

# Control States and Motivated Agency

Steve Allen<sup>1</sup>

Deduction and Multiagent Systems Group  
German Research Centre for Artificial Intelligence (DFKI)  
Stuhlsatzenhausweg 3, D-66123 Saarbrücken  
Germany

<http://www.dfki.de/~allen>  
Steve.Allen@dfki.de

## Abstract

One of the challenges faced by researchers in the behaviour modelling of life-like characters is the need to develop a systematic framework in which to ask questions about the types of internal state life-like characters might possess, and how those different states interact. We propose a solution based on a cognitively inspired multi-layered agent architecture (composed of reactive, deliberative and meta-management layers), and a recursive “design-based” research methodology – wherein each new design gradually increases our explanatory power and allows us to account for more and more of the phenomena of interest. By describing a variety of “broad but shallow” complete agents at the information-level, and showing how these designs realise mental states and processes, we aim to provide a rich and deep explanatory framework from which to explore motivated autonomous agency. Early experiments have concentrated on: (a) the requirements of goal-processing; (b) the emergence of perturbant (emotional) states; and (c) the relationship between motives, goals, emotions, and personality.

## 1 Introduction

### 1.1 What we are trying to do

Our conjecture is that human “higher-level” mental concepts (beliefs, desires, intentions, moods, emotions, personality, etc.) are grounded in implicit assumptions about an underlying information processing architecture. The aim of this paper is to provide a framework in which to explore this conjecture, and make the underlying architectural assumptions more explicit. For example, *rage* (anger characterised by a breakdown in self-control) and *grief* (characterised by an inability to concentrate due to recurrent thoughts of a lost loved one) place a requirement on the architecture for self-monitoring and deliberation – you cannot lose what you do not have. Whereas *startle* places a requirement on the architecture for reactive mechanisms that respond quickly to potentially serious unexpected events.

---

<sup>1</sup> This research is also part of the Cognition and Affect Project at Birmingham University, UK.

## 1.2 How we intend to do it

Although we are interested in building autonomous agents in order to elucidate “higher-level” mental concepts, we will not restrict ourselves to intentional models [Dennett 87] that rely on knowledge level descriptions. These approaches pre-suppose rationality – *If an agent has knowledge that one of its actions will lead to one of its goals, then it will select that action* [Newell 82, page 102] –, whereas many of the mental states we are interested in include emergent and/or automatic (i.e. neither rational or irrational) processes. We therefore propose using information-level descriptions within a cognitively inspired multi-layered agent framework.

The partitioning of multi-layered architectures is often made on functional grounds, with little regard to the emergent properties of the architecture (an exception is the *GLAIR* [Hexmoor et al 93] architecture which is partitioned to investigate the learning of emergent behaviours). But it is often these very properties that we want to capture within our motivated / life-like characters – i.e. *moods, emotions, and personality*. If we are to understand the emergence of such properties, then we must systematically study the attributes of these emergent states and their relation to the different layers of the architecture. This needs to be done by specifying, analysing, and building complete agents that meet a human-like (our idealised subject) requirements specification, i.e. have the ability for both rapid and considered response, with incomplete / inconsistent knowledge, and multiple competing concerns, in a dynamic and possibly hostile environment.

## 1.3 Structure of this paper

This paper is organised as follows: section 2 focuses on the role played by motivational Control States within an agent architecture, trying to answer the basic “what” and “how” questions about control states; section 3 introduces the Motivated Agent Framework, describing our design-based research methodology and the cognitively inspired multi-layered architecture; section 4 discusses a range of motivational control states (motivators, emotions, and personality) in terms of our framework; section 5 discusses the theoretical aspects of these motivational control states within a concrete Society Of Mind [Minsky 87] architecture; and finally section 6 presents a summary of our approach.

## 2 Control States

Thinking of complex systems as collections of interacting, partially-independent, sub-systems (or *control states*) provides two useful functions. Firstly, it allows us to ask questions about what types of *control state* complex systems might possess, and secondly, it allows us to ask how those different *control states* might interact. A thermostat can be represented by three *control states*: belief-like; desire-like; and intention-like – the thermostat can hold a belief that “the room is at 20°C”, a desire to “make the temperature 23°C”, and an intention to “turn the radiator on”. But, *control states* are not restricted to the classic BDI (Belief, Desire, Intention) formalism. *Control states* can operate asynchronously, at different rates, and at different / multiple levels within the architecture.

## 2.1 What Are Control States?

Complex control systems (such as the minds of humans or life-like characters) are capable of supporting many different types of control states over and above the beliefs and desires of a simple thermostat – people regularly base their predictions upon observations such as “he is in a bad mood” without referring to the desires or beliefs of the observed agent. The fact that we can use “mood” as a predictor of behaviour is a good indicator that it is a control state – however a more scientific form of qualification is called for: *control states* are information-bearing representations of an information processing control system.

Formally, we define two types of attribute for a control state [Beaudoin 94, section 3.1.1]: (a) *Dimensional attributes* refer to the quantitative attributes such as duration, and intensity; and (b) *Structural attributes* which describe the “virtual machines” through which control states are realised (these structural attributes are often linked to the agent’s ontology). For example, *plans* may have dimensional attributes of importance, status (active, suspended, partial expansion, etc.), and relevance; with structural attributes such as valid predicates, relations, and propositions.

An important subset of structural attributes are those that reflect mechanisms which modify other representations, i.e. mechanisms which transform *beliefs* into either *standards* or *attitudes*, or modify existing *beliefs* on the basis of new *goals*. Finally, we allow values of dimensional attributes to be expressed in term of the structural attributes – i.e. the duration of an emotion can be expressed in terms of the emergence of a perturbant state in which a motivator repeatedly grabs and holds attention.

The flexibility of this classification scheme allows us to explore the requirements of *control states* without committing ourselves to rigid representational forms. Figure 1 shows the variation in scope and duration we must cope with for some common *control states* of human-like motivated agency.

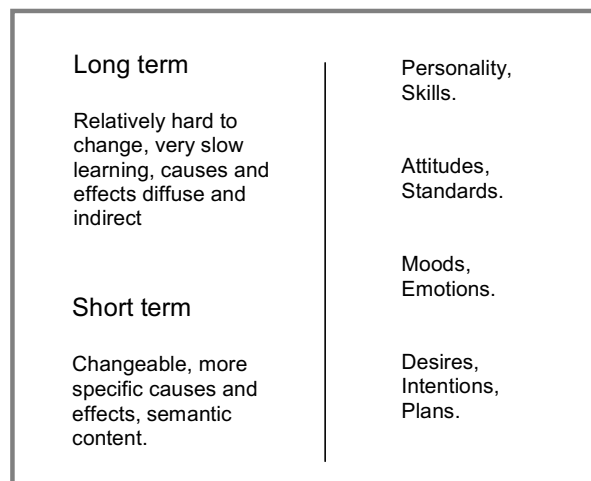


Figure 1 - Scope and Duration of Control States

## 2.2 How are Control States Realised?

Our view of a “control system” is clearly at odds with the standard notion of a control system used by physicists and control engineers. Conventional control systems have a fixed degree of complexity, allowing their behaviour to be completely described by a system of partial

differential equations. However, the intelligent control systems that we wish to describe do not have a fixed architecture, and are capable of evolving during the lifetime of the agent. Further, within such intelligent systems, many of the control states exhibit changes that are more like changing structures than like changing values of numeric variables – *beliefs* become rigid *attitudes*, and *learnt behaviours* become homed *skills*.

Control states are realised as “virtual machines” which operate on information-level representations within the agent architecture. A belief-like control state implies the existence of mechanisms for belief generation, representation, storage, evaluation and execution. In the case of a mechanical thermostat; the differential expansion of two metal strips provides the belief generation, the curvature of the bimetallic strip provides the belief representation and storage, and the making or breaking of an electrical contact provides the belief evaluation and execution.

However, not all control states require specific supporting mechanisms within the architecture. Some control states emerge from the interaction of lower level mechanisms, whereas others may share common components, only differing in the way they interact (i.e. *beliefs*, *standards*, and *attitudes*).

### 2.3 What Control States are Needed for Motivated Agency?

There are no hard and fast rules for determining the number and nature of control states needed for motivated agency. Many successful agents have been built within the classic BDI (Belief, Desire, Intention) framework, and in principle all agents can be reduced to purely reactive architectures. However we believe that there are definite benefits to be gained in working with appropriate levels of abstraction, for example it may be useful to distinguish *reflexes* from *skills* or *behaviours* within the architecture. Figure 2 gives a feel for the typical range of control states we would like to investigate within the context of life-like characters.

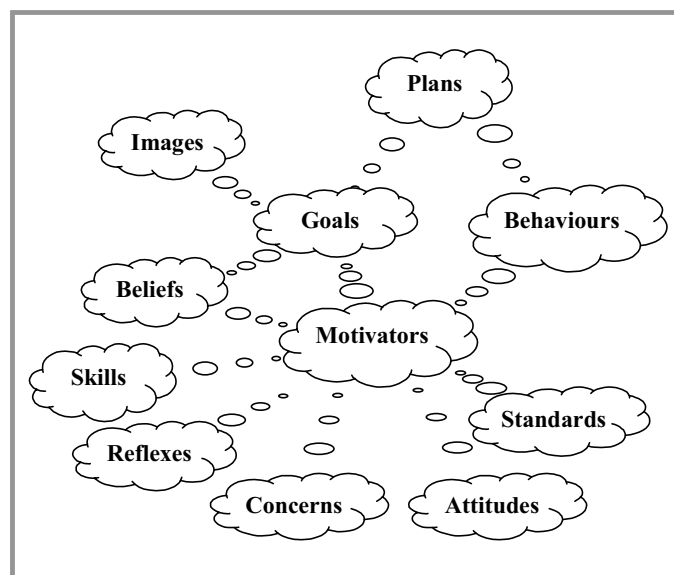


Figure 2 - Possible Control States of a Motivated Agent

### 3 Motivated Agent Framework

Our initial descriptions of motivational control states will inevitably be rather vague and somewhat unsatisfactory. We therefore propose building and analysing many different “broad but shallow” agents to give us a deeper understanding of their attributes, and their functional and emergent properties. These agents will all be defined within a common framework, and based on a requirements analysis for human-like control systems attempting to support multiple motives in a rapidly changing hostile environment.

#### 3.1 Design-Based Research Methodology

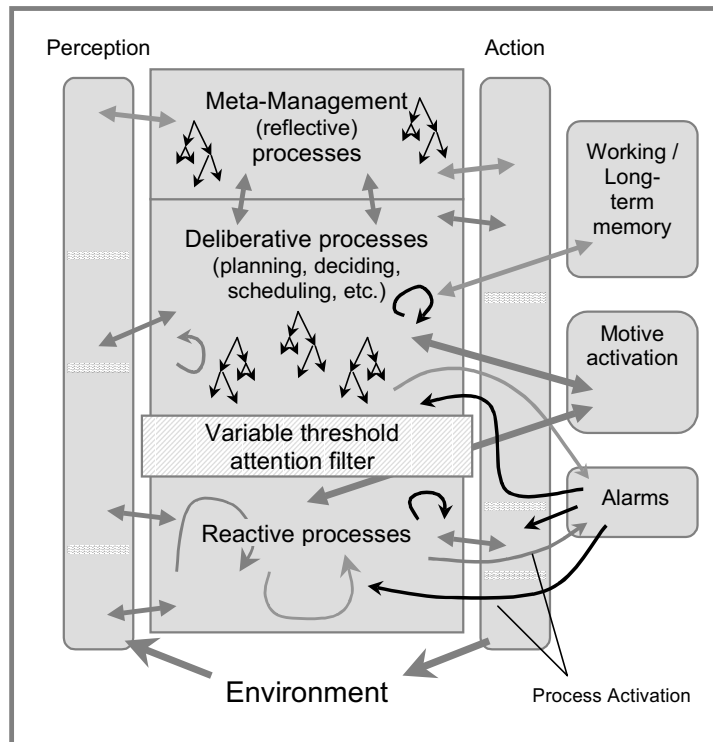
Producing and understanding complex systems requires an iterative, exploratory, design process in which both understanding of the problem and proposed solutions systematically improve. This involves five parallel threads of execution, using a mixture of top-down and bottom-up analysis and design. Threads 1-3 represent common engineering practices, and threads 4-5 add scientific rigour.

1. *A requirements analysis of the system of interest*, including its main capabilities, its environment, hard and soft constraints of various kinds, and so on.
2. *A design specification for a working system*, describing major components and their functional relationships. Designing each component can recursively replicate all five threads described here.
3. *A succession of working implementations*, whose scope and amount of detail will typically increase as the project develops.
4. *A theoretical and (where mathematical techniques are inadequate) empirical analysis of how far the design meets the currently articulated requirements*. This will help to identify trade-offs and suggest improvements to requirements and designs for future iterations.
5. *An analysis of similar designs in “design-space”*. By considering alternatives to a particular design, we can often obtain a deeper understanding of that design, e.g. the trade-offs it involves.

#### 3.2 Motivated Agent Architecture

We will use the term “motivator” to refer to motivational control states that move an agent towards a desired (or away from an undesired) physical / mental state in the light of agent *beliefs* and *concerns*. In other words, *motivators* are a subclass of information structures with dispositional powers to determine action (both internal and external), and which subsume *emotions*, *desires*, *goals*, and *intentions*.

Our motivated agent architecture is based around four basic mechanisms (co-existing sub-systems), we believe a *human-like* motivational agent is likely to possess [Sloman and Logan 98]. These mechanisms are described below, and shown pictorially in Figure 3.



**Figure 3 - Motivated Agent Architecture**

The four co-existing sub-systems are:

- a) *Pre-Attentive (Reactive) management processes* that use dedicated circuits to respond automatically to triggering conditions in the environment. There is no considered construction of new plans or explicit evaluation of alternative options. New behaviours and concepts may form through modification of association strengths or relative weights in automated processes such as reinforcement learning. It is likely that reactive processes form hierarchical control structures, especially in the perceptual and action sub-systems. At low levels of the hierarchy, reactive circuits may be continuous and analogue in nature (using simple feedback and feed-forward connections to achieve high levels of information processing speed). As you move up the hierarchy, the circuits take on a more digital nature in the form of discrete behaviours or more abstract concepts. Some genetically determined circuits may act as alarm signals, triggering emergency response patterns or behaviours. Conflicts over shared resources (action selection) may be handled by relatively simple mechanisms such as spreading activation, vector addition or winner-takes-all networks. The agent can survive even if it has only genetically determined behaviours, provided the environment does not present many problems for which the generically determined solutions fail.
- b) *Attentive management processes* that use general purpose resources to focus and address the current primary concerns of the agent. As reusable mechanisms and space are dynamically allocated, many of the processes are inherently serial and resource limited. Access to concurrent long-term memory may also be inherently serial due to problems of cross-talk. *Deliberation* (a sub-state of attentive processing) is the process whereby possible world models are constructed and used for the evaluation of plans and goals before actions are selected. Deliberation

requires working memory to facilitate the comparison of options, and long-term memory to store the individual steps used in the construction of the plan. Perception may require deliberation to resolve ambiguities and constrain the search path of possible candidate concepts. Action may require deliberation to carry out novel or complex tasks for which behaviours have yet to be established. Attentive / Deliberative processes can be thought of as threads of a “virtual machine” running on the reactive substrate.

- c) *The Attention Filter* is proposed as a mechanism to protect the resource limited attentive processes from excessive interruption by reactive motivators. The filter threshold is set by meta-management processes, and reflects the perceived importance/urgency/difficulty of the current attentive task. Only motivators (events with motivational attributes) with *insistence* levels higher than the threshold can pass through the filter and grab attention. Insistence assignment (propensity to penetrate the filter) is based on heuristic measures of motivator importance and urgency.
- d) *Meta-management processes* which are responsible for agent adaptation by monitoring and controlling reactive / attentive management mechanisms. It is likely that approaches that work well in an agents early development may become less-than-optimum as the agents environment (including its internal environment) change. Meta-management processes enhance the agents adaptability and robustness by continually evaluating the agents current performance against long-term generic objectives. These long-term objectives (motivational *attitudes*) could include such things as not failing in too many tasks, not allowing the achievement of one goal to interfere with other goals, not wasting a lot of time on problems that turn out non-solvable. Meta-management is achieved through a process of inner perception and action acting on the attentive state of the agent.

#### 4 Motivational Control States

So far we have used rather vague terminology when describing the attributes of control states. The following discussion attempts to situate these attributes within the motivated agent framework of Section 3, building on the general definitions given below.

- *Motivators* are mechanisms and representations that tend to produce, or modify, or select between action in the light of *concerns* and *beliefs*.
- *Concerns* are dispositions to desire occurrence or non-occurrence of a given kind of situation [Frijda 86, page 335].
- *Goals* are states of affairs towards which the agent is motivationally directed.
- *Standards* are beliefs about what ought to be the case as opposed to what one simply wants – or would like – to be the case.
- *Attitudes* are dispositions, or predispositions, to like some things and to dislike others without reference to standards or goals [Ortony, Clore & Collins 88]. Attitudes may change over time, but tend to change slowly.

## 4.1 Motivators

In moving the agent towards a desired physical/mental state, motivators need to perform three functions [Cañamero 97]: (1) a *directing* function – they steer behaviour towards alleviating a particular concern; (2) an *activating* function – they animate the agent into action, and (3) an *organising* function – they combine individual behaviours into a coherent, goal-orientated response.

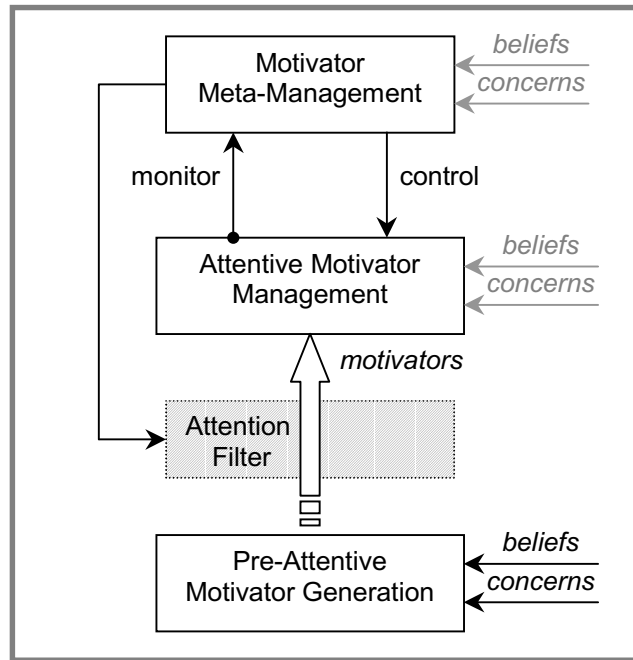
### *Directing Function of Motivators*

By matching events against agent *concerns*, motivators are able to *direct* behaviour towards alleviating those concerns (see Figure 4; also Norman [96]). In this sense, we can also think of *concerns* as “*motivational attitudes*, serving as *standards* against which situations are tested for compliance or non-compliance with desired norms”.

### *Activating Function of Motivators*

Within our motivated agent framework, motivators have an *activating* effect on two levels: (1) at the attentive / deliberative management level motivators generate new sub-goals in response to deliberative reasoning or as part of pre-constructed *plans*; and (2) at the reactive level, motivators can (a) initiate automatic behaviour such as *reflexes* in response to epistemic events, (b) prepare our agent for action through *action readiness change* (discussed in relation to affective states in Section 5), or (c) generate new management *goals* which interrupt attentive management processes (discussed below). As our agent should be able to survive with just its reactive competencies, we make an implicit assumption that all high-level concerns are encoded within the reactive layer (where high-level means those concerns with a high survival value).

Real-time performance is achieved by reactive motivator generactivators (generators/re-activators) which interrupt ongoing management processes when new events match agent *concerns*. As motivators are generated by reactive processes largely ignorant of current state of management processing, management attention must be protected from excessive and/or irrelevant diversion by an attention filter (under the control of meta-management processes). The reactively generated motivators are assigned an *insistence* level (the propensity for evaluations to pass through the filter) proportional to the perceived urgency/importance of the motivator. Motivators with insistence levels greater than the current filter threshold, penetrate the filter and capture management resources. This is the *activating* function of motivators, depicted graphically in Figure 4.

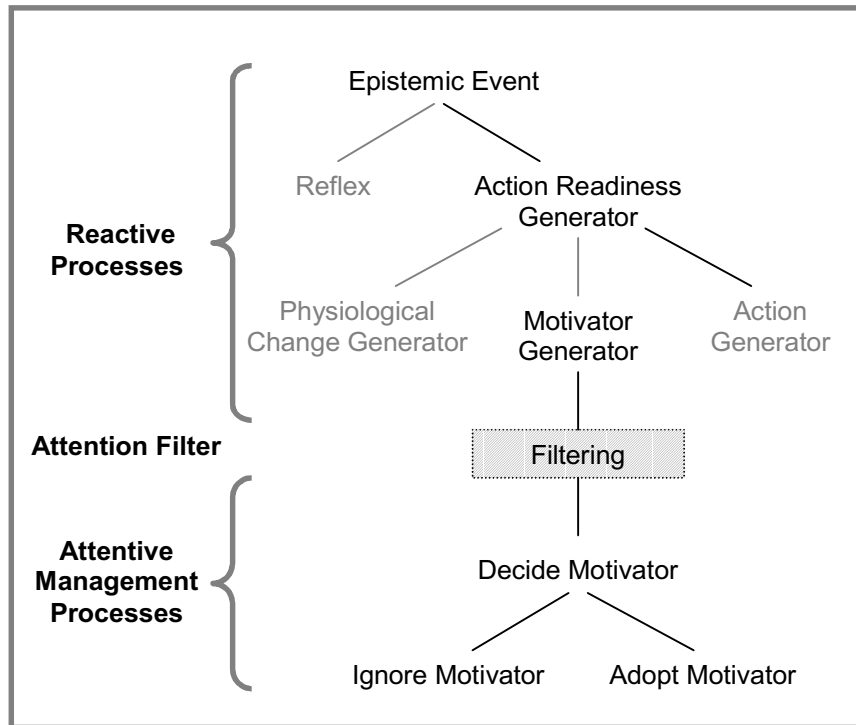


**Figure 4 - The Role of The Attention Filter**

The ability of reactive motivators to interrupt attentive processes depends on the interaction of two separate mechanisms: (1) reactive insistence assignment - representing the urgency / importance of the new reactive motivator, and (2) filter threshold assignment – representing, amongst other things, the urgency/importance of the adopted motivator. Meta-management processes can raise or lower the filter threshold in response to the needs of management processing. Factors that affect the filter threshold are: (a) the number of motivators already under management consideration; (b) the assessed urgency/importance of the current management process; and (c) the ability of management processes to cope with the current situation.

### ***Organising Function of Motivators***

Motivators *organise* behaviour when adopted by management processes. Motivators with insistence levels higher than the filter threshold pass through the filter, but simply passing through the filter is not a guarantee of motivator adoption by management processes (see Figure 5). Motivators that *surface* through the filter are first *decided* (assessed to see whether the motivator should remain surfaced, i.e. whether management resources should continue to be devoted to it) before being *adopted*. With a deliberative layer, it is therefore possible for a motivator to temporarily distract attention without changing ongoing plans and actions.



**Figure 5 - Motivator Adoption**

## 4.2 Emotions

Emotions form a powerful, but ill-defined class of motivational control states. We can make some inroads into resolving this situation with the general observation that: those theorists who: (a) stress emotions based on the limbic system are primarily studying effects of the reactive layer; (b) stress emotions such as apprehension, disappointment and relief, related to phases in the execution of plans, are studying effects of the attentive/deliberative layer; and (c) stress emotions involving loss of control of thought processes are studying processes involving the reflective, or meta-management layer. We can therefore identify three main classes of emotional state:

- 1) *Primary* emotional states: such as being startled, terrified, or sexually stimulated, which are typically triggered by patterns in the early sensory input and detected by a global alarm system. These emotional states are sometimes called primes or primary emotions [Buck 85; Damasio 96; Picard 97].
- 2) *Secondary* emotional states: such as being anxious, apprehensive, or relieved, depend on the existence of a deliberative layer in which plans (for future states) can be created and executed with relevant risks noticed, progress assessed, and success detected, etc. An alarm system capable of detecting features in these cognitively generated patterns is still able to produce global reactions to significant events in the thought process that impinge on the concerns of the agent (person). Damasio terms such cognitively generated emotional states – secondary emotions.

- 3) *Tertiary* emotional states: such as feeling humiliated, ashamed, or guilty, can be further characterised by a difficulty to focus attention on urgent or important tasks. These emotions cannot occur unless there is a meta-management layer to which the concept of “loosing control” becomes relevant. Without meta-management, which provides some sort of evaluation and control of thought processes, there cannot be any loss of control: you can’t lose what you don’t have [Sloman 99]. Tertiary emotions correspond to secondary emotions which reduce self-control.

This emotion classification scheme does not form a one-to-one mapping with our common usage of emotion labels. Fear can be generated as an innate response to a situation / event – a *primary* emotion; by cognitively identifying a threat – *secondary* emotion; or, in an extreme case, it can even involve loss of control of our attention/deliberation mechanism – a *tertiary* emotion. However, this scheme maps nicely onto research in the fields of psychology [Frijda 86], and neurology [Damasio 96; LeDoux 96].

### 4.3 Personality

The Webster’s on-line dictionary defines personality as: “the complex of characteristics that distinguishes an individual or a nation or group; *especially*: the totality of an individual’s behavioural and emotional characteristics.” Here the phrase “totality of ...” is important, in so far as it acknowledges that personality refers to many different non-specific behavioural and emotional characteristics of the individual. *Emotion* and *Personality* are cluster concepts.

#### *Psychological Models*

There are many competing psychological models which attempt to identify the fundamental dimensions of personality – where the “dimensions of personality” are seen as patterns of covariation of traits across individuals, and not as the specific organisation of attributes within an individual [McCrae and John 91]. These models are typified by the Five Factor Model (FFM) which uses the dimensions of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Although, no real consensus exists, it is generally agreed that some variation of these five dimensions (i.e. Eysenck’s [91] Psychoticism, Extraversion, and Neuroticism (PEN) three factor model - where psychoticism blends agreeableness and conscientiousness) will be necessary for an adequate description of individual differences.

Being able to classify personality types offers some hope for the specification and direction of like-like characters through the development of intentional models of personality - to complement the OCC [Ortony, Clore, and Collins 88] model of emotions. However, we feel that it is just as important to identify the information processing structures and dependencies that lie beneath these patterns, both from the perspective of understanding the interaction of control states better, and in order to build more consistent models of human-like agents. As a starting point, the Process and Component centric models being developed in cognitive computer science offer an promising compromise.

#### *Process Centric Models*

Process centric models of personality [Rousseau 96] take as their basis the personality trait theories mentioned above, and attempt to marriage the different dimensional traits with the information processing process of the agent architecture. There is good reason to believe that such a marriage is possible, simply because personality trait theories (i.e. FFM) are often

based on an analysis of the relationship between commonly used personality terms (and so map naturally onto the human information processing processes they describe).

Process	Inclination		Focus	
	Level	Value(s)	Aspect	Value(s)
Perceiving	Low High	Absentminded Alert	Expectations Reality	Imaginative Realistic
Reasoning	Low	Silly	Undesired effects Facts	Pessimistic Objective
	High	Rational	Desirable effects	Optimistic
Learning	Low	Incurious	What is learned only Both what is learned and what is already known	Gullible Open-minded
	High	Curious	What is already known only	Selective Intolerant
Deciding	Low	Insecure	First reaction	Impulsive
	High	Self-confident	Good decision	Thoughtful
Acting	Low	Passive	Anything besides the task	Indifferent
	Intermediate	Active	Task	Diligent
	High	Zealous	Result of the task	Perfectionist
Interacting	Low	Introverted	Addressee as a threat Exchange of information	Hostile Neutral
	High	Extroverted	Addressee as a help	Friendly
Revealing	Low	Secretive	Lie	Dishonest
	High	Open	Truth	Honest
Feeling Emotions	Low	Emotionless	Self	Selfish
	High	Sensitive	Others	Unselfish

**Figure 6 - Dimensions of a Personality [Rousseau 96]**

Rousseau's dimensions of personality (see Figure 6) bear a strong family resemblance to the Five Factor Model (with a slightly different emphasis). Although the dimensions are viewed as distinct processes, architectural constraints are likely to confer dependencies between them – *perceiving* and *reasoning* might both compete for the same deliberative resources, whereas *revealing* and *interacting* are both associated with coping strategies. Process centric models provide an intermediate step between the personality traits of psychological theories and the architectural requirements of an agent architecture.

### ***Component Centric Models***

Elliott [92, page 29] identifies two components of personality: (i) **interpretive personality** – the *rudimentary “personality” which gives agents individuality with respect to their interpretations of situations* (i.e. their uniquely individual concerns); and (ii) **manifestative personality** – the *rudimentary “personality” which gives agents individuality with respect to the way they express or manifest their emotions*. Table 1 lists some typical determinates of agent personality.

What determines an agent <i>personality</i> ?	Type
a) Agent <i>concerns</i> (motivators, goals, standards and attitudes).	<i>interpretive</i>
b) The motivational profile of the agent.	<i>interpretive</i>
c) The sensitivity of emotional states	<i>interpretive</i>
d) The balance between spontaneous and planned responses to affective events.	<i>manifestative</i>
e) Persistence in attaining goals or longevity of moods.	<i>manifestative</i>
f) Choice of words and style of scripts.	<i>manifestative</i>

**Table 1 - Agent Personality**

- a) *Agent Concerns*: The emotional and behavioural characteristics of an individual are essentially determined by the individual's concerns (or sources of motivation). These concerns might be explicitly encoded as active *plans* and *goals*, lie dormant within *standards* and *attitudes*, or even be dispositionally encoded in autonomic responses such as *reflexes* and *skills*.
- b) *Motivational Profile*: The relative weightings of the different classes of motivation give the broad *motivational profile* of the agent. Morignot and Hayes-Roth [94, 96] base their motivational profile on the work of Maslow [54] (see Table 2). By creating a profile where  $W_{\text{aff}} > W_{\text{ach}} > W_{\text{learn}}$ , we could imagine an "altruistic" agent that held the well-being of the user above either its own desires to achieve goals or learn user requirements.

Motivations of human agents	Interpretation for An Agent	Motivational Profile
Physiological	Energy	$W_{\text{phys}}$
Safety	Feeling Threatened	$W_{\text{safe}}$
Affiliation	Affective State of User	$W_{\text{aff}}$
Achievement	Achieving own goals	$W_{\text{ach}}$
Learning	User Requirements	$W_{\text{learn}}$

**Table 2 - Motivational Profile**

In a similar vein, Rizzo et al. [97] offer a model of personality based on the relative relevance of events to six high-order concerns (see also Ford [92]):

*Resource Provision*: Giving approval, support, assistance, advice, or validation to others. Avoiding selfish or uncaring behaviour.

*Material Gain*: Increasing the amount of money or tangible goods one has. Avoiding the loss of money or material possessions.

*Social Responsibility*: Keeping interpersonal commitments, meeting social role obligations, and conforming to social and moral rules. Avoiding social transgressions and unethical or illegal conduct.

*Belongingness*: Building or maintaining attachments, friendships, intimacy, or a sense of community. Avoiding feelings of social isolation or separateness.

*Image*: receiving positive evaluations from others, both in the realm of competence – being evaluated as skilful, intelligent, resourceful, etc. – and in the moral sphere – being evaluated as honest, generous, considerate, etc. Avoiding social disapproval.

*Entertainment*: Experiencing excitement or heightened arousal. Avoiding boredom or stressful inactivity.

Using such a classification scheme, the four personality types of Altruist, Normative, Selfish, and Spiteful can be mapped on to the high level concerns, i.e. Altruist = +Resource Provision, +Belongingness, -Material Gain, and +Image; Normative = +Social Responsibility, and +Image; Selfish = +Material Gain, -Resource Provision, -Social Responsibility, and +Image; whereas Spiteful = +Entertainment, -Social Responsibility, -Resource Provision, and -Image. Different combinations of motivational classes will of course lead to different personality traits.

- c) *The Sensitivity of Emotional States (Neuroticism)*: By differentiating between those concerns that lead to the emergence of emotive and non-emotive states it is possible to support the *neuroticism* dimension of personality. However, as emotions are an ill-defined motivational class, the definition of an emotive state is somewhat subjective and highly dependent on the underlying information processing architecture of the agent. Within our motivated agent framework, the emergence of perturbant states (*tertiary emotions*) is a natural outcome of the interaction of heuristic reactive insistence assignment and subsequent re-evaluation by the deliberative management process.
- d) *Reactive and Deliberative Processes*: The balance between reactive and planned responses is often termed the impulsiveness / thoughtfulness of the agent. By raising the attention filter threshold we can create a thoughtful agent (at the expense of being attentive – i.e. the classic stereo-type of an absentminded professor). Assigning high insistence levels at the reactive level can create agents that have a tendency to over-react to events. Finally self-monitoring demons in the meta-management layer can determine how quickly the agent abandons goals or becomes *frustrated* with current progress.
- e) *Persistence in Attaining Goals (Conscientiousness)*: Meta-management processes can be used to determine how conscientious the agent is, i.e. the hysteresis assigned to the active top-level motivators.
- f) *Choice of Expressive Behaviour*: Expressive behaviours include choice of actions, words, gestures, facial expressions, and vocal intonations. The use of motivational control states such as *motivators*, *emotions*, and *moods*, allows an agent to co-ordinate expressive behaviour in a believable manner.

## 5 Society of Mind Agent Architecture

As a starting point in our investigation of concern mediation mechanisms, we have adopted Cañamero's [97] Motivated Society of Mind architecture [see also Minsky 87]. The Abbott architecture already captures a number of different concern processes within a unified framework – using the twin approach of homeostatic drives, and non-homeostatic affect amplifiers. A modified Abbott architecture – see Figure 7 – has been implemented in Pop11 using the SIM\_AGENT toolkit [Sloman and Poli 96], and is being used to explore the inter-relationship between agent motivators, emotions and personality.

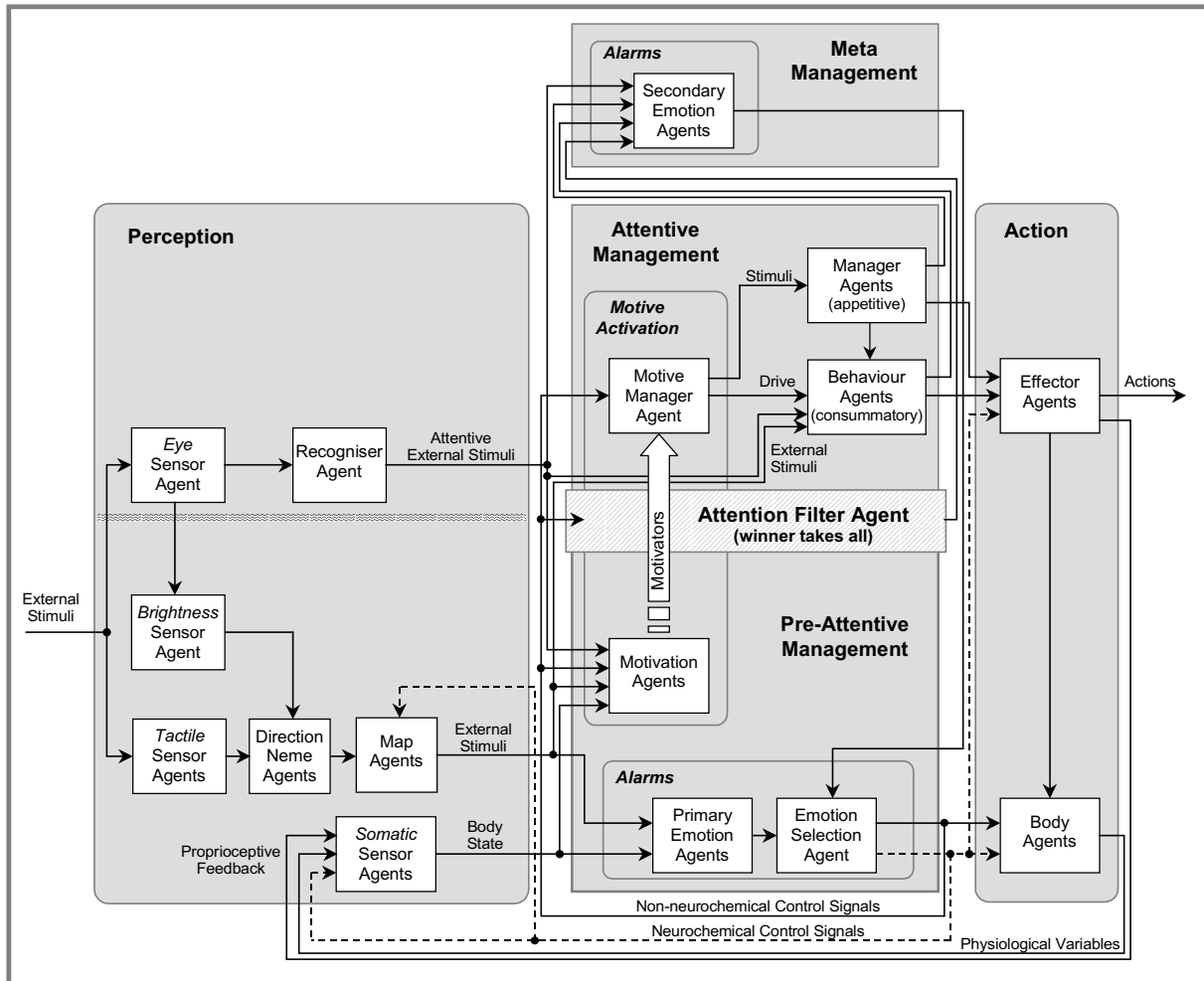


Figure 7 - Modified Abbott Architecture

### Biological Model

The adoption of an ethologically inspired drives-based model (see Tyrrell [93] for a review of drive-based action-selection models) allows us to simplify the task of differentiating between emotional and non-emotional drives, and retain a “pseudo-biological” flavour for the architecture. This “biological” angle is not unimportant, as it allows us to more easily relate our work to research in psychology and neuroscience. Furthermore, although an explicit drive

mechanism is not a pre-requisite for a life-like agent, it does make the implementation of reactive motivator generators cleaner.

The main modifications to Cañamero's original architecture have centred on clarifying the role of *emotion* agents and their associated chemical control signals. Separating the homeostatic drive mechanism from the affect amplification mechanism allows us to investigate the added value of affect based alarm and management / meta-management mechanisms on an already functioning system. *Emotion* agents were also partitioned into those lending themselves to primary and secondary emotional states – i.e. responding to pre-attentive and cognitively generated events respectively. The inclusion of the *emotion selection* and *attention filter* agents allowed the new architecture to support the complete action selection algorithm within the Society of Mind framework. Finally, the link between a behaviour and the results of executing a behaviour was made dispositional. *Behaviour* agents no longer directly alter physiological variables (this is done by *effector* and *body* agents), and are now selected on the basis of their expected effect – *BehaviourDrink* is expected to increase vascular volume by instructing the *mouth* agent to drink, and – by the same token – *BehaviourAttack* is expected to dispositionally decrease adrenaline levels by subduing the *anger* agent.

### ***Motivations, Emotions, and Personality***

Abbott's main motivational path consists of: (1) perception (*sensor*, *direction neme*, *map*, and *recogniser* agents); (2) reactive motivator generation and activation – Abbott is equipped with eight action tendencies with which to maintain its body state (Aggression; Cold; Warmth; Curiosity; Fatigue; Hunger; Thirst; and Self-protection). Each action tendency is represented in the architecture by a *motivation* agent which monitors the status of a single controlled variable (energy, temperature, ...). However, not all the controlled variables can be maintained within the desired range at the same time – in order to increase blood sugar Abbott needs to walk to find food, thus decreasing energy and increasing temperature; (3) motivator selection by the winner-takes-all *attention filter* agent; and (4) active motivator management by the *motive manager* agent (which selects the appropriate behaviour to satisfy the current active drive).

This main motivational path can be modified by the “emotional” path. Within Abbott, the emotional path has three roles: (1) to generate motivators from external events (i.e. converting one-off discrete events in to continuous drive-based signals); (2) to modify perception and prepare for action (i.e. reduce sensitivity to pain or make certain types of action more likely); and (3) to perform simple motivator management (i.e. switching tasks when the current action is not working).

Finally, by modifying the different components of the Abbott architecture, a rudimentary personality can be bestowed on our agent. For example, increasing the hysteresis associated with the attention filter will make Abbott more conscientious, and modifying the relative “gains” of the motivation agents will change Abbott's motivational profile.

## 5.1 Ongoing and Future Research

We are currently building a successions of complete “broad but shallow” agent architectures to integrate higher-level control states into our motivated agent framework. Our immediate research goal is the implementation of a more refined attentive layer which starts to capture the requirements of a true deliberative layer – allowing multiple sources of motivation to be considered at the attentive level. We would then like to integrate emotion based learning in to the architecture in order to investigate some of the longer-term motivational control states.

## 6 Conclusion

When building artificial autonomous agents and lifelike characters, we will often find it useful to refer to an agent’s underlying architecture using familiar mentalistic terms. In applying these mentalistic terms we implicitly assume a certain type of information processing architecture. This architecture does not need to contain specific mechanisms for each concept or property, as many concepts refer to emergent states, but the mechanisms employed must satisfy similar requirements if the use of mentalistic concepts is not to be misleading [Sloman and Logan 98]. For example, the mentalistic concept “frustration” places a requirement on the architecture to support a motivational attitude towards goal achievement.

We argue for an information-level “design-based” approach to the study of motivated autonomous agents. By adopting information level descriptions, we feel we are able to offer a rich explanatory framework for exploring human-like mental states in terms of the information processing and control functions of the agent architecture.

## 7 Acknowledgements

The author would like to thank Aaron Sloman and members of the Cognition and Affect Project at Birmingham University for the inspiration, background, and motivation for this work; and members of the Multiagent Systems and Intelligent User Interfaces groups at the German Research Centre for Artificial Intelligence for their ongoing support and lively discussions.

## 8 References

- Beaudoin, L. (1994). *Goal Processing in Autonomous Agents*. PhD Thesis, School of Computer Science, University of Birmingham.  
([ftp://ftp.cs.bham.ac.uk/pub/groups/cog\\_affect/Luc.Beaudoin\\_thesis.ps.Z](ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Luc.Beaudoin_thesis.ps.Z))
- Buck, R. (1985). Prime Theory: An Integrated View of Motivation and Emotion. *Psychological Review*, Vol. 92, No. 3, pages 389-413.
- Burt, A. (1998). Modelling Motivational Behaviour in Intelligent Agents in Virtual Worlds. In *Proceedings of the 1998 Conference on Virtual Worlds and Simulation*.  
(<http://www.dfki.de/~burt/papers/mrvw-ws.ps>)
- Cañamero, D. (1997). Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In *Proceedings of the First International Symposium on Autonomous Agents, AA'97*, Marina del Rey, CA, February 5-8, The ACM Press.  
(<http://www.ai.mit.edu/people/lola/aa97-online.ps>)

- Damasio, A. R. (1996). *Descartes' Error*. London: Papermac. (first published 1994, New York: G. P. Putman's Sons.)
- Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.
- Elliott, C. (1992). *The Affective Reasoner: A process model of emotions in a multi-agent system*. PhD Thesis, Northwestern University, Institute for the Learning Sciences Tech. Report #32  
(<ftp://ftp.depaul.edu/pub/cs/ar/elliott-thesis.ps>)
- Eysenck, H. J. (1991). Dimensions of Personality: 16, 5, or 3? – Criteria for a Taxonomic Paradigm. *Personality and Individual Differences*, Vol. 12, pages 773-790.
- Ford, M. E. (1992). *Motivating Humans. Goals, Emotions, and Personal Agency Beliefs*. Newbury Park, Sage
- Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.
- Hexmoor, H., Lammens, J., Caicedo G., and Shapiro, S. C. (1993). *Behavior Based AI, Cognitive Processes, and Emergent Behaviors in Autonomous Agents*. Technical Report 93-15, University of Buffalo. April 1993.  
(<ftp://ftp.cs.buffalo.edu/pub/tech-reports/93-15.ps.Z>)
- Minsky, M. (1987). *The Society of Mind*. London: William Heinemann Ltd.
- Maslow, A. H. (1954). *Motivation and Personality*, Harper. (3<sup>rd</sup> Edition, Addison-Wesley, 1987).
- Moffat, D. (1997). Personality parameters and programs. In R. Trappl, and P. Petta (Eds.), *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*. New York: Springer-Verlag, pages 120-165.  
(<http://d056.psyc.cf.ac.uk/papers/emotion/Vienna>)
- Morignot, P. and Hayes-Roth, B. (1994). Why does an agent act? In *Knowledge Systems Laboratory Report KSL-94-76*, December 1994.  
([http://ksl-web.stanford.edu/KSL\\_Abstracts/KSL-94-76.html](http://ksl-web.stanford.edu/KSL_Abstracts/KSL-94-76.html))
- Morignot, P. and Hayes-Roth, B. (1996). Motivated Agents. In *Knowledge Systems Laboratory Report KSL-96-22*, July 1996.  
([http://ksl-web.stanford.edu/KSL\\_Abstracts/KSL-96-22.html](http://ksl-web.stanford.edu/KSL_Abstracts/KSL-96-22.html))
- Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.
- Newell, A. (1982). The knowledge level. In *Artificial Intelligence*, Vol. 18, pages 87-127
- Norman, T. J. (1996). *Motivation-based direction of planning attention in agents with goal-autonomy*. PhD thesis. DAI Unit, Department of Electronic Engineering, Queen Mary and Westfield College  
(<http://www2.elec.qmw.ac.uk/~tim/thesis/thesis.ps.gz>)
- Picard, R. W. (1997). *Affective Computing*. Cambridge, Mass: The MIT Press.
- Rizzo, P., Veloso, M. V., Miceli, M., and Cesta, A. (1997). Personality-Driven Social Behaviors in Believable Agents. *AAAI 1997 Fall Symposium on "Socially Intelligent Agents"*, AAAI Press Technical Report FS-97-02, pp. 109-114.  
(<ftp://pscs2.irmkant.rm.cnr.it/pub/paola/papers/SIA97.ps.gz>)

- Reilly, W. S. (1996). *Believable Social and Emotional Agents*. Ph.D. Thesis. Technical Report CMU-CS-96-138, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. May 1996.  
(<http://almond.srv.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/papers/CMU-CS-96-138-1sided.ps.gz>)
- Rousseau, D. (1996). Personality in computer characters. In *Working Notes of the AAAI-96 Workshop on AI/ALife*, AAAI Press, Menlo Park, CA, 1996.  
([ftp://ftp.ksl.stanford.edu/pub/pdoyle/personality\\_AAAI96.ps](ftp://ftp.ksl.stanford.edu/pub/pdoyle/personality_AAAI96.ps))
- LeDoux, J. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.
- McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. *Special Issue: The five-factor model: Issues and applications. Journal of Personality* 60: 175-215, 1992.
- Sloman, A. (1993). The mind as a control system. In C. Hookway, and D. Peterson (Eds.), *Proceedings of the 1992 Royal Institute of Philosophy Conference 'Philosophy and the Cognitive Sciences'*. Cambridge: Cambridge University Press, pages 69-110  
([ftp://ftp.cs.bham.ac.uk/pub/groups/cog\\_affect/Aaron.Sloman\\_Mind.as.controlsystem.ps.Z](ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_Mind.as.controlsystem.ps.Z))
- Sloman, A. (1999). Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?). To appear in K. Dautenhahn (Ed.) *Human Cognition And Social Agent Technology*, John Benjamins Publishing.  
([ftp://ftp.cs.bham.ac.uk/pub/groups/cog\\_affect/Sloman.kd.ps](ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Sloman.kd.ps))
- Sloman, A. and Logan, B. S. (1998) Architectures and Tools for Human-Like Agents, In F. Ritter and R. M. Young (Eds.), *Proceedings of the 2nd European Conference on Cognitive Modelling*. Nottingham: Nottingham University Press, pages 58-65.  
([ftp://ftp.cs.bham.ac.uk/pub/groups/cog\\_affect/Sloman.and.Logan.eccm98.ps.gz](ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Sloman.and.Logan.eccm98.ps.gz))
- Sloman, A. and Poli, R. (1996). SIM\_AGENT: A toolkit for exploring agent designs. In *Proceeding IJCAI Workshop on Agents Theories Architectures and Languages ATAL'95*, Springer-Verlag Lecture Notes in Computer Science, 1996.  
([ftp://ftp.cs.bham.ac.uk/pub/groups/cog\\_affect/Aaron.Sloman\\_Riccardo.Poli\\_sim\\_agent\\_toolkit.ps.Z](ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_Riccardo.Poli_sim_agent_toolkit.ps.Z))
- Tyrrell, T. (1993). *Computational Mechanisms for Action Selection*. PhD Thesis, University of Edinburgh.  
(<ftp://ftp.ed.ac.uk/pub/lrtt/>)
- Wright, I. P. (1997). *Emotional Agents*. PhD Thesis, School of Computer Science., University of Birmingham.  
(<http://www.cs.bham.ac.uk/~ipw/thesis.ps.Z>)