See [a, d, g, h].

2. The difficulty of accepting the fact of the friend's illness and death. See [b, e, g].

3. The disruptive effect on normal, day-to-day functioning. See [c, e, g, k, p].

4. Periods of relative normality when grief is 'backgrounded', sometimes because external factors help one regain 'normal' control. See [f, i, j, m].

5. Attempts to 'fight' the grief. [g: 'try to go to bed', l: 'hold it inside, slog along and try to get on with life'].

6. Second order motivators, some of them involving evaluation of the grieving state as good or bad, including, in some cases, wishing the grief to continue. See [n, o].

7. The subjective 'pain' experienced by the mourner. There may be mental pain as well as bodily disturbances. See [a, h, p].

8. Crying. See [o].

These are not all the possible symptoms of long-term grief and neither are they unique to grief: excited anticipation of a long awaited event could also disrupt normal day-to-day functioning. Anger can also be 'backgrounded' when other demands control one's attention. Guilt feelings, or an undesirable infatuation may also be fought against.

Many theories of emotion concentrate either on the neural substrate, external behaviour or externally observable changes (facial expression, posture, muscular tension, sweating, etc.), i.e. evolutionarily old mechanisms shared with other animals. We are deliberately ignoring most of these, indeed regarding them as only marginally relevant, since, in principle, the mental phenomena that we are concerned with could occur without these other accompaniments, for instance, in beings from another planet whose mental functioning and social life were much like ours despite considerable bodily differences (see Sloman 92).

We do not wish to argue over whether these phenomena are or are not part of the definition of the word 'emotion'. Many words of folk psychology are notoriously ill-defined (Read & Sloman, 93; Kagan, 78), both in colloquial use, and in scientific contexts where different theorists offer different definitions. So instead of arguing over definitions we merely identify certain types of familiar phenomena and then ask what sorts of mechanisms might explain them. Since we are interested in information processing explanations of information processing phenomena we are not concerned that there are no behavioural or physiological definitions of the phenomena: for example, there are no behavioural or physical definitions of many of the states and processes in information systems that are of interest to software engineers. They are rightly not concerned that their work does not conform to narrow (and often misguided) criteria of what constitutes 'science'.

Our interest is in information processing theories of affective states because we expect they will be both theoretically and practically enormously fruitful in the long run, even though most of the people who should be most concerned, such as therapists, counsellors and educationalists, are usually unaware of them on account of current training regimes.

For those that remain uncomfortable with a functional analysis of affective states the following distinction may be helpful: many mental terms have, besides their physical basis, two non-physical aspects – phenomenological (subjective feelings states) and psychological (functional role) (Chalmers, 96). For example, 'pain' certainly denotes an unpleasant feeling state, but it also refers to a functional role of a pain control signal that leads to the withdrawal or avoidance of an aversive stimulus. In our discussion the 'phenomenological problem' is factored out and placed to one side. Our emphasis is on what causal mechanisms *do*, not *what it is to be like* these causal mechanisms.

## 3.2 Self-control and perturbance

A grieving person is partly 'out of control'. Explaining this requires an analysis of what it is to be in control. The ordinary notion of 'self-control' is not a unitary concept admitting of a unique analysis. Like many mental concepts it covers a variety of cases and further study is required to investigate how many of them can be accommodated within the design-based framework. In particular, not all agent architectures can support a distinction between being in control and not being in control of one's thought processes. For instance it is not clear that a rat ever has control of thought processes. In that case it cannot lose control. If that is so, the sorts of processes we are discussing cannot occur in a rat.

What kind of system can have control of its thought processes? A partial answer is that the system must be able to have goals relating to its own thought processes, and it has control when everything that occurs in it is consistent with all the currently adopted management and meta-management goals. This does not imply that everything that happens is *generated* by those goals, for that would rule out intrusions, such as feeling hungry, or a new desire to help someone in trouble. These can arise without contradicting one's view of what one should be like.

Our partial answer needs to be expanded in various ways. One is to allow that there need not be only *one* coherent global set of goals, preferences, etc. since some people seem to change personality from one context to another, like the kind father who is an aggressive car driver. The architecture might allow a number of *different* sets of mutually consistent high level dispositions that co-exist, though only one set is active at a time (as sketched in Ch. 10 of Sloman 78).

Detection of incongruent states requires some sort of self-monitoring of the global 'picture' and an explicit evaluation of it as fitting or not fitting the agent's ideals, long term objectives, or previous decisions regarding (for example) what to think about, or which desires are unacceptable. If a substantial amount of what is happening at any time is inconsistent with the agent's dominant evaluations and preferences, then the agent is, to that extent, partly out of control, even if all the disturbances and disruptions are generated entirely within the system, e.g. from lower level automatic, non-attentive, processes, or reflexes in the management system. Addictions are an extreme case.

We use the term *perturbant* for a state in which partial loss of control is due to the continual surfacing of postponed or rejected, or unwanted, motivators (Beaudoin, 94), or possibly disruptive thoughts, images, and the like (e.g., a catchy tune that won't 'go away'). Such disruption can interfere with the management of other, important goals. This is the type of information processing state that the 'attention filter penetration' theory posits as characteristic of many emotional states (Simon 67, Sloman & Croucher 81, Sloman 87, 92). It is what the filter normally prevents.

Perturbant *states* differ in several dimensions: duration; whether the source is internal or external; semantic content (what is referred to); type of disruption (it could be due to a goal, thought, or recollection); effect on management processes; frequency of disruption; positive or negative evaluation (compare grieving with being unable to stop thinking about the victory one has recently won — grieving and gloating have much in common); how the state develops; whether and how it decays; how easily it can be controlled, and so on.

Perturbances (like 'thrashing' in an overloaded computer operating system) are side-effects of mechanisms whose major role is to do something else (just as thrashing arises from the paging and swapping mechanisms in the operating system). Perturbances arise from the interactions between (a) resource-limited attentive processing, (b) a subsystem that generates new candidates for such processing and (c) a heuristic filter mechanism. These design elements arise from the requirements for coping with complex and rapidly changing environments. Perturbances do not arise because of some special perturbance generating (or emotion generating) mechanism. Thus it is misguided to ask what the *function* of perturbant states is or to postulate a perturbance mechanism.

Our architecture was described above as concerned with processing goals. However, other things should also be able to penetrate the filter mechanism and divert management processes, such as items of factual information about current or past events. Sometimes new information generates motivators as a result of high level reasoning which interacts with generactivators in the management system, for instance, discovering that one's bank balance is very low. Also new factual information may be required to interact with current plans and goals, for instance, showing that a goal has been satisfied and is now redundant, or that a short-cut is now available (Ch 6 of Sloman 78, Pryor & Collins 92). This means that perturbant states can be produced not only by new motivators but also episodic recollections, Simon's 'cognitive associations' (Simon, 67), or motivationally neutral thoughts.

Can perturbant states be controlled? Self-control requires self-knowledge. Control of motive management processes requires both knowledge of what they are and motivators expressing what they should be: a partly normative theory of the self, which can be used in detecting and categorising internal states, just as knowledge and desires are used in external perception and action. In humans there are wide differences in degree and kind of self-knowledge, though mourners usually know that they are mourning. People are aware of some aspects of their own mental state which they can describe with more or less precision and artistry.

Monitoring of management processes may be based on detection of such variables as the rate of surfacing of new motives, aspects of their semantic content, their effects on processing, the current filter threshold level, the rate of success of plans. Self-descriptions thus produced may vary in scope and content, as in 'My last goals were not achieved', 'I am usually calm', 'I am happy now', 'I am grieving over my lost brother', 'I feel depressed about life', 'I am worried about a job interview but looking forward to a hot bath'. Information sampled from the stream of management processing can be stored to provide data for the self-control mechanism.

Such 'internal perception', where the self-control mechanism is the perceiver and motive processing is the environment, requires non-intrusive or 'transparent' monitoring that does not change the processes observed, just as looking at a scene does not change the scene. This is also a requirement for 'software agent' architectures in computing systems, where agents need to be able to observe and react to actions of other agents in the system. Full transparency is difficult, or even impossible to achieve: architectures are always partly opaque. Attempts to 'trace' everything in a software system break down when the tracing processes are traced. Self-knowledge is always incomplete, for otherwise there would be an infinite regress, for instance, of beliefs about beliefs about beliefs ...

The requirement for transparency of motive management processes may be satisfied to different degrees by different designs (e.g., a neural net monitoring a lower level neural net by 'sampling' connections, or procedures that can be interrogated about their functionality without disrupting their operation (Rozas, 93), or blackboard architectures with globally readable information spaces, or towers of self-referential processors (Bawden, 88), and so forth). Exactly which forms of transparency exist in which organisms is an empirical question. Which forms are useful for which purposes is a design question.

Access to information is not enough. The system must be able to produce adequate descriptions or categorisations. So self-reflective concepts are required for self-monitoring, including both concepts learnt from the utterances of others ('Don't be angry', 'Are you upset?') and concepts induced internally via generic concept-learning mechanisms. Thus both socially acquired descriptions and more or less idiosyncratic categories will be used to refer to internal states.

Besides self-monitoring, self-control requires selection of remedies and their application. Like medical treatment self-control uses a three-stage process of diagnosis (identifying the problem), selecting strategies (selection of medication and prognosis), and applying them within the system (treatment). There are numerous possible strategies, with details differing between architectures. Some management strategies will be socially inculcated while others are induced by generic learning mechanisms from consequences of self-management decisions.

Such a meta-management system will have, or develop, ways of classifying internal states and associating appropriate strategies with them — often different strategies in different contexts. For example, raising the attention filter threshold to counter perturbance may be useful when it is difficult to keep up with new goals, but could be disastrous in situations where some highly urgent and important new information could turn up.

Some control strategies treat symptoms: for example, counting imaginary sheep to counteract insomnia. The sheep-counting strategy works by concentrating deliberative resources on a motivationally neutral task that requires full attention. The cause of insomnia remains but its ability to divert resources is diminished by a rival task. (This often fails.) Another case of treating symptoms is turning to drink in order to dull painful thoughts and emotional responses: a strategy which depends on close coupling between chemical processes and high level control states.

More powerful self-control strategies may be more difficult to apply: removing the cause of the problem (dismantling an attachment structure (defined below) or abandoning an unfulfilled desire — often difficult or impossible), suppressing external symptoms while internal turmoil continues (the stiff upper lip syndrome), diverting the system via tasks and external contexts that cause the attention filter threshold to be raised, undergoing explicit or implicit training (or therapy) that alters some of the links in the system (e.g. disabling some generactivators or changing them so that they assign lower insistence levels). Some of these are temporary palliatives, whilst others have long term effects. Some may lead to new re-integrated personalities, while others leave potentially harmful tensions within the control system.

Self-control is always limited. The difficulty in controlling perturbances may have different causes. (a) Self-control mechanisms have limited causal powers (hence the tension between the ease of representing what you want to become and the difficulty of becoming what you have represented (Smith, 86)); (b) self-monitoring is limited by architectural opacity; (c) the system may lack the concepts required to categorise internal states and express appropriate remedial strategies; or (d) adequate control strategies may not yet be available.

Powers of self-control mechanisms can be extended by various kinds of training and practice. Opacity can sometimes be overcome by extra reasoning (after the event), like inferring the shape of an occluded object during visual perception. New meta-management strategies can be learnt through trial and error, social influences, or possibly therapy.

Psychological evidence exists for self-modelling, reflection and self-repair, e.g. (Kuhl & Kraska, 89). Selecting or creating control strategies requires a decision mechanism that can synthesise different kinds of information — current state of motive management, environmental context, current goals, theory of the self and so on — and then decide on a strategy. This ability will vary between individuals.

## 3.3 Attachment structures

A raised surface can leave an impression on human skin; similarly, interaction with another person will leave an 'impression' on mentality. Before the advent of information processing architectures this metaphor could not be unpacked.

Bowlby's theory of attachment (Bowlby, 79 & 88) attempts to explain how affectional bonds are created and the effects that occur when such bonds are broken. Although criticised in recent times (particularly the emphasis on maternal deprivation in childhood to explain subsequent problems in adulthood; see (Smith & Cowie, 91), Ch.3) attachment theory is still used to account for both childhood and adult mourning in clinical psychology. We shall explore this theory within our proposed architecture, showing how processes described above as percolation, circulation and diffusion allow a distributed multi-component 'structure of attachment' to an individual to develop and influence subsequent processing. Being deeply entrenched at many levels within the control hierarchy it manifests itself in multifarious ways when its object dies.

The perceptual system and belief systems of an agent will include information about other agents, including information about how to recognize them and what behaviour to expect from them in various situations. This may also include evaluations such as 'X is a good person' or 'X is dependable'. Interaction with X will lead to creation of motive generactivators expressing motivational attitudes towards X. Over time, enduring control states pertaining to X will be generated that interact with higher level attitudes and personality traits within the hierarchy of dispositional control states.

For example, various preferential mechanisms may be set up ('prefer to be in the company of X'), which could function as motive comparators; or unfocused and abstract wishes ('wish X is always happy and well'); also desires ('desire to holiday with X sometime soon' or 'desire to spend more time with X'), hopes ('hope X likes me', 'hope X enjoys my company', 'hope that X will remain close') and aims ('maintain friendship with X', 'avoid arguments with X'). High level preferences may generate lower level motive generators: for example, preferring to be in the company of X could generate the aim to maintain the friendship of X. In other words, a diverse collection of control states with complex interrelations and dispositional powers will be created alongside factual information collected through interactions with X.

The evaluations, generactivators and motivators will be positive towards some individuals, negative or neutral towards others, and with varying degrees of strength, possibly involving the pleasure and pain mechanisms. Depending on the particular combinations of evaluations and other attitudes towards an individual, the death of that individual may cause grief or some other kind of emotional state, or no emotional state. A pre-requisite for grief is strong positive evaluation, though that is not sufficient, for the death of a person whom one admires or respects greatly need not cause grief. Something more is required, namely the sort of entanglement of personalities commonly labelled as 'love' (another highly ambiguous term).

The kind of loving that potentially leads to grief, which we are calling 'attachment', is a very complex mixture of states that develop over time through mutual interaction. It will involve many dispositions, including dispositions that produce pleasurable feelings in the company of the person, displeasure when the person is absent or harmed, and so on. Besides feelings, attachment structures can generate new motivators relating to the person, e.g. when information is received about that person's needs, successes, failures, suffering, etc. All these new (dispositional) control states generated by the process of attachment will, over time, integrate into the existing control network: the process we have called 'diffusion.'

These control states involve many dispositions, including potential influences on both pre-attentive and attentive processes. The former include (a) a tendency for new motives to be generated pertaining to X; (b) assigning relatively high insistence values to motives concerned with X, particularly if there is a serious problem involving X that needs urgent attention; (c) allowing the filtering mechanism to give preferential surfacing conditions to X-related motives, as all things relating to person X are deemed important (compare a mother and her baby); and (d) new links between phenomena involving X and the pain and pleasure mechanisms.

Effects on attentive, management processes include: (e) new dedicated decision procedures with regard to X (e.g., skewed importance, urgency and cost-benefit computations that raise the priority of X-related motives); (f) creation of unusually detailed (possibly unrealistic) predictive models about X's behaviour and preferences; (g) clusters of management procedures that manage X-related motives by combining model-based information, current and new goals to form new intentions; (g) a relatively high proportion of items in the goal database concerned with long, medium and short-term intentions relating to X, in various states, such as conditionally suspended, postponed, ongoing etc.; (h) an unusually high proportion of intentions that are long term mutual or joint plans predicated on the co-operation and continued proximity of X; and possibly (i) new motive conflicts pertaining to X, e.g. the combination of preferring to be in the company of X, wishing that X is happy and believing that X wishes to be alone, or loves another. Great novels and real human tragedies often depend on such conflicts and the processes they generate.

Meta-management procedures may be generated or altered during the growth of attachment. For example, management tasks of the form 'decide whether to adopt motive M' may come to be be handled as soon as possible if M pertains to X. Relatively more computational resources may come to be allotted to any decision procedure concerned with X. A host of plan libraries expressing the utility of certain actions for achieving goals with regard to X will be formed, which facilitate planning relating to X; and 'chunks' of actions that appear to be efficacious when dealing with X may be abnormally strongly reinforced and lead to stereotypical and positively valenced patterns of interaction.

Summary: an 'attachment structure' relating to an individual is a highly distributed collection of information stores and active components embedded in different parts of the architecture and linked to many other potential control states. When an attachment structure concerning individual X exists in an agent, almost any information about X is likely to trigger some internal reaction. In particular, information about good things or bad things happening to X may trigger reactions whose strength and pervasiveness depends on how good or bad they are. Death is a particularly bad event.

In this paper we shall not attempt to describe the process of *detachment*, in which the attachment structure is gradually dismembered and possibly replaced by a new complex set of beliefs and motives relating to X, consistent with X no longer being alive. This drawn-out process is part of a self-control strategy for overcoming perturbance, albeit a long-term strategy that attempts a design change to achieve its ends. This process can be analysed into many sub-problems — for example, the structure of attachment would need to be inspected for the sources of perturbance, blame assigned, a modification of the structure selected, and repair work effected followed by some kind of verification process to check whether the modification had resulted in an improvement.

These are all extremely sketchy ideas that need to be developed in the light of a more detailed specification of the architecture and its 'learning' capabilities. Yet all the proposed control states are of a type that we claim could be implemented (with difficulty) in a suitably rich architecture based on AI mechanisms, possibly using a mixture of neural nets and symbolic processes.

## 3.4 An architecturally grounded analysis of grief

We now return to the symptoms of grief and attempt an architecturally grounded interpretation of the surface phenomena in terms of an attachment structure in the griever towards the deceased.

1. The continual and repeated interruption of attention by memories or thoughts relating to the friend's illness and death.

We have already described (Section 3.2) how the architecture permits perturbant states when heuristic mechanisms designed to prevent disturbance of resource limited processes continually 'let through' motivators and thoughts that divert attention from highly valued activities. Following bereavement, cyclic processes could occur, involving, among other things: motives relating to the dead person, generated by long term attachment structures, including desires for the person to be alive, or present, or unharmed; wishing one had done things that might have prevented the death; recalling that the person is dead; rejection of the motives as therefore inappropriate or futile; evaluating such rejection as undesirable; reminders of relevant information concerning the person, such as might be important if the rejected goals were being acted on. These and other interactions might all reverberate throughout the system because of the deeply entrenched information structures and the powerful triggering effect of news that the worst possible harm has already happened to the person.

Some of these events may set off a stream of deliberative thought (or meta management processes) attempting to re-orientate extant desires, intentions and plans, to cope with the changed circumstances. This process could also trigger the recall of associated memories in the form of different sensory modalities (images, smells, sounds etc.), as well as triggering a host of embedded generactivators waiting in the wings.

The structure of attachment explains why motives relating to X are likely to disrupt attention. (a) X-related motives will be given high insistence values because the relationship with X is strongly positively valenced and X has suffered great harm. (b) Exception fields within the filtering mechanism may provide preferential surfacing conditions for X-related motives and thoughts. (c) Higher level, abstract control states expressing attitudes or preferences towards X may influence lower level motive generators not directly concerned with X, producing 'partially X-related' motives. (d) Meta-management control processes ensure that motives and thoughts pertaining to X are always decided as soon as possible, so that such motives tend to grab attentive resources immediately. (e) Dedicated evaluation procedures rate X-related motives preferentially, assigning skewed importance, urgency and cost-benefit measures. (f) Predictive models, triggered by X-related motives, will consume computational resources by attempting to reason about about X's needs and possible reactions to things. (g) In a resource-limited system, the proliferation of motives pertaining to X may 'crowd out' other motive generators.

Besides internal processes that spontaneously occur following the news of X's death, external reminders may trigger additional X-related processes: e.g. driving past X's favourite restaurant or accidentally finding an old photograph, or hearing X mentioned in a conversation. In some environments such reminders will be frequent. Perceptual schemata looking out for the 'lost' individual may misidentify strangers as X. The association of places, objects and events with memories of the deceased may be powerful triggers for perturbant episodes, sustaining the period of mourning and making recovery difficult without a change of location.

To summarise: *If a structure of attachment to X exists then motives and thoughts pertaining to X will surface and successfully compete for attentive computational resources; news relating to X's death will therefore have a strong tendency to generate perturbant states. The agent's thought processes will be partly out of control.*

2. The difficulty of accepting the fact of the friend's illness and death.

Updating many entries in a large database of information can take time, including the time for restructuring and propagation. This is one notion of the 'difficulty' of accepting new information.

Another factor is resistance to change. The agent has 'affective' grounds for wanting to believe that information about X's death is false. This could include long term high commitment intentions pertaining to X, involving mutual plans that have had resources expended on them, which the agent does not wish to regard as wasted. Besides uncomfortable evaluations, complete assimilation of the new information may require extensive resource-consuming cognitive reorganisation because the attachment structure is distributed and interwoven with other control states. Humans often seem to reject information, however reliable, if it requires extensive reorganisation of control states and value systems. This may be part of a good engineering design for intelligent agents in a mostly stable world.

Finally, the agent may know from past experience that the acceptance of such beliefs entails a long process of suffering and pain. Holding out hope that the information may turn out to be false is a management goal to delay the onset of this process.

3. The disruptive effect on normal, day-to-day functioning.

Perturbance involves disruption of the processes of motive management, and day-to-day goal processing may be adversely affected by management overload. It is difficult to plan a shopping trip or attend to what others are saying when distracted by futile regrets or painful thoughts and memories.

Besides cognitive disturbance, bereavement can cause physiological changes in the mourner, such as weight loss and excessive tiredness, which will contribute to a lack of efficacy; however, this is not explained by our architecture.

4. Periods of relative normality when grief is 'backgrounded', sometimes because external factors help one regain 'normal' control.

When the management system is involved in new important and urgent tasks it sets the interrupt filter threshold so high that the conditions discussed above no longer hold, like the soldier or football player who is injured and yet feels no pain. When the external demands are removed, the threshold drops, and processes relating to the bereavement regain control.

Another factor may be a general mood of depression that 'colours' motive processing during grief. A depressed mood is a global control state that 'scales-down' interaction with the environment. (We do not have space for a full discussion of moods.)

5. Attempts to 'fight' the grief.

How can the mourner 'fight' the grief and 'try' to get on with life? 'Fighting' here refers to a kind of mental striving or conscious self-control, which is not always easy. It requires some way of suppressing perturbant states, which is often much harder than control of emotional expression (external symptoms).

When the bereaved is attempting to work yet thoughts are continually drawn to the recent death, the self-control mechanism described above may detect the perturbance and attempt to negate its disruptive effects. Until detachment has been achieved, such self-control is partial and transient: the mourning returns when fragments of the attachment structure are next triggered, by external or internal (possibly subconscious) processes.

One form of self-control uses the artificial 'deadening' of cognitive activity by alcohol or anti-depressants. Chemicals can alter the functioning of abstract machines through their effect on the neuro-physiological substrate. Although understandable, this strategy could exacerbate the problem (drink can make people maudlin) or possibly slow down the process of detachment.

People are sometimes exhorted to try intentional suppression of perturbing thoughts, using internal imperatives: 'don't think about that', 'ignore that', 'put these thoughts out of mind' etc. This could also flow from a meta-management process, spawned by a self-control mechanism, which rejects new motives pertaining to the deceased. Some people may have learnt how to raise their attention filter threshold deliberately. In practice instructions to oneself often fail.

Even if temporarily successful this may lead to a build up of motives waiting to surface, since the cause of perturbance remains. Sudden surfacing of these suppressed motives could produce breakdown of control if there is a drop in the filter threshold due to low management load. Thus a person experiencing grief might function normally at work during the day (when high management load sets a high filter threshold) only to break down at home later (when both load and threshold drop).

Even at home one can try to absorb oneself in attention grabbing, computationally expensive tasks. 'And then there are details like writing obits, funeral arrangements, meeting his family, etc., which keep you busy. And then the grief resumes.' Arranging the funeral *has* to be done and can divert attention for a while, whereas attempting to read an interesting book fails.

However, as the intensity of grief lessens through gradual dismantling of the attachment structure, thoughts and motives relating to X have lower insistence, allowing 'normal' tasks with low importance and urgency to hold attention and prevent perturbance, and permitting enjoyment of activities such as listening to music, playing games, reading books, or conversing with others.

Another coping strategy is the formation of a new affectional bond to replace the old one ('on the rebound') — an option that is not always available, especially to an older person who has lost a spouse. The formation of such a bond might be a way to avoid the lengthy and painful process of detaching the structure of attachment by finding a new use for it. This involves replacing the original referent and possibly other things, and will not be achieved easily because of the distributed mechanisms and links that constitute attachment. This might be connected with the phenomenon of 'projection', where the grieving person views the new person in terms of the old.

Some of these strategies may be ineffective, or may have undesirable side effects, including hindering the long term process of detachment. The pressure to find quick fixes may come from the culture, for example through the necessity to keep one's job. This could cause harm if the only satisfactory strategy for achieving internal reorganisation following death of a loved one is through the natural process of grieving.

Examples of strategies that might aid this process are acceptance (as opposed to suppression), via a meta-level goal to interpret the experience of grief positively, by understanding that grief is necessary and worthwhile. A supportive social circle may be required for this to work. The self-control of emotional expression ('hold it inside, slog along and try to get on with life') becomes necessary when friends or work colleagues are less prepared to make allowances. It is difficult to control facial expression and general demeanour. People can usually see through the attempt. Limitations on our ability to dissimulate may arise from requirements for successful social co-operation (Sloman 92).

> 6. Second order motivators, some of them involving evaluation of the grieving state as good or bad, including, in some cases, wishing the grief to continue. See [o, q].

Disruptive and painful processes can trigger a second order motive to end the state. But the mourner quoted wishes to preserve his grief: 'I don't want to let go of the grief. Sometimes I think it's all I have left of XXX.' And 'I don't want to stop crying every day as I'm reminded of another piece of our time together. Well, I *do*, eventually, and I know I will, but I am not comfortable with this yet.' Why should the mourner — paradoxically — wish the grief to continue? There are a number of elements at work here: (a) the knowledge that the period of intense grieving may be coming to an end, (b) the association between grieving and the recollection of memories of the deceased, (c) a conflict of meta-level motivation between wanting to stop grieving and not wanting to.

The architecture described can support such conflicting processes. For example, detachment may be occurring concurrently with a meta-management process that ensures that the deceased is not forgotten, but remembered with the appropriate sadness. The meta-management process may have been constructed by a collection of high level control states constituting a self-image; for example, the mourner may view himself as somebody who loved the deceased very much and consequently *should* experience the appropriate amount of grief and heartache. (Cultural norms will affect this.)

Second order effects are to be expected within the framework of the architecture. The meta-management system includes self-monitoring processes that allow high level motive generators to be triggered by the detection of internal states that require some change in management strategy. Some of these simply redirect attention and cause sensible evaluating, reasoning, deciding, and planning to occur. Others generate some new motives that are, for one reason or another, hard to achieve, and some that are rejected yet go on being reactivated and interrupting processing. Exactly how all this develops will vary from individual to individual and within an individual from one situation to another. Second order processes may be strongly influenced by a culture.

> 7. The subjective 'pain' experienced by the mourner. There may be mental pain as well as bodily disturbances.

Why is grief a subjectively painful 'soul-grinding' experience? Some perturbant states are very pleasurable (such as excited anticipation); consequently, a structure of attachment perturbing attentive processing is not a *sufficient* condition for grief. The pain presumably comes via the fact that the death violates so many of the motives and preferences associated with the attachment structure. Further research is in progress to explain the roles of pain, pleasure and evaluation in the architecture.

> 8. Crying. Can we provide a design-based account of the onset of crying? Why does wailing or howling occur in times of extreme emotional distress?

Infants cry when their desires are unsatisfied. This grabs the attention of an adult who will then normally attend to the baby's needs. For a baby in the helpless altricial stage crying is a *basic, genetically determined, plan*. It is the only way to satisfy its needs for food, warmth, milk etc. Later, crying is no longer necessary because there is a repertoire of plans and the agent can act independently.

Consider the following scenario. A loved one suddenly died a few weeks ago and you are in mourning. You were thinking about going shopping but your attention was disrupted by thoughts of the lost one. A perturbant state is manifest. Your thoughts are out of control and you cannot stop thinking about what you were going to do together, the things you miss sharing with them, that you are lonely and without a close friend, and so on. You stare at the wall and ruminate. In terms of the architecture the structure of attachment is generating motives that are surfacing through the attention filter to disrupt management processes. These motives, pertaining to the dead person, are unsatisfiable and therefore rejected. But their high insistence makes them surface again, foiling any plans for regaining control. In this situation, the basic plan of crying may be invoked because (a) no other plans from the repertoire exist for the surfacing motives, (b) the invocation conditions of the basic plan are situations where 'nothing can be done', i.e. identical to the altricial stage of development, (c) the basic plan has worked in the past to satisfy desires in such situations, and (d) it has general applicability, i.e. was used to satisfy diverse and basic wants such as food, warmth, milk, and proximity to adults. *Crying is the plan of last resort,* and can be triggered by negatively valenced perturbant states. There may be other 'basic' plans, some concerned with internal processes only.

This explains Sartre view of 'emotions' as an attempt to 'magically' transform the world (Sartre, 48). From an infant's standpoint, crying achieves things by magic. As children grow their helplessness diminishes, along with the need to cry. (The picture becomes complicated later on, when the possibility of emotional deception is discovered — faking crying to manipulate others.)

Also, the notion of 'basic plans' as control mechanisms of last resort bears important similarities to Kraemer's *cascade hypothesis* (Kraemer, 92), which states that control will 'cascade' down to 'genetically programmed neurobiological adaptive behaviours' if the organism is faced with 'disasters' (problematic situations) that its 'acquired behaviours' cannot deal with. However, Kraemer's approach is very different from ours, namely psychobiological as opposed to cognitive, and leads to seemingly different answers (biogenic amine system function as opposed to information processing). There is no space for a full comparison here, but a synthesis should be possible. For example, our notion of a basic plan makes no commitment to its implementation details. The integration of 'bottom-up' approaches with 'top-down' requirements-driven design work should yield fruitful, and not necessarily contradictory, results.

## 3.5 Discussion

Using the postulated architecture, we have offered a partial and provisional design-based interpretation of many aspects of human grief. However, our theory is sketchy, incomplete and offered only as an initial step towards a comprehensive theory of emotions informed by the exploration of agent architectures in which control mechanisms are information based.

**1** For each loved person, a structure of attachment develops, consisting of diverse distributed mechanisms and representations with varying powers of persistence and dispositional causal roles in determining behaviour. This is the affectional bond.

**2** Removal of the referent renders the structure of attachment inappropriate for the control of behaviour.[6] Nevertheless, triggered by the news of the death and reverberating associations, the attachment structure generates thoughts and motives that surface and divert attentive resources, producing negatively valenced perturbant states.

[6] Compare Oatley's treatment of grief: 'a whole repertoire of subplans and knowledge becomes useless' (Oatley, 92).

**3** Perturbant states disrupt normal functioning in resource-limited management processes.

**4** The futility of other plans may trigger regression to a basic plan of crying.

**5** Various self-control strategies may be instigated to overcome the perturbant states; however, the phenomenology of grief suggests that the causal powers of self-control mechanisms are limited.

**6** Detachment takes time due to the deep and diffuse embedding of the attachment structure in the architecture. Extensive cognitive reorganisation and re-learning is required before the generation of perturbant states ceases, or drops to a manageable level. How long it takes will depend on details of the case. Some grief lasts as long as the griever.

Much work is still required, to deal with many difficulties and gaps in the design and the interpretation we have based on it. Problematic areas include mechanisms for: learning, pleasure and pain, self monitoring, the hierarchy of dispositional control states, self-controlling abilities, the kinds of global control indicated by mood changes, and the processes which assemble and disassemble attachment structures,

Terms such as 'attention' and 'conscious' have been used without precise definitions. When fully specified the architecture will be used as the basis for a host of new definitions of classes of mental states and processes (like basing the descriptions of types of physical stuff on a theory of the architecture of matter).

Questions arise about the possibility of using the architecture to explain other perturbant states such as distress at the ending of a relationship, excited anticipation, sexual infatuation, obsessive love of a child, various pathologies of motivation or attention such as obsessive-compulsive disorder and attention deficit disorder. These are all possible directions for future work.

It is hard to think about the multifarious states and processes that can occur in such a complex paper design. A working implementation can aid analytical thinking, by exposing consequences of the design. An implementable scenario for investigating perturbance is outlined in (Wright, 94), and further work is in progress. This has led us to extend the architecture, to overcome limitations of early versions which assumed a fixed succession of processes to be applied to new motivators. This proved unworkable (Beaudoin 94).

It may be objected that the use of an architecture as explanatory ground merely restates the obvious in a more complicated way, particularly if it generates no new testable predictions. This expresses a naive view of complex information processing systems. Prediction may be impossible because some determinants of behaviour are inaccessible internal information states.

Our work can be classified as belonging to the 'creative modelling' phase of science (Bhaskar, 78 & 94), where competing models of generative mechanisms of surface phenomena are explored. Until we have a good 'design-space' of broad and increasingly deep agent architectures it will be difficult to move from hypothetical explanation to justified selection between hypotheses. The research community operating as parallel search in both empirical and theoretical directions will ultimately select between competing theories, using a variety of criteria including generality, simplicity, implementability, evolvability and applicability to other animals. But the theories need to be developed. Using an architecture as explanatory ground is a first step.

The interpretation of grief given here may be compared with other design-based models of emotionality, which tend to reason about emotional labels based on an operational semantics of emotion concepts (for example, see (Dyer, 87), (Frijda & Swagerman, 87) and (Pfeifer, 92)). On our model such emotion concepts would arise from internal perception of motive processing states, self-control mechanisms and social interaction. Another point in favour of the architecturally grounded interpretation is that it has greater explanatory power than that of folk psychology. For example, we begin to give answers to the following questions: *What is grief?* – Grief is (often) an extended process of cognitive reorganisation characterised by the occurrence of negatively valenced perturbant states caused by an attachment structure reacting to news of the death. *What causes grief to endure?* – Grief persists because of the time required to disassemble a complex, distributed and deep attachment structure. *Why does grief consume the mourner?* – Attentive processing is resource-bound and becomes swamped by highly insistent motives generated by a structure of attachment to a highly valued individual; in addition, the requirements for re-learning and detachment entail extensive rumination that can also generate perturbant states. *Why do we cry during grief?* – In processing situations where strong but unsatisfiable motives surface, a time may come when otherwise helpless management processes resort to a basic plan. Crying is the plan of last resort. *Why can't grief be overcome easily?* – Processes of self-control are limited by architectural opacity, the lack of adequate categorisations of internal states, the lack of good meta-management strategies, and the disruptive effects of perturbances.

# 4 Concluding comments

We have sketched an architecture of a sort that might be useful for an intelligent agent with multiple asynchronously generated goals in a complex and changing environment. We have shown how perturbant states can emerge in such an architecture and how a structure of attachment can generate ongoing perturbance after bereavement. This extends some previous theories of emotions and draws attention to new research problems, some of which require more detailed design specifications and some more detailed empirical investigations of phenomena involving grief and other perturbant states.

Our architecturally grounded interpretation of human grief has been provided within a design-based framework. We expect development of the design to lead to more fine-grained analysis of well-known phenomena, generating hard cases that will force further development of the theory. Detailed implementations may reveal unexpected insights, along with hidden flaws in the theory and also with opportunities to extend it.

No strong claim is made that the interpretation is *the* theory; currently, we are only exploring theoretical possibilities and illustrating a new approach to the study of complex mental phenomena.

We have extended our earlier (pre 1995) design, particularly in the direction of self-monitoring and self-controlling abilities, and alluded to the need for further extension, e.g. learning mechanisms. Sartre's intuition that 'emotions' are magical attempts to transform the world has been architecturally grounded via the idea of a fall-back plan in situations of helplessness. We have not yet said anything about links with neural structures and processes or emotional mechanisms shared with other animals.

We must stress that whether the actual interpretation of grief outlined here is correct in detail is of secondary importance. A stronger claim can be made for the methodology that generated the interpretation. In order fully to understand the complexity of human behaviour we have to posit new ontologies at the information processing level of description just as a designer specifies an architecture to meet some collection of behavioural requirements. By exploring families of related architectures and their strengths and weaknesses in relation to various niches, we may be able eventually to explain not only current human capabilities but also the evolution of human and non-human mental architectures in a variety of organisms. The work will also give us new insights into what to expect if and when we design autonomous artificial agents, whose internal complexity will make detailed prediction of their behaviour very difficult.

That all this will be a long and difficult process, with continual revision of ideas, is not in question.

## Acknowledgements

This work was supported by the UK Joint Council Initiative and the Renaissance Trust.

Thanks to Christian Paterson, Chris Complin and past and present members of the Cognition and Affect project at Birmingham. Additional papers by the group can be found at the ftp site:

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

and in Aaron Sloman's Web page

**http://www.cs.bham.ac.uk/~axs/cogaff.html**

## References

Bates, J., Loyall, A. B., & Reilly, W. S. (1991). Broad agents. Paper presented at the *AAAI spring symposium on integrated intelligent architectures.* Stanford, CA: (Available in SIGART BULLETIN, 2(4), Aug. 1991, 38-40).

Bawden, A. (1988). Reification without evaluation. MIT AI Memo 946; also in *Proceedings of the 1988 ACM Conference on Lisp and Functional Programming.*

Beaudoin, L. P. (1994). *Goal Processing in Autonomous Agents.* PhD Thesis, School of Computer Science, University of Birmingham.

Beaudoin, L. P. & Sloman, A. (1993). A study of motive processing and attention. In *Proceedings of AISB93*, A. Sloman, D. Hogg, G. Humphreys, A. Ramsay & D. Partridge (Eds), 229-238, Oxford: IOS Press.

Bhaskar, R. (1978). *A Realist Theory of Science.* The Harvester Press Ltd.

Bhaskar, R. (1994). *Plato, Etc.* Verso.

Brooks, R. A. & Stein, L. A. (1993). Building brains for bodies. MIT artificial intelligence laboratory, AI Memo No. 1439.

Bowlby, J. (1979). *The Making and Breaking of Affectional Bonds.* Tavistock Publications Ltd.

Bowlby, J. (1988). *A Secure Base.* Routledge.

Chalmers, D. (1996). *The Conscious Mind.* Oxford University Press.

Dodd, B. (1991). Bereavement, in: *Psychology and Social Issues* R. Cochrane & D. Carroll (eds), London: The Falmer Press, 1991 pages 63-72.

Dyer, M. G. (1987). Emotions and their computations: Three computer models. *Cognition and Emotion* 1(3), 323-347.

Fehling, M. R., Altman, A. M. & Michael Wilber, B. (1989). The Heuristic control virtual machine: An implementation of the schemer computational model of reflective, real-time problem-solving. In *Blackboard Architectures and Applications.* Academic Press, Inc.

Firby, R. J. (1987). An investigation into reactive planning in complex domains. *Proceedings of the Sixth National Conference on Artificial Intelligence*, (202-206). Seattle: AAAI.

Frijda, N. H. (1986). *The Emotions.* Cambridge: Cambridge University Press.

Frijda, N. H., & Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. *Cognition and Emotion*, 1, 235-257.

Georgeff, M. P., & Ingrand, F. F. (1989). Decision-making in an embedded reasoning system. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 2 (972-978). Detroit, MI: IJCAI.

Hanks, S., & Firby, R. J. (1990). Issues and architectures for planning and execution. In *Proceedings of a Workshop on Innovative Approaches to Planning, Scheduling and Control*, San Diego, CA: DARPA.

Hayes-Roth, B. (1993a). An architecture for adaptive intelligent systems (KSL Report No. 93-19). Knowledge Systems Laboratory, Department of Computer Science, Stanford University.

Hayes-Roth, B. (1993b). Intelligent control. *Artificial Intelligence*, 59, 213-220.

Hayes-Roth, B. (1991a). Evaluation of integrated agent architectures. SIGART Bulletin, 2(4), 82-84.

Hayes-Roth, B. (1991b). An integrated architecture for intelligent agents. SIGART Bulletin, 2(4), 79-81.

Johnson-Laird, P. N. (1988). *The Computer and the Mind: An Introduction to Cognitive Science*, Fontana.

Loyall, A. B., & Bates, J. (1991). Hap — A reactive, adaptive architecture for agents (Technical report No. CMU-CS-91-147), School of Computer Science, Carnegie Mellon University.

Kagan, J. (1978). On emotion and its development: a working paper. In *The Development of Affect* edited by M. Lewis and L. A. Rosenblum. Plenum Press, New York and London.

Kraemer, G. W. (1992). A psychobiological theory of attachment. *Behavioural and Brain Sciences* 15, 493 – 541.

Kuhl, J., & Kraska, K. (1989). Self-regulation and metamotivation: Computational mechanisms, development, and assessment. In R. Kanfer, P. L. Ackerman, & R. Cudek (Eds.), *Abilities, motivation, and methodology: The Minnesota Symposium on Individual Differences* (343-374). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Lakatos, I. (1970), Falsification and the methodology of scientific research programmes, in *Criticism and the Growth of Knowledge* I. Lakatos and A. Musgrave (eds), Cambridge University Press, 921-196.

Miller, Galanter, & Pribram (1970). *Plans and the Structure of Behaviour.* Holt International Edition.

Moffat, D. & Frijda, N. H. (1995). Where there's a Will there's an agent. To appear in: *Intelligent Agents – Proceedings of the 1994 Workshop on Agent Theories, Architectures and Languages*, M. J. Wooldridge and N. R. Jennings (Eds.), Springer Verlag (LNAI Series) 1995.

Oatley, K. (1992). *Best laid schemes, the psychology of emotions.* Studies in Emotion and Social Interaction, Cambridge University Press.

Palmer, S. E., & Kimchi, R. (1984). The information processing approach to cognition. In *Approaches to Cognition: Contrasts and Controversies*, T. J. Knapp & L. C. Robertson (eds.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Paterson, C.J. (1995) The use of ratings for the integration of planning and learning in a broad but shallow agent architecture. MPhil Thesis, Computer Science Department, University of Birmingham.

Penrose, R. (1989) *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics.* Oxford University Press.

Pfeifer, R. (1992). The new age of the fungus eater: Comments on artificial intelligence and emotion. Extended and revised version of an invited talk at the AISB-91 conference in Leeds, U.K.

Pryor, L., & Collins, G. (1992). Reference features as guides to reasoning about opportunities. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Bloomington, Lawrence Erlbaum Associates.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-Architecture (Technical Note No. 14). Australian Artificial Intelligence Institute, 1 Grattan Street, Carlton, Victoria 3053, Australia.

Rao, A. S., & Georgeff, M. P. (1992). An abstract architecture for rational agents. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning.* Boston: KR92.

Read, T., & Sloman, A. (1993). The terminological pitfalls of studying emotion. Paper presented at the *Workshop on Architectures Underlying Motivation and Emotion - WAUME 93*, Birmingham.

Reilly, W. S. (1993). Emotions as Part of a Broad Agent Architecture. Talk given at WAUME93, Birmingham.

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

Rozas, G. J. (1993). Translucent procedures, abstraction without opacity. Phd thesis: MIT AI Technical Report No. 1427.

Sartre, J. P. (1948) *Esquisse d'une théorie phénomenologique des émotions (The Emotions)*, Paris: Hermann (New York: Philosophical Library), 1934, Translated 1948

Searle, J.R. (1980), Minds Brains and Programs, *The Behavioral and Brain Sciences* 3,3.

Simon, H. A. (1967). 'Motivational and Emotional Controls of Cognition', reprinted in *Models of Thought*, Yale University Press, (1979) 29–38.

Simon, H. A. (1969). *The Sciences of the Artificial*, Second Edition, MIT Press (1981).

Simon, H. A. (1995). 'Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration', *Proceedings 14th International Joint Conference on Artificial Intelligence* Montreal, August 1995, 939–945.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind.* Harvester Studies in Cognitive Science, Harvester Press.

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

Sloman, A. (1992) Prolegomena to a theory of communication and affect, *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues,* A. Ortony, J. Slack, and O. Stock, eds. Springer: Heidelberg, 229-260.

Sloman, A. (1993a). Prospects for AI as the general science of intelligence. In *Prospects for Artificial Intelligence* Proceedings of AISB93, Oxford IOS Press, Editors: A. Sloman, D. Hogg, G. Humphreys & D. Partridge.

Sloman, A. (1993b). The mind as a control system, in *Philosophy and the Cognitive Sciences,* (eds) C. Hookway and D. Peterson, Cambridge University Press, 69-110.

Sloman, A. (1994) Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering,* Vol 349, 1689, 43-58

Sloman, A, (1995) Exploring design space and niche space. In *Proceedings 5th Scandinavian Conf. on AI,* Trondheim May 1995, IOS Press, Amsterdam,

Sloman, A., Beaudoin, L.P., & Wright, I. P. (1994) Computational modeling of motive-management processes, *Proceedings of the Conference of the International Society for Research in Emotions, Cambridge, July 1994.* (ed) N.Frijda, ISRE Publications. 344-348.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, (197-202). Vancouver.

Sloman, A. & Poli, R. (1995). 'SIM_AGENT: A toolkit for exploring agent designs' in *ATAL-95, Workshop on Agent Theories, Architectures, and Languages,* IJCAI-95, Montreal, August 1995. (Also Cognitive Science technical report: CSRP-95-3, The University of Birmingham.)

Smith, B. C. (1986). Varieties of self-reference. In J. Y. Halpern (Ed.), *Proceedings of the First Conference on Theoretical Aspects of Reasoning About Knowledge.* Morgan Kaufman.

Smith, P. K. & Cowie, H. (1991). *Understanding Children's Development*, chapter 3: Parents and families. Blackwell Publishers.

Wooldridge, M. & Jennings, N. R. (1995). Agents: Theories, Architectures and Languages, *Knowledge Engineering Review.*

Wright, I. P. (1994). An emotional agent: the detection and control of emergent states in autonomous resource-bounded agents. Cognitive Science Research Report RP-94-21, School of Computer Science and Cognitive Science Research Centre, University of Birmingham.