Concern Processing in Autonomous Agents

by

Stephen Richard Allen

A thesis submitted to
the Faculty of Science of
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
Cognitive Science Research Centre
The University of Birmingham
January 2001

Abstract

Concerns are broadly defined as *dispositions to desire the occurrence or non-occurrence of a given kind of situation*. In this thesis we present an information-level analysis of the mechanisms that render the *concerns* of intelligent agency in the symbolic, situated, and "emotional" programming paradigms – to give an account of the functions, constraints and types of concern processes, and to investigate cognitive architectures that are capable of supporting such processes.

Part I introduces the research topic and describes the *motivated agent framework* used within the Cognition and Affect Project, and this thesis, to elucidate the architectural requirements for intelligent autonomous agency. Part II focuses on the issue of concern-processing in autonomous agency. We identify weaknesses in current deliberative and behaviour-based design approaches, and provide two case studies of our concern-centric information-level design-based approach to intelligent autonomous agent design. Part III applies our design methodology to the requirements for human emotional states. We present a information-level analysis of leading theories of emotion, and describe a series of broad agent architectures for elucidating emergent infant-like emotional states. Part IV summarises the different strands of research presented within this thesis, and identifies some fertile areas for future research.

By describing a variety of functions using the design stance at the information-level, and showing how they account for human-like mental states and processes, we aim to provide a rich explanatory framework for intelligent autonomous agency.

Acknowledgements

I would like to express my deep gratitude to Aaron Sloman for his inspiration, guidance, and energetic commitment to my research, and for establishing the tradition of the Cognition and Affect Project at the University of Birmingham.

I would like to thank Matthias Klusch, Elisabeth André, Martin Klesen, Catriona Kennedy, Luc Beaudoin, David Moffat, Juan Velásquez, Stan Franklin, Lee McCauley, Lola Cañamero, Michael Schillo, Alastair Burt, Steffen Hantschel, Patrick Gebhard, Klaus Fischer, Jörg Siekmann, Brian Logan, John Fox and Donald Peterson for their support, encouragement, and valuable critiques of earlier drafts of this thesis. I would also like to thank Manfred Kerber, Ricardo Poli, and Peter Hancox for their continued support as members of my thesis group.

Finally, I would like to thank Thorsten Bohnenberger and Susanne Baier-Allen for proof-reading the final drafts.

This research was partially supported by a University of Birmingham Studentship.

Table of Contents

PART III

List of Illustrations

List of Tables

List of Definitions and Abbreviations

Abbott is the name of the agent initially developed by Cañamero [97] and extended within the confines of this thesis – chapters "Emotional" Agents, 7, and 8.

Affects are the innate biological drives, or motivating systems most commonly associated with emotion – chapter 5.

AFP – Attention Filter Penetration theory [Sloman 92] – section 4.1.

ANA – Agent Network Architecture [Maes 89] – section 3.2.3.

Appraisal is the process of evaluation of significance or meaning.

BDI – Belief-Desire-Intention architecture [Bratman 87] – section 3.1.1.

CAP – Cognition and Affect Project at Birmingham University – chapter 4.

CNS – Central Nervous System.

Cognitive Appraisal is the process of appraisal at a complex information-processing level.

Concerns are broadly defined as dispositions to desire the occurrence or non-occurrence of a given kind of situation – chapter 3.

Control states are information-bearing representations of an information-processing control system – section 2.1.

Core-self refers to the innate concern-processing mechanisms of the agent – section 7.1.2.

Dispositional States are cognitive states dispositionally related to action (i.e., connected, but not as a necessary and sufficient condition). Dispositions can be related to one another hierarchically, some having very indirect links with external behaviour.

Effectance Motivation is the intrinsic need for an effective interaction with the environment, which makes exploration, stimulus seeking, and a wide variety of related behaviours rewarding.

Emotional Arousal is the dispositional state that results from affective appraisal. States of emotional arousal share a number of special features: valence; change in action readiness; intensity; insistence; and persistence.

Emotional Episodes are states of emotional arousal which include interruption of attentive processing.

Extended-self refers to the concern-processing mechanisms that evolve as the agent interacts with the environment – section 7.1.2.

FFHA – Free-Flow Hierarchical Architecture [Tyrrell 93a] – section 3.2.4.

FGCA – Fine-Grained Cognitive Architecture [Rosenblatt and Payton 89] – section 3.2.2.

Information-level descriptions refer to descriptions of the representational form (i.e. *motivations* and *goals*) that the information takes at each stage of the control process – section 2.1.

Insistence is defined as the propensity to pass through the attention filter and thereby divert and hold attention – section 4.1.

KAs – Knowledge Areas – section 3.1.1.

MINDER1 is the name of the agent developed by Wright [97] for elucidating proto-emotional states – section 4.4.

Motivational Control States move an agent towards a desired physical/mental state.

NML1 is the name of the agent developed by Beaudoin [94] for elucidating goal-processing in autonomous agents – section 4.2.

Perturbance is an emergent dispositional state in which an agent temporarily loses attentive control over some of its management processes – section 4.3.

Primary Affects are based upon specific hard-wired systems in the brain involving general response tendencies to the environment.

Primary Appraisal is the process of relevance evaluation within the reactive layer of an agent architecture. Appraisal at this level must rely on quick and simple heuristics, not complex information-processing – section 5.2.

Primes are the primary motivational systems – reflexes, instincts, primary drives, primary affects, and effectance motivations.

PRS – Procedural Reasoning System [Georgeff and Lansky 86] – section 3.1.1.

SoM – Society of Mind [Minsky 87] – section 6.1.4.

Subsumption Architecture [Brooks 86] argues that the design of an agent should be constructed along the lines of behavioural competence levels wherein the higher competence levels are able to completely subsume the behaviours of the lower levels. The architecture can be partitioned at any level, with the levels below forming a complete and self-contained agent – section 3.2.1.

Valence is the positive or negative 'character' of an appraisal.

Ventromedial Prefrontal Cortex in neuroanatomical terminology is the underbelly (ventral) of the frontal lobe in the proximity to the midline (medial) or inside surface of the brain (sometimes also referred to by the less specific term orbital cortex). Damage to this region has a number of serious effects on cognition and emotion – appendix C.

Part I

Introduction

1 Introduction

> *"The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions. I suspect that once we give machines the ability to alter their own abilities we'll have to give them all sorts of complex checks and balances."*
>
> – Minsky, *The Society of Mind* (section 16.1)

In the following scenario, consider the tasks and abilities of a nursemaid in charge of four toddlers, Tommy, Dicky, Mary, and Chloe.

> "One morning, under the nursemaid's supervision the four children are playing with toys. Mary decides that she wants to play with Dicky's toy. So she approaches him and yanks the object out of his hands. Dicky starts to sob, as he cries out "mine! mine!" The nursemaid realises that she ought to intervene: i.e., to take the toy away from Mary, give it back to Dicky, and explain to Mary that she ought not to take things away from others without their permission. This task is quite demanding because Dicky continues crying for a while and needs to be consoled, while Mary has a temper tantrum and also needs to be appeased. While this is happening, the nursemaid hears Tommy whining about juice he has spilt on himself, and demanding a new shirt. The nursemaid tells him that she will get to him in a few minutes and that he should be patient until then. Still, he persists in his complaints. In the afternoon, there is more trouble. As the nursemaid is reading to Mary, she notices that Tommy is standing on a kitchen chair, precariously leaning forward. The nursemaid hastily heads towards Tommy, fearing that he might fall. And, sure enough, the toddler tumbles off his seat. The nursemaid nervously attends to Tommy and surveys the damage while comforting the stunned child. Meanwhile there are fumes emanating from Chloe indicating that her diaper needs to be changed, but despite the distinctiveness of the evidence it will be a few minutes before the nursemaid notices Chloe's problem." – [Beaudoin 94, page 1]

This human scenario highlights some of the many challenges future researchers must face as we attempt to integrate autonomous agents into our complex human world. As a valuable first step towards meeting these challenges, we propose the development of an explanatory framework within which to explore and describe the human actions and mental states we hope to emulate. Using this framework we can then start to develop an understanding of the architectural requirements that underlie such mentalistic terms as *motives*, *goals*, *intentions*, *concerns*, *attitudes*, *standards* and *emotions*, and how they relate to reactive and resource-bounded practical reasoning. Finally, by building complete agents, and testing them in realistic scenarios, we will then be in a position to start to learn how these mentalistic control states interact.

The research described within this thesis takes a number of decisive steps towards developing such a framework, and an understanding of the architectural requirements and design trade-offs that underlie some of our more common mentalistic terms and concepts.

1.1 Research Contributions

This research makes a number of contributions towards developing a framework to describe and elucidate concern-processing in intelligent autonomous agents – a more detailed description of these contributions is provided in chapter 9.

Framework for Analysing/Designing Intelligent Autonomous Agents

(Parts I, II, and III)

Consolidating earlier work by the Cognition and Affect Project, we argue for a motivated agent framework consisting of three strands: (i) a concern centric view to the requirements of intelligent autonomous agency; (ii) a cognitively inspired three-layered agent architecture for analysing and building intelligent autonomous agents; and (iii) an information-level, design-based research methodology. Within the context of this framework, we present an analysis of concern-processing in both the symbolic and situated AI programming paradigms – i.e. those of resource-bounded practical reasoning and behaviour-based architectures.

Analysis of Human and Artificial "Emotional" States

(Parts I, II, and III)

Dismissing the wholesale adoption of the intentional stance [Dennett 87], we argue that the use of certain mentalistic concepts can still be justified by referring such concepts to the underlying information-level processing mechanisms of the system. Within our motivated agent framework, we present an analysis of the control mechanisms associated with the emergent mental phenomena we normally term emotion. Supportive evidence for this approach is provided by mapping leading cognitive theories of affect from psychology and neuroscience [Frijda 86; Damasio 94; LeDoux 96] on to our framework.

Design of an Intelligent Autonomous Agent for Elucidating "Emotional" States

(Part III)

Using Cañamero's [97] motivated *Society of Mind* architecture as a starting point (see also [Minsky 85]), we develop a series of broad agent designs that systematically address different aspects of concern-processing identified in **part II**. These designs culminate in Abbott3, an implementation of a cognitively inspired intelligent autonomous agent architecture for elucidating emergent "emotion-like" states.

Toolkit for Building Intelligent Autonomous Agents

(Appendix)

Extending the SIM_AGENT toolkit [Sloman and Logan 98], we add a graphical front-end and development environment for building, testing, debugging, and analysing intelligent autonomous agents. This toolkit forms the heart of the Gridland and Nursemaid Scenarios used extensively in the development of the intelligent autonomous agent architectures described in this thesis.

1.2 Research Methodology

One of the challenges faced by researchers in the construction of intelligent autonomous agents is the need to develop a systematic framework in which to answer questions about the types of control mechanisms such agents might need, and how those different control mechanisms might interact. In this section, we argue for an information-level design-based approach to the study of intelligent autonomous agents – wherein each new design gradually increases our explanatory power and allows us to account for more and more of the phenomena of interest. These broad designs help to build our understanding of the different attributes of information-level representations, their functional roles, and their causal relationships. Further, by adopting information-level descriptions, 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 underlying architecture.

1.2.1 Intentionality

"Intentionality" is a philosophical term for aboutness. Something exhibits "intentionality" if its competence is in some way *about* something else. A thermostat is an "intentional" system – it contains representations of both the current temperature (the curvature of the bimetallic strip) and the desired temperature (the position of the dial). Autonomous agents are also "intentional" systems, but at levels of richness and complexity orders of magnitude greater than the humble thermostat.

Treating agents (people, animals, objects, or machines) as "intentional" systems is one of the techniques we use in our everyday lives to understand the behaviour of complex systems [Dennett 78, 87, 96]:

1) *The physical stance*. We apply the physical stance to objects when we refer our predictions to the classic laws of physics, i.e. objects fall to the ground because they are subject to the law of gravity. The physical stance affords us a great deal of confidence in our prediction.

2) *The design stance*. When we wish to understand and predict features of *design*, we need to adopt the design stance. The design stance allows us to ignore implementation details and make predictions based on *designed for* characteristics, i.e., that the alarm clock will make a loud noise at 7:15.

3) *The intentional stance*. We adopt the intentional stance whenever we treat observed systems *as if* they were rational agents who governed their "choice" of "action" by a "consideration" of their "beliefs" and "desires." The intentional stance is the most powerful, and yet the most risky of Dennett's predictive stances. Its riskiness stems from two connected problems: (i) we are non-privileged observers having to infer intention (in the philosophical sense of *aboutness*) from observed behaviour; and (ii)

> complex systems are inherently resource-bounded, and as such can only approximate rationality (without rationality there can be no basis for inferring intention from observed behaviour). But even with these caveats, the intentional stance is still a remarkably robust tool. It allows us to make workable predictions about the external behaviour of very complex systems such as animals and other human beings.

Dennett suggests that "*if done with care*, adopting the intentional stance is not just a good idea, but offers the key to unravelling the mysteries of the mind" [Dennett 96, page 27]. However, such an approach extorts a heavy price: (a) care must be taken not to confuse the philosophical term "intentionality" (*aboutness*) with the common language term referring to whether someone's action was intentional or not – as in the case of intentional control states [Bratman 87] (and section 3.1.1); and (b) care must also be taken to recognise the limits of agent rationality. Much behaviour is simply *automatic* (neither rational or irrational), and devoid of any form of "consideration". Such behaviour often appears rational because we are adept at spotting patterns and regularities in our environment. Some of these regularities are derived from the *designed for* characteristics of the system, be that a chess playing machine designed to win, an animal designed to carry genes from one generation to the next, or a stressed nursemaid designed to handle multiple goals. Other regularities emerge from the *physical* characteristics of the system, i.e. the resource constraints of the architecture, or the temperature of the room.

In reality, the limits of agent rationality, and the requirement of balancing multiple competing concerns in an unknowable environment, ensures that the "intentional stance" is at best a methodology of approximation rather than one of design and analysis. By assuming that systems behave *as if* they were rational agents the "intentional stance" allows us to approximate behaviour by approximating the "intentionality" (*aboutness*) of the system. However, these approximations invariably mask the real "intentionality" of the constituent components, leading to an overestimate of the complexity of the system in what Braitenberg calls the "law of uphill analysis and downhill invention" [Braitenberg 84, page 27].

1.2.2 The Design-based Approach

There is another approach. Complex systems can also be understood through a *succession* of designs, in the downhill mode of invention. Here, each design gradually increases our explanatory power and allows us to account for more and more of the phenomena of interest.

The design-based approach [Sloman 93b; Beaudoin 94; Wright 97] takes the stance of an engineer attempting to build a system to exhibit the phenomena/behaviour of interest. Formally, this can be represented as a recursive methodology with five parallel threads of execution. Threads 1-3 represent common engineering practices, and threads 4-5 give the methodology the rigour needed for scientific validity:

1) *A requirements analysis of the system of interest*, i.e. a specification of the capabilities of the autonomous agent using information-level descriptions. These should include: the key features of the environment; the resource constraints within the agent; the behaviours the agent must exhibit and their causal links; and a description of the agent's concerns and coping strategies. A preliminary requirements analysis is given section 1.3, with more detailed requirements specifications given in subsequent chapters.

2) *A design specification for a working system to meet those requirements*. This is an architectural analysis of the design, to include its major components and the causal links between these components. A design can be recursive, replicating threads 1-5 at individual component levels, i.e. a low-level implementation specification of one component and a theoretical analysis of another.

3) *A detailed implementation or implementation specification of the working system*. Depending on the objectives of the research, this can take the form of a simulation with predictive power, or a realistic model, accurate to some level of detail. In this thesis we will develop a cognitively inspired agent architecture for elucidating "emotional" states. Our agents will initially be developed in the Gridland Scenario (see sections 6.1.4 and appendix A).

4) *A theoretical analysis of how this design meets the initial requirements*. It is more than likely that an implementation will not meet all the requirements set out in the requirements analysis. A design verification analysis is therefore required to determine the extent to which: (a) the design meets the requirements; and (b) the implementation/simulation embodies the design. Ideally this should take the form of a rigorous mathematical proof, but in practice we must rely on intuitive analysis combined with systematic testing of the implementation.

5) *An analysis of similar designs in design-space*. By considering the implications of alternative options to a particular design, we can often obtain a deeper understanding of that design. The literature review in **parts II** and **III** can be seen as part of this process of exploration. The experimental results described in chapter 8 provide a further exploration of the design-space.

1.3 Requirements of Autonomous Agency

Before starting on our quest towards a better understanding of concern-processing in autonomous agents, we must first establish exactly what we mean when we talk about intelligent autonomous agents:

1) An *autonomous agent* is a system situated within, and as part of, an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to affect what it senses in the future [Franklin and Graesser 96].

2) An *intelligent agent* is a versatile and adaptive system that performs diverse behaviours in its efforts to achieve multiple goals in a dynamic, uncertain environment [Morignot and Hayes-Roth 94].

By combining these definitions we get the outline of a set of basic requirements for our intelligent autonomous agent. Namely, it must be capable of: (i) handling multiple sources of motivation with limited resources; (ii) having and pursuing an agenda; and (iii) being robust and adaptable in the face of a hostile and uncertain environment.

Autonomous agents have multiple sources of motivation. These sources vary in their nature, importance, urgency, duration, and range of associated behavioural responses. Motivations need to be generated asynchronously to each other, and they must be capable of interrupting/diverting ongoing activity (mental and/or physical). Autonomous agents have limited resources with which to satisfy these motivations. They move at finite speeds, they have a fixed number of manipulators/sensors, their processing is bounded, and they have limited knowledge of the environment.

Autonomous agents must be capable of having and pursuing an agenda. That is to say, they must have a purpose or "mission" in life. This agenda might simply be to preserve its own well-being, or it might be required to balance its own needs against those of its supervisor/programmer/provider.

Autonomous agents must be robust and adaptable in the face of uncertain and dynamic environments. In particular, their beliefs may be out-dated, false or even contradictory, their internal processes may operate asynchronously and at different speeds, and their intentions/ actions might fail. Robustness and adaptability require action on two levels: (i) at a motivation processing level to select alternative behaviours when initial behaviours fail to satisfy a concern; and (ii) at a motivation generation level to modify the agent's motivational profile to better match its environment (reducing or increasing the sensitivity to certain concerns).

1.4 Thesis Structure and Guide

This thesis is presented in the engineering style of the "design-based" research methodology [Sloman 93b] to guide the reader towards a greater understanding of the types of mechanisms that render the *concerns* of intelligent autonomous agents.

Part I introduces the research topic and describes the *motivated agent framework* used within the Cognition and Affect Project, and this thesis, to elucidate the architectural requirements for intelligent autonomous agency; **part II** focuses on the issue of concern-processing in autonomous agency. We identify weaknesses in current deliberative and behaviour-based design approaches, and provide two case studies of our concern-centric information-level design-based approach to intelligent autonomous agent design; **part III** applies our design methodology to the requirements for human emotional states. We present a information-level analysis of leading theories of emotion, and describe a series of broad agent architectures for elucidating emergent infant-like emotional states; **part IV** summarises the different strands of research presented within this thesis, and identifies some fertile areas for future research; the **references** section provides pointers to the primary and secondary sources of literature used within this research; the **appendices** provide supportive background information for the thesis itself.

Although we have written each chapter as a self-contained module, the earlier chapters do provide useful background material for the concepts presented later. We would therefore recommend that at least some of this earlier material is read before launching into the heart of the thesis described in **part III**. However, we also recognise that readers are in the best position to decide on the relevance of each chapter to their own particular interests, and so a brief guide to each chapter is provided to aid this navigation process:

Part I
Introduction

Chapter 1 The first chapter provides a general introduction into the problem area by establishing: (i) the research objectives; (ii) the research methodology; and (iii) a requirements specification for intelligent autonomous agency.

Chapter 2 The second chapter presents the main strands of the motivated agent framework used within the Cognition and Affect Project. We introduce the idea of a mind as an information-processing control system, and identify some of the control states that are likely to play an important role in intelligent autonomous agency. We also take the first steps towards elucidating these control states by describing their *functional* attributes, and proposing a three-layered model within which to explore the *structural* and *dimensional* attributes.

Part II
Concern Processing

Chapter 3 The third chapter provides a design-based analysis of concern-processing in existing deliberative and behaviour-based autonomous agent designs. We argue that many of the identified weaknesses in existing designs can be addressed by taking a concern-centric stance towards intelligent autonomous agent design.

Chapter 4 The fourth chapter analyses previous work completed within the Cognition and Affect Project in relation to concern-processing in intelligent autonomous agent architectures. We introduce Sloman's Attention Filter Penetration theory of emotions [Sloman 92], and explain how the architectural requirements imposed by a dynamic and uncertain environment can lead to the emergence of proto-emotional states [Beaudoin 94; Wright 97]. This chapter forms the initial design specification for an agent architecture to meet the basic requirements of intelligent autonomous agency.

Part III
"Emotional" Agents

Chapter 5 The fifth chapter presents an information-level design-based analysis of the phenomena we commonly call emotion. We start by arguing that a lot of the confusion surrounding the term emotion can be attributed to the fact that different theorists focus on different concern-processing mechanisms (*reactive*, *deliberative*, or *reflective*) active in the emotion process – this is related to our argument that emotions are emergent mental states. We then extend our analysis by mapping leading cognitive theories of emotions [Frijda 86; Damasio 94; LeDoux 96] on to our motivated agent framework, and identify the different mechanisms active in *primary*, *secondary* and *tertiary* emotions.

Chapter 6 The sixth chapter presents an information-level design-based analysis of "emotional" agent architectures. We start with a brief overview of related work on emotional agents [Moffat and Frijda 95; Velásquez 96; McCauley and Franklin 98; and Cañamero 97]. We then present two implementations of broad-but-shallow "emotional" agent architectures – integrating different control states active in the emotion process into an extended motivated *Society of Mind* (based on Cañamero [97] and Minsky [85]). These implementations look at both deliberative and reactive mechanisms of concern mediation within our motivated agent framework.

Chapter 7 The seventh chapter presents an abstract design of a cognitively inspired agent architecture for elucidating "emotional" states – integrating the different research strands explored in chapters 1 through 6. We describe how the different concern-processing competence levels of our three-layered architecture co-evolve, and identify the different processes active in the emergence of "emotional" states.

Chapter 8 The eighth chapter presents an implementation of our agent design, and an analysis of similar designs in design-space. We also present a critique of our design, and address some of the architectural requirements needed to support basic human emotions.

Part IV
Conclusions

Chapter 9 Chapter nine summarises the contributions this research makes to the field of understanding concern-processing in intelligent autonomous agents, and points to new directions in which the research can be taken in the future.

Chapter 10 Chapter ten provides a list of references to the primary and secondary literature sources used within this thesis.

Appendices

Appendix A describes the extensions we made to the Sim_Agent [Sloman and Poli 96] toolkit to provide the test and development environment for this thesis.

Appendix B explains how to run the source code provided with each of the Abbott agent architectures developed in the thesis – described in chapters 6 and 8.

Appendix C provides a brief overview of the important structures involved in both reasoning and emotion in the human brain. This appendix provides useful background information for our analysis of the neurological basis for emotions in chapter 5.

Appendix D provides an overview of the different types of chemical messengers (hormones) active in the human brain – giving useful background information for our analysis of emotional agents in chapters 6, 7, and 8.

Appendix E describes the evolution of mind from the perspective of our "selfish" genes and "selfish" memes – providing the context for future work described in chapter 9.

1.5 Summary

In this chapter we have introduced the research objectives, the research methodology, and a requirements specification for a cognitively inspired intelligent agent. In the next chapter we will provide some scaffolding for this framework by introducing the terminology of *mentalistic control states*, and a cognitively inspired three-layered agent architecture. In **parts II** and **III**, we will further extend the framework by: (a) analysing case studies on the requirements of goal-processing [Beaudoin 94] and proto-emotions [Wright 97] in autonomous agents; (b) using the framework to describe the *functional*, *dimensional*, and *structural* attributes of the mentalistic concept we call "emotion"; and finally (c) building an agent that supports emergent "emotional" control states.

2 Motivated Agent Framework

When attempting to understand human behaviour we often find it useful to build a partial model of a person's internal mental state – in a sense we adopt an extended "intentional stance" (see section 1.2.1). In applying mentalistic concepts such as "frustrated", "happy", or "stressed", we implicitly make certain assumptions about the type of information-processing architecture required to support such states. Although the architecture does not need to contain specific mechanisms for each mentalistic concept, as many concepts refer to emergent states, the mechanisms employed within the architecture must satisfy certain requirements if the use of mentalistic concepts is not to be misleading [Sloman 97]. For example, the mentalistic concept of "frustration" places a requirement on the architecture to not only support a motivational attitude towards goal achievement, but also a mechanism to detect when a goal is not being achieved.

The notion of using mentalistic concepts to describe the internal state of artificial agents will inevitably meet with some opposition – we are not claiming that at anytime in the foreseeable future our agents will have the same degree of richness of mental states as we have ourselves. However, the use of *certain* mentalistic concepts can be justified by referring such concepts to the underlying information-level processing mechanisms of the system (see Sloman [97] for a more detailed discussion on this point). Further, in exploring mentalistic concepts in artificial agents, we can start to make valuable inferences as to the types of mental information-processing mechanisms that are necessary to support similar states in humans.

This chapter lays the foundation for such an approach by presenting a motivated architectural framework for an intelligent autonomous agent. We use the label "motivated" to reflect our concern-centric design stance – focusing on the requirements of explicit motivational control states within the architecture.

2.1 The Mind as a Control System

In the following discussion we will start from the assumption that *human minds are incredibly complex information-processing machines*. If we also adopt the definition of information as "something to which some process is causally sensitive" (or to use a more memorable slogan, "information is *a difference that makes a difference*" [Bateson 72 in Chalmers 97, page 281; also Franklin 95, page 34]), we can start to see the problem we face when attempting to understand the information-processing machinery of the mind – How do you describe processes that act on an intangible commodity which itself only exists in relation to a *choice* of process within the machinery you wish to describe?

Different scientific disciplines take different approaches to the problem of studying the mind: neurologists study the physical structures of the brain and make educated guesses as to each structure's function by looking at brain activity with modern scanning tools, the way in which the structures are inter-connected, and the disabling effects of disease; psychologists treat the mind as a black-box and make inferences as to its internal structure by observing how people and animals react to experiments from without; and philosophers dream up elaborate thought experiments to dissect the mind from within. Our solution is to: (a) treat the mind as an incredibly complex information-processing *control system*; and then (b) adopt the engineering style of the design-based approach (section 1.2.2) to study mental phenomena by actually building systems that meet similar requirements (i.e. explore the design space of artificial minds).

What Are Control States?

> "[T]he idea of a mind as [a] control system leads to a new analysis of the concept of 'representation': a representation is part of a control state: and different kinds of representations play different roles in control mechanisms." – [Sloman 93a, page 7]

At the start of this chapter we talked of the need to refer mentalistic concepts (such as "frustrated", "distracted", and "stressed") to information-level descriptions of the underlying information-processing architecture. But with a substance as ethereal as information, what do we really mean by *information-level descriptions* of the architecture?

Being "frustrated" places a requirement on the architecture to support a motivational attitude towards goal achievement. Our information-level descriptions therefore refer to descriptions of the representational form (i.e. *motivations* and *goals*) that the information takes at each stage of the control process – we do not actually need to model each neuron to create an artificial mind, we just need to capture the flow of information created by those neurons (although the former might help the latter). When we talk about information-level descriptions of the minds of autonomous agents, we refer to descriptions of information-bearing representations, or – if we treat the mind as a control system – *dispositional control states* (adopting the terminology of a *control state* also allows us to grow our agent architectures step-by-step without having the burden of accounting for the full richness of human *mental attitudes* and *mentalistic concepts* right from the beginning).

Complex control systems (such as the minds of humans or intelligent autonomous agents) are capable of supporting many different types of control state – we regularly base our predictions of human behaviour upon observations such as "he is in a bad mood", without referring to the *desires* or *beliefs* of the observed person. The fact that we can use "mood" as a predictor of behaviour is a good indicator that it refers to some concrete configuration/state of the underlying information-processing structure that is called a mind. 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; whereas (b) S*tructural attributes* 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 (at every step of the control process information has the potential to be transformed into other forms of representation) – for example, mechanisms which transform *beliefs* into either *standards* or *attitudes*, or modify exiting *beliefs* on the basis of new *goals*. Finally, we allow values of dimensional attributes to be expressed in terms 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.

In addition to dimensional and structural attributes, we also find it useful to talk about control states at a more abstract level – and so we will add *functional attributes* to the above list. *Functional attributes* simply describe the role a control state plays within a control system, without attempting to describe how this is achieved.

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. Furthermore, 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*).

What Control States are Needed for Intelligent Agency?

There are no hard and fast rules for determining the number and nature of control states needed for intelligent 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 with dedicated circuits for every function. 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). The flexibility of our approach allows us to explore the requirements of *control states* without committing ourselves to rigid representational forms. Figure 2.1-1 shows the variation in scope and duration of some likely *control states* of intelligent agency.

Long term	Personality, Skills.
Relatively hard to change, very slow learning, causes and effects diffuse and indirect.	Attitudes, Standards, Preferences.
Short term	Moods, Emotions.
Changeable, more specific causes and effects, semantic content.	Desires, Intentions, Plans.

Figure 2.1-1 Control States of varying Scope and Duration

Classification of 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 formalism. Control states can operate asynchronously, at different rates, and at different/multiple levels within the architecture. Table 2.1-1 gives the functional attributes of some of the common control states used in the design of intelligent autonomous agents.

Control State	Functional Attributes
Goals	*Goals* are states-of-affairs towards which the agent is motivationally directed. Goals form the junctures of *plans* and sub-plans, and are generated in response to *concerns*. Goals are considered purely deliberative structures within the context of this research, marking a boundary between reactive and deliberative concern representations. Reactive goal-processing mechanisms (such as Nilsson's [94] Teleo-Reactive Programs) are more akin to skills – Table 2.1-2.
Intentions	*Intentions* are states-of-affairs towards whose achievement the agent has made some form of a commitment [Bratman 87; Cohen and Levesque 90].
Plans	*Plans* are semantically rich structures that guide an agent towards a goal. Plans make use of a number of deliberate control states such as *images* and *goals*, as well as beliefs about reactive control states such as *skills* and *competencies*.
Beliefs	*Beliefs* are states-of-affairs which the agent holds to be true or false, but due to the dynamics of the environment and the limits of perception of the internal/ external world model, may or may not be so. Beliefs are probabilistic "statements of fact" that range from high probability beliefs about new sensor data, to beliefs about beliefs, and beliefs about inferred reactive beliefs. Wright [97] distinguishes two types of belief: sensor-based beliefs, which are generated with regard to sense data, and agent-based beliefs, which are generated with regard to agent actions.
Behaviours	*Behaviours* are arbitrary complex action patterns made in response to external stimuli or inferred internal motivational states. We distinguish between appetitive and consummatory behaviours: (i) *Consummatory* behaviours provide an immediate direct benefit to the agent; whereas (ii) *Appetitive* behaviours require further expenditure of time and energy before any benefit is realised.

Table 2.1-1 Functional Attributes of Common Control States of Intelligent Agency

In addition to the common control states already used in autonomous agents, we can also identify some more esoteric states (see Table 2.1-2) that we feel are likely to play a role in future intelligent autonomous agents – for example, Ortony et al. [88] differentiate emotional states on the basis of whether events match/mismatch *goals*, *standards*, and *attitudes*; whereas Damasio [96] uses *images* and *somatic markers* to connect deliberative reasoning with reactive processing.

Control State	Functional Attributes
Motivators	*Motivators* are mechanisms and representations that tend to produce, or modify, or select between actions, in the light of beliefs [Sloman 97]. Motivators are present in both the deliberative and reactive domains. In the reactive domain, motivators have the disposition to generate action through *reflexes*, *behaviours*, or modes of action readiness, whereas in the deliberative domain, motivators have the disposition to generate action through *images* and *goals*. In humans it appears that the reactive motivators are not directly accessible from the deliberative domain and must be inferred through actions and felt modes of reactive action readiness – we are only sometimes aware of the real reasons for our actions.
Emotions	*Emotions* are complex modes of action readiness change that redirect attentive processing towards situations or events that impinge on agent concerns. Damasio defines an emotion as "the combination of a *mental evaluative process*, simple and complex, with *dispositional responses to that process*, mostly *toward the body proper*, resulting in an emotional body state, but also *toward the brain itself* (neurotransmitter nuclei in brain stem), resulting in additional mental changes." [Damasio 96, page 139]
Images	*Images* are internal multi-modal representations of potential, present, or past states-of-affairs. Images utilise many of the mechanisms of perception and are accessible to reactive mechanisms. Humans use *images* to examine the consequences of actions – providing a feedback path to sense reactive responses to *goals*, or infer reactive *beliefs* (see Damasio's *Somatic Marker Hypothesis* – section 5.2.5).
Somatic Markers	*Somatic Markers* are a "*special instance of feelings generated from secondary emotions*. Those emotions and feelings *have been connected, by learning, to predicted future outcomes of certain scenarios*." [Damasio 96, page 174]
Standards	*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	*Attitudes* are dispositions, or predispositions, to like some things and to dislike others without reference to standards or goals [Ortony et al. 88].
Reflexes	*Reflexes* are ballistic forms of action that can be specified by a narrow set of rules based on simple input integration [Beaudoin 94]. Like emotions, reflexes can also cause redirection of attention, but they differ from emotions in that they are not accompanied by the feeling state associated with *somatic markers*, or an actual change in physiological arousal.
Skills	*Skills* are learnt action patterns which can be executed with simple perceptual (visual, auditory, touch, or proprioceptive) feedback.

Table 2.1-2 Functional Attributes of Future Control States of Intelligent Agency

Having established the initial functional attributes of some useful forms of information representation for intelligent autonomous agency, we must now provide an architectural framework on which we can hang our information-level descriptions of the *dimensional* and *structural* attributes of these cognitively inspired mentalistic control states.

2.2 A Three-Layered Cognitive Model

Figure 2.2-1 shows an architectural framework for an intelligent autonomous agent [Sloman 97; Beaudoin 94; Wright 97]. This framework identifies three basic types of concurrent concern-processing mechanisms (co-existing sub-systems) such an agent is likely to possess – with an attention filter to protect the deliberative motivator management layer from excessive interruption. We make no commitment to specific information-level structures or processes at this stage.

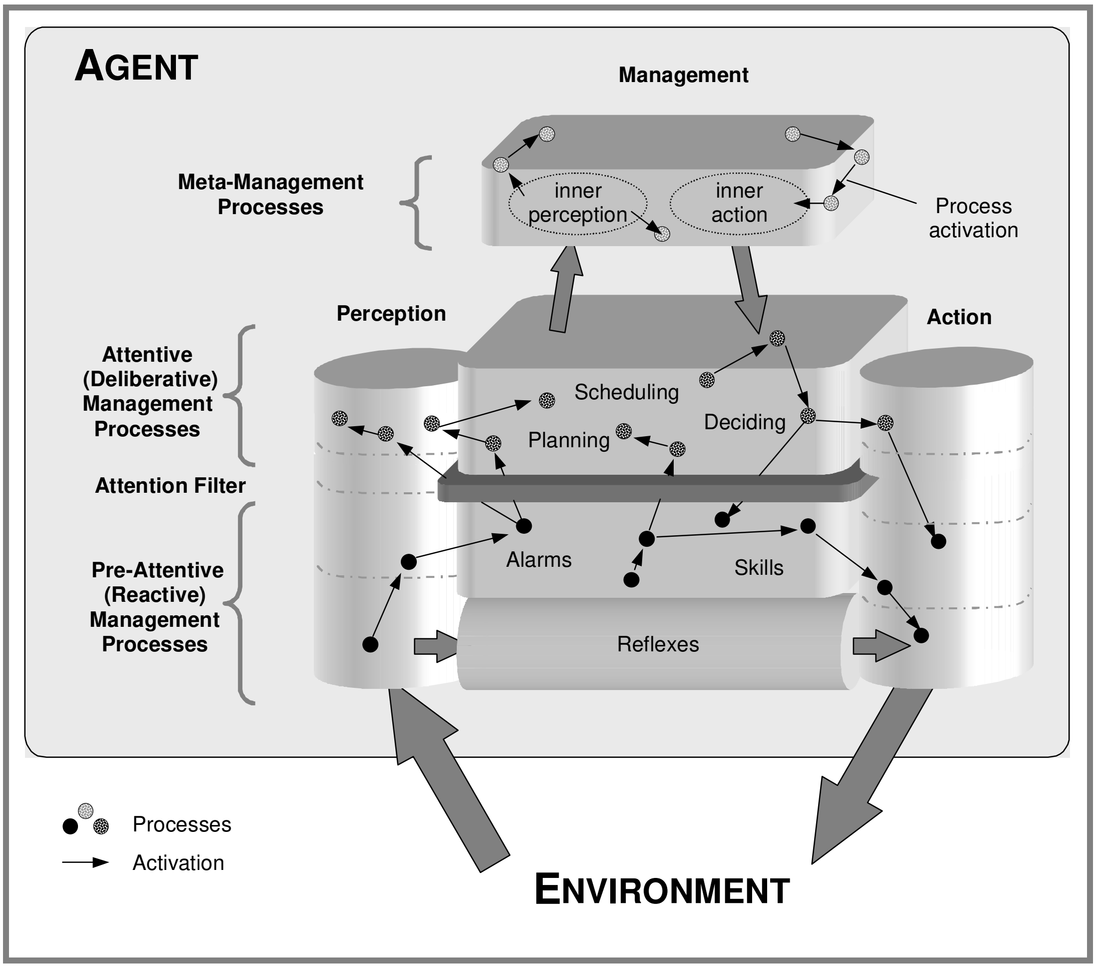


Figure 2.2-1 Towards an Architecture for an Intelligent Autonomous Agent

Pre-Attentive (Reactive) management processes 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.

Attentive management processes 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-class 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.

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.

Meta-management processes are responsible for monitoring and controlling motivator management mechanisms. It is likely that approaches that work well in an agent's early development may become less-than-optimum as the agent's environment (including its internal environment) change. Meta-management processes enhance the agent's adaptability and robustness by continually evaluating the agent's 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 operating primarily through the attentive state of the agent – although the reactive concern-processing mechanisms are also subject to longer-term adaptive change.

2.3 Summary

In this chapter we have introduced the concepts of (a) the mind as a *control system*, and (b) representations as *information-bearing sub-states* of that control system. Armed with this conceptual framework, we identified three attributes (*functional*, *structural*, and *dimensional*) with which to describe the control states of intelligent agency, and proceeded to describe the *functional* attributes of the control states we feel will be needed in an intelligent autonomous agent that is able to deal with the real-time constraints and complexity of human environments. Finally, we introduced an architectural framework within which we can describe the *structural* and *dimensional* attributes of these control states.

Part II

Concern Processing

3 Concern Processing

> *"Concern is defined as a disposition to desire occurrence or nonoccurrence of a given kind of situation. Such dispositions are assumed to exist when an individual initiates activity to achieve given kinds of situations and spends time, effort, or money in doing so; or when he explicitly expresses desire to achieve such a kind of situation; or when there is emotional response upon events implying achievement or nonachievement of such a kind of situation."*
>
> – Frijda, *The Emotions* (page 335)

Information exists in relation to a *choice* of process, and in intelligent autonomous agency that choice is made in relation to the *concerns* of the agent. In this chapter we argue for a concern-centric approach to the design of intelligent autonomous agents.

The architectural requirements for the expression of agent concerns have received very little attention in the research literature on autonomous agent design (a notable exception being Bratman [87]). This is somewhat surprising given the fact that concern-processing lies at the heart of autonomous agency, but can in part be attributed to the success that has so far been achieved with our existing repertoire of concern-processing control states proposed by Bratman (beliefs, desires, and intentions) and others. However, as we indicated in the last chapter, *belief*, *desire*, and *intentional* control states alone are unlikely to be sufficient to build the types of complex intelligent autonomous agents that populate environments such as our Nursemaid scenario – a view partly born out by current research trends towards hybrid and affective agent architectures (where the emphasis is on utilising different types of control structures at different levels of the agent architecture), and partly by noting the semantic richness of the mentalistic terms we use in our everyday life to describe human behaviour and action. In this chapter, we will attempt to further elucidate the requirements for intelligent autonomous agency (see section 1.3) by providing a *concern-centric* analysis of existing deliberative and behaviour-based agent architectures.

3.1 Deliberative Concern Processing

The traditional "symbolic" AI approach to rational agency has been to treat control states as semantically rich information structures to be manipulated according to Newell's principle of 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]). Naturally, this approach has concentrated on those state-of-affairs type control states with clearly defined structural attributes that can be *approximated* at a single high-level of abstraction (i.e. the short-term states of Figure 2.1-1 – *beliefs*, *desires*, *plans* and *intentions*). The *concerns* of such rational agents are either expressed as active *goals*, or else lie dormant within Scripts [Schank and Abelson 77], Teleo-Reactive Programs [Nilsson 94], Reactive Action Packages [Firby 89], or

Knowledge Sources/Areas [Hayes-Roth 85, 95; Georgeff and Lansky 86]. In this section we will concentrate on the concern-processing credentials of the Procedural Reasoning System.

3.1.1 The Procedural Reasoning System

The Procedural Reasoning System (PRS) [Georgeff and Lansky 86; Georgeff and Ingrand 89] is a generic BDI (belief-desire-intention) architecture for representing and reasoning about actions and procedures in dynamic environments. BDI architectures are so named as they view *beliefs*, *desires*, and *intentions* as being necessary (and to some extent *sufficient*) control structures for resource-bounded practical reasoning agents [Bratman 87] – see Georgeff and Rao [95] for an argument as to the necessity (although not adequacy) of beliefs, desires and intentions in domains where real-time performance is required from both a quantitative decision-theoretic perspective and a symbol reasoning perspective.

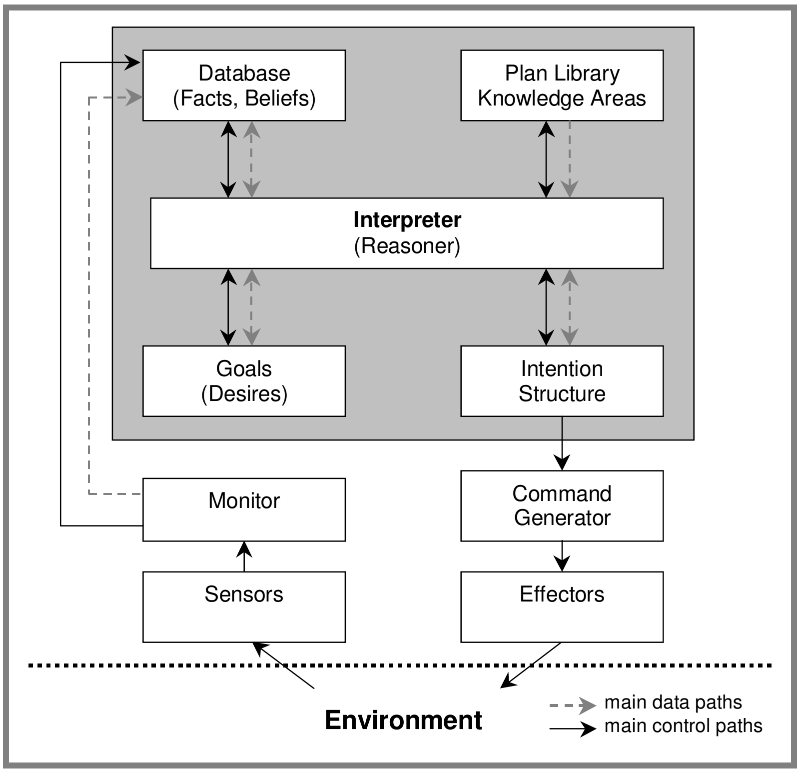


Figure 3.1-1 PRS Architecture

The Procedural Reasoning System (PRS) architecture, shown in Figure 3.1-1, can be sub-divided into peripheral and reasoning components. The peripheral components are: (i) *sensors*; (ii) a *monitor* to translate sensor information into agent beliefs; (iii) a *command generator* to translate primitive actions into effector commands; and (iv) *effectors* to generate events in the outside environment. The reasoning components are: (i) a *database* containing current beliefs and facts about the world which is automatically updated; (ii) a library of plans, called *Knowledge Areas* (KAs) used to achieve given goals or react to particular situations; (iii) a set of current *goals* (PRS uses the term *goals* to refer to *desires* that are consistent and obtainable); (iv) an *intention structure* containing a partially ordered set of all plans (KA stacks) chosen for execution at runtime; and (v) an *interpreter* (or inference mechanism) to select appropriate plans in response to system beliefs and goals, commit the selected plans to the intention structure, and execute them.

Operation

The operation of the PRS interpreter can best be explained in two steps: (i) as an idealised BDI interpreter; and (ii) by invoking meta-level KAs to explain the **Deliberate()** process.

The idealised BDI interpreter loop is shown in Figure 3.1-2: (i) the interpreter checks for new events; (ii) for every new event (belief and/or goal) in the event queue, the **OptionGenerator()** attempts to unify the event against the invocation conditions of each of the knowledge areas (KAs). Unified KAs are added to the list of possible plan options; (iii) the interpreter then uses **Deliberate()** to select an appropriate KA for insertion into the intention structure; (iv) the chosen KA is inserted into the intention structure (the intention structure contains multiple intention stacks, which are either running in parallel, suspended until some condition occurs, or ordered for execution in some way). If the KA arose out of the acquisition of a new goal or belief, it is inserted as a new intention stack. If the KA arose out of the processing of an existing intention, it is pushed on to the stack of KAs associated with that intention.; (v) the interpreter selects an intention from the root of one of the intention stacks, and executes one step of that intention. This results in either the performance of a primitive action (including mental actions of believing some new proposition), or the establishment of a new subgoal; (vi) successful and impossible attitudes (in this case, goals and intentions) are removed from the goal database and intention structure; (vii) the interpreter returns to the start of the loop.

BDI – interpreter

```
InitialiseState();
repeat
  EventQueue = GetNewEvents();
  Options = OptionGenerator(EventQueue);
  SelectedOptions = Deliberate(Options);
  UpdateIntentions(SelectedOptions);
  Execute(SelectIntention());
  DropSuccessfulAttitudes();
  DropImpossibleAttitudes();
end repeat
```

[Rao and Georgeff 95, section 5]

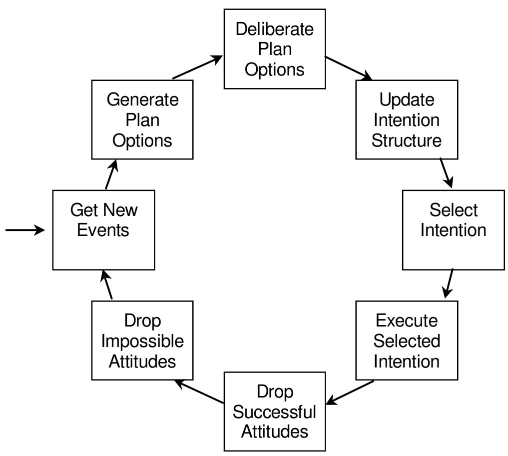


Figure 3.1-2 Idealised BDI Interpreter Loop

The actual PRS interpreter loop is shown in Figure 3.1-3. The PRS interpreter implements the **Deliberate()** function as a meta-level KA. The meta-level KA is executed from the intention structure in the same way as any other intention, and is therefore interruptible, maintaining the reactivity of the PRS. The actual operation of the interpreter loop is as follows: (i) the interpreter checks for new events; (ii) **FindSoak()** returns a Soak (Set Of Applicable KAs) which unify with the new events. If the Soak is empty the interpreter jumps to iv (below), selecting an intention from the existing intention structure and executing one step of that intention; (iii) if the Soak is not empty, the interpreter *recursively* (a) posts meta-facts about the Soak on to the database (generating new events); and (b) unifies these events against meta-level KAs to create a new Soak; *until* (c) the new Soak returned is an empty list. The previous (non-empty) Soak is retrieved, and a meta-level KA (to perform the **Deliberate()** function) is chosen at random for insertion into the intention structure; (iv) the interpreter loop inserts the intention into the intention structure and continues as described in the idealised BDI interpreter loop above.

PRS – interpreter

```
InitialiseState();
repeat
  Soak = FindSoak(GetNewEvents())
  while (ListEmpty(Soak) == False)
    PostSoakMetaFacts(Soak);
    PreviousSoak = Soak;
    Soak = FindSoak(GetNewEvents())
  end while
  if (ListEmpty(PreviousSoak) == False)
     PostSoakMetaFacts(Soak);
     Intend(SelectRandomly(PreviousSoak));
  end if
  Execute(SelectIntention());
  DropSuccessfulAttitudes();
  DropImpossibleAttitudes();
  PreviousSoak = Soak;
end repeat
```

(Soak = Set of Applicable KAs)

[Ingrand and Coutance 93, section 3.3]

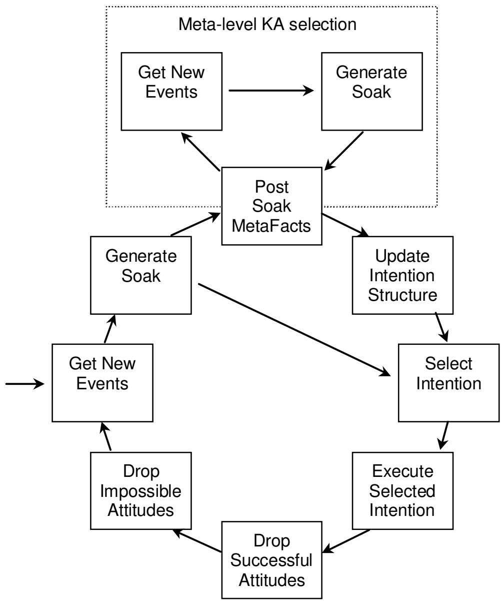


Figure 3.1-3 PRS Interpreter Loop

PRS Operators

The PRS operators allow the system to represent a wide variety of goals, including goals of achievement, goals of maintenance, and goals to test for given conditions. The basic PRS operators [Ingrand et al. 96] are: (i) **achieve C** (sometimes written as **!C**) which causes the system to attempt to achieve the goal **C**; (ii) **test C** or (**?C**) which tests for the condition **C**; (iii) **wait C** or (**^C**) which causes the system to suspend the current intention until the condition is true; (iv) **preserve C** or (**#C**) which passively preserves the goal **C**; and (v) **maintain C** or (**%C**) which actively maintains the goal **C** (interrupting the current activity to re-establish the goal should it fail). The PRS language also includes two operators to modify beliefs about the environment (including beliefs about the internal state of the system): **assert C** or (**=>C**), and **retract C** or (**~>C**).

Event- and Goal-Driven Behaviour

Although the posting of beliefs and goals are both treated as events, the notion of goal- and event-driven behaviour is very different in PRS. Goals (stored in the goal database) are treated as objectives that the interpreter attempts to satisfy *by any means*, i.e. by systematically trying one KA after another. Goal-driven behaviour is tested for success/failure on each successive adoption of a KA, and only considered to have failed after all valid KAs (i.e. those that unify with the goal) have been adopted and failed. By comparison, event-driven behaviour proceeds without any analysis of the eventual success or failure of the unified KAs, selected KAs are simply placed on the intention structure and left for later adoption and execution.

Analysis

PRS offers a very promising approach to concern-processing within resource-bounded deliberative reasoning. It supports shifting attention (of intentions), event- and goal-driven behaviour, and parallel execution of actions and planning. A tight interpreter loop and step-wise intention execution maintains the reactivity of the system (within provable limits), and meta-level KAs can be used to create a layer of “virtual applications” on top of the basic interpreter – to augment the in-built reasoning capabilities of the system.

The real strength of PRS’s concern-processing credentials lies in its ability to extend the basic goal- and event-driven behaviour through the judicial application of meta-level KAs. As we mentioned above, meta-level KAs give the designer the freedom to implement many different types of architecture on top of the basic PRS cycle – we will discuss one such architecture (NML2) in chapter 4.

However, a number of problems have been identified with the standard PRS approach when evaluated against the concern-processing requirements of intelligent autonomous agency: (i) new events are unified against a monolithic KA database, for complex domains requiring many KAs and meta-level KAs this could drastically reduce the responsiveness of the system; (ii) Myers [96] points out that the success/failure semantics of PRS’s goal-driven behaviour can become very restrictive when attempting to monitor continuous action processes; (iii) the position of KAs in the intention structure orders them in time, but says very little about the importance of the intention or how it was derived. UM-PRS [Huber et al. 95] goes some way to address this by including the facility to explicitly assign a priority value to the achievement of a goal; (iv) PRS uses an exhaustive strategy to test the success/failure of goals. PRS-CL [Wilkins and Myers 95; Wilkins et al. 95] includes an **Achieve_By** operator to filter candidate KAs and constrain the goal-processing process.

PRS was never conceived of as a generic architecture for autonomous agency – its main application was intended to be in the area of resource-bounded practical reasoning. Most of the identified problems can be resolved by applying suitable indexing techniques on the KA plan library, and through the judicial use of meta-level KAs. Unfortunately, the more reflective the interpreter loop becomes (selecting meta-level KAs to arbitrate meta-level KAs), the less reactive the whole system is made – no matter how efficient the indexing is. Although PRS is capable of reacting to new events in its belief database, it can only respond to them by generating an intention to do so, and thus at the expense of its ongoing reasoning. In fast changing environments it becomes essential to filter events before they reach the attention of the PRS cycle (making them inaccessible for considered attention later). This has meant that PRS increasingly finds itself in a supervisory role in autonomous agent designs. Here responsiveness is needed, but the flow and breadth of information is much reduced. The procedural nature of most supervisory tasks has also favoured the PRS approach.

Finally, from a concern-centric design stance, the adequacy of the semantic richness of goals within the standard PRS architecture is also in doubt. Work by Beaudoin [94] has identified the need for a rich representation of motivators when dealing with complex human-like environments – i.e. our Nursemaid scenario described in the introduction. We will discuss Beaudoin's enhancements to the standard PRS architecture in section 4.2.

3.1.2 Conclusions

The advantage of using a bounded rational agency approach (as typified by the Procedural Reasoning System) to concern-processing, is the ability to completely describe the behaviour of the agent at a human-intelligible level – giving a high degree of confidence in the responsiveness and correctness of the design. However, many of the requirements of intelligent autonomous agency are in stark contrast to the neat world of bounded rational agency – agents hold inconsistent beliefs, have multiple competing concerns, and must be robust in the face of hostile and unknowable environments. These requirements place additional real-time constraints on the concern-processing mechanisms of the agent – for example, new operators may be needed; new intermediate processing steps (i.e. through learning, or manipulation within specialist working memories) may be required; completely new forms of representation (i.e. affective states) may need to be added to the system; and asynchronous reactive processes may be required to filter the incoming percepts. With this added complexity comes an inevitable breakdown in our ability to completely capture the design of an agent at a "single" high-level of abstraction (although good approximations at different levels of abstraction are often achievable – as in the "intentional stance").

In recognition of the special problems of intelligent autonomous agency, various solutions have been proposed centring around multi-layered architectures. These multi-layered architectures generally combine a behaviour-based reactive layer with traditional planning and goal-based reasoning layers – achieving responsiveness and robustness at the expense of a certain degree of predictability. For example: *Touring Machines* [Ferguson 92] has a reactive, planning, and modelling layer; *GLAIR* [Hexmoor et al. 93] has a Sensori-Actuator, a Perceptual-Motor, and a Knowledge layer; *CYPRESS* [Wilkins et al. 95] has an executor, a planner, and an action library; *Saphira* [Myers 96; Konolige et al. 97] has an effector, a behaviour, and a task layer; *InteRRaP* [Müller 96; Jung 99] has a behaviour-based, local planning, and co-operative planning layer; *Architecture for Autonomy* [Alami et al. 98] has a functional, an execution control, and a decision layer.

The partitioning of multi-layered architectures is often made purely on functional grounds, with little attention paid to the emergent properties created by the interaction of the short-term control states encoded in the many layers (an exception is the *GLAIR* architecture which is partitioned to investigate the learning of emergent behaviours). When designing multi-layered autonomous agent architectures, we must not only think about the processes that happen within a layer, but also pay careful attention to the attributes of those emergent control states created by the processes that operate between the different layers. Until we recognise that these emergent properties form an integral part of the system – in humans *moods*, *emotions*, and *personality* often serve as the basis for our predictive stances –, we will never be able to attain a reasonable degree of confidence in the responsiveness and correctness of our multi-layered agent designs in the types of open environments we described in the introduction.

3.2 Behaviour-Based Concern Processing

A different approach to concern-processing is championed by the behaviour-based (or situated AI) school of thought. Behaviour-based architectures adopt the stance that agent control systems should be decomposed according to the desired external behaviours of the system, rather than divided along functional lines based on internal workings of the solution [Rosenblatt and Payton 89]. This emphasis on behaviours has the positive effect of embedding agent architectures in the environment, but leaves wide open the question of the mechanisms through which behaviours are selected. Concerns are not explicitly manipulated within behaviour-based systems – the manipulation of explicit representations is avoided altogether –, but processed implicitly within the confines of the behaviours themselves, and the network of connections between behaviours.

Behaviour-based architectures generally address the issues of concern-processing in one of two ways: (a) Subsumption architectures (sections 3.2.1 and 3.2.2) use the physical structure of the architecture to determine the priority of competences, and thus implicitly the assignment of motivational attitude; whereas (b) Command Fusion architectures (sections 3.2.2 to 3.2.5) calculate motivational attitude in a dynamic manner through spreading activation energy models.

The behaviour-based movement has opened up fertile new avenues of research into the requirements of concern-processing in autonomous agency. In the following analysis, we will focus on two issues: (i) the concern-processing mechanisms active in the behavioural action selection process itself; and where relevant, (ii) how the different architectures attach *valence* and *motivational attitude* to situations and events in service of these concerns. This process should be viewed as part of the ongoing requirements specification for the reactive layer of an intelligent autonomous agent architecture.

3.2.1 Subsumption

A central tenet of the subsumption architecture is that behaviours are arranged in a vertical structure as *levels of competence* – see Figure 3.2-1. This structure can be partitioned at any level, with the levels below forming a complete and self-contained specification of the agent.

Once a level has been debugged it is ideally never altered. New competences are provided by building higher levels on the competences of the levels below (subsuming behaviours where conflicts exist). These higher level modules have access to the intermediate results of the lower levels at the inputs and outputs of the modules. A typical layered structure for a simple mobile robot could consist of: (i) level 0 – avoid objects; (ii) level 1 – wander; (iii) level 2 – explore; (iv) level 3 – build maps; (v) level 4 – monitor changes; (vi) level 5 – identify objects; (vii) level 6 – plan changes to the world; and (viii) level 7 – reason about behaviour of objects.

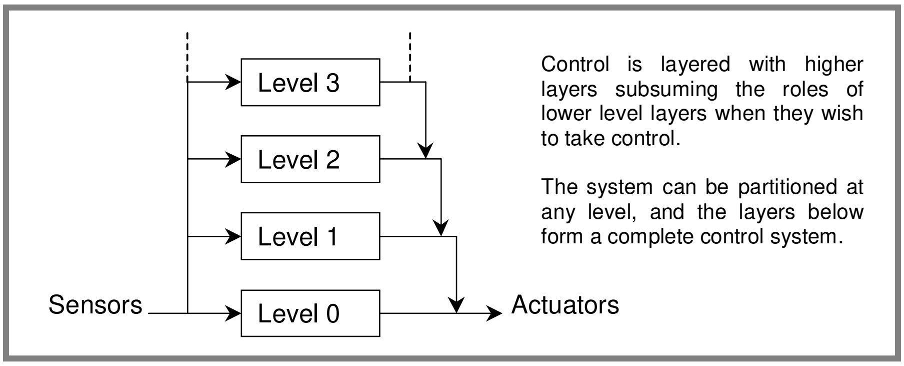


Figure 3.2-1 The Control Layers of a Subsumption Architecture [Brooks 86]

An important feature of the *vertical* decomposition of the subsumption architecture is that control is distributed among a number of behaviours operating asynchronously and in parallel. Each behaviour is concerned with a narrow aspect of the control process, and only needs to receive sensory data which is directly relevant to its particular decision making needs. By appropriately fusing behaviour commands through arbitration, a robot control system can quickly respond to events in its environment without the delays imposed by centralised sensor fusion. But, as low-level modules are able to process data faster than modules higher up in the structure, the system must be able to cope with conflicting messages being transmitted when low-level modules manage to send their messages out before the subsuming behaviours have had time to react.

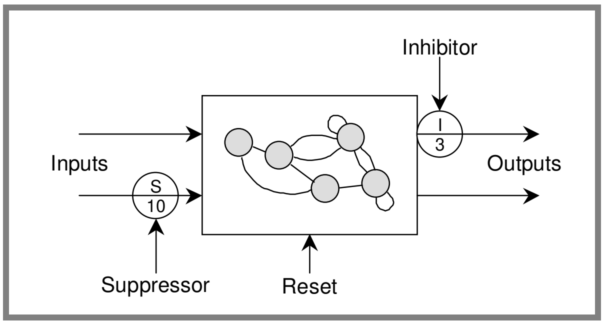


Figure 3.2-2 Subsumption Behaviour Module [Brooks 86]

Each behaviour module is implemented as a finite state machine, with internal variables as well as states (Figure 3.2-2). Behaviour modules are interconnected through a network of message carrying lines. Each module is capable of receiving messages along multiple input lines and generating messages on multiple output lines. Inhibitors may be attached to the outputs of a module and suppressors on the inputs. Any signal arriving on an inhibitor or suppressor control line blocks the messages on the corresponding output/input line of the module for a pre-determined period of time (shown by the number in the circle). In addition to blocking input messages, the suppressor control line also injects its own message on to the input line of the module. If more than one suppressor is present on a line, the injected messages are effectively "or"-ed together.

Analysis

As the only means a behaviour has of influencing another behaviour is by completely subsuming or inhibiting it, there is a real problem with command fusion between levels in a subsumption architecture. In an extreme case this can result in an architecture where a higher level completely subsumes the functionality of a lower level (duplicating the lower levels' behaviours) making the vertical structure redundant. Command fusion can occur if we use the dimension of time to represent the strength of a signal. The degree of subsumption can be controlled by varying the time constant and transmission rate of the subsuming signal. By comparing the number of conflicting messages to arrive in a fixed period of time, a module can accomplish a form of command fusion. However, this is not an ideal solution as it fails to account for signals arriving from two or more subsumption sources. It can also be argued that command fusion is not in the true spirit of subsumption.

The subsumption of a behaviour module results in the loss of information contained within the inhibited module. This loss may have significant repercussions on the other modules within the same level, causing a serious disruption to that level of competence. In practice the design of a particular level must be carried out with the competences of higher levels in mind at the start of the design process. This is necessary to ensure that the right hooks are present for the higher level behaviours when needed. Simply adding inhibitors to a low-level module may alter the functionality of a level to the extent that it needs redesigning.

Finally, the *motivational attitude* of behaviours are reflected by their relative position within the subsumption hierarchy. A level 0 design must account for the primary concerns of the agent – these will normally be related to self-preservation. As the levels of abstraction and representation increase through the vertical structure, the subsumption architecture can account for more and more specific concerns (such as achieving goals or minimising energy expenditure). The problem then becomes one of integrating these increasingly sophisticated (but potentially lower priority) concerns, with the primary concerns of levels 0 and 1. Ideally we would like our level 0 concerns to take precedence over the level *n* concerns in a reverse subsumption arrangement – at present this is achieved by relying on the speed of the lower layers, and the ability of the higher layers to recognise a high-priority concern and arbitrate accordingly.

The problems with the subsumption architecture can be summarised as: (a) inadequate command fusion; (b) inaccessibility of internal state; (c) disruption of levels of competence; and (d) apparent reversal of concern-processing priorities. We will initially discuss these problems within the context of Rosenblatt and Payton's [89] fine-grained connectionist architecture (see section 3.2.2 below), before returning to them again within the context of our motivated agent framework in chapter 7.

3.2.2 Fine-Grained Subsumption and Command Fusion

Rosenblatt and Payton [89] proposed a fine-grained connectionist architecture (FGCA) as a solution to some of the problems identified in the classic subsumption architecture. The FGCA uses the same vertical philosophy as the subsumption architecture. However, behaviours are not constructed from self-contained modules, but from fine-grained networks of smaller decision making units. Each unit represents a specific concept the designer is striving to implement. These concepts are achieved by transforming a set of input activations into a set of output activations as shown in Figure 3.2-3. The individual units are not allowed to contain internal variables, and so are completely transparent to the outside world.

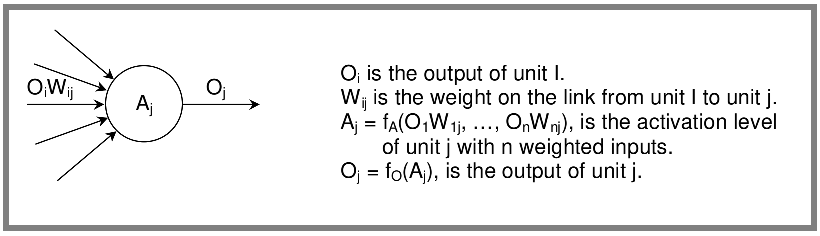


Figure 3.2-3 Flow of Activation in a Fine-Grained Connectionist Architecture

Although nodes are not allowed to contain internal variables, it is still possible to build sophisticated state machines out of node networks through the "selective unit update mechanism". When a new value is entered into an input unit, the links emanating from that unit not only carry a numeric value, but they also transmit a unique tag for one time step.

Receiving units can use this tag to selectively update their outputs. This simple mechanism converts a node into a latch, and thereby allows a node network to provide the same functionality of a standard subsumption state machine, should it be desired.

The activation and output functions for each unit can be any mapping from real numbers to a single real number value. The only constraint is that these functions are defined for inputs between –1 and 1, inclusive, and that the outputs be within this same range. There are no constraints on how the units are connected together, except a desire to maintain a structured behaviour-based subsumption architecture.

Analysis

In many ways the FGCA represents a universal agent architecture (with learning supported through selective updates). By reducing the granularity of the nodes, the FGCA approach goes a long way to solving many of the deficiencies of the subsumption architecture – but at some cost to the complexity of the design (it is important to get the concepts represented by the nodes that make up behaviour modules right): (a) command fusion is achieved by allowing the individual nodes within a behaviour module to directly communicate with the nodes of another behaviour module; (b) the inaccessibility of internal state problem is also addressed by the same mechanism – "[s]ince the processing elements in the system are extremely simple, behaviors are defined by the connections between elements rather than by the properties of the elements themselves. Since no information is hidden within these elements, the internal representations used in one behavior are completely accessible to all other behaviors" [Rosenblatt and Payton 89]; and (c) as behaviours are no longer completely subsumed, there is less chance of a higher-level behaviour causing a disruption within the lower levels of competence.

Although the FGCA utilises the vertical structure of competence layers, it is arguable as to whether it can still be called a subsumption style architecture – as behaviours within layers no longer completely subsume lower-level behaviours. It is therefore useful to distinguish between strong and weak forms of the subsumption style philosophy. The strong form of a subsumption architecture insists that behaviours within higher levels of competence completely subsume behaviours in lower levels. However, we can also adopt a weaker form in which the different levels of competence are allowed to co-evolve through processes such as learning and command fusion. We will return to the issue of co-evolving levels of competence in section 7.1.1.

Rosenblatt and Payton use a weaker form of the subsumption style framework in which behaviours "are constructed of units which maximize the amount of information available to other behaviors, so that the communication barriers between behaviors do not exist" [Rosenblatt and Payton 89]. The free-flow of command information between behaviours has been adopted by Tyrrell [93b] in an ethologically inspired autonomous agent design that addresses the issue of drive-based concern-processing. We will discuss Tyrrell's architecture in section 3.2.4.

3.2.3 Agent Network Architecture

The Agent Network Architecture (ANA), or spreading activation, approach to action selection [Maes 89] is closely allied to Minsky's [85] idea of a *Society of Mind* (see also section 6.1.4). Maes' agent is defined by a set of competence modules with an underlying network of connections. The approach is termed "spreading activation" to reflect the manner in which planning evolves as the result of the flow of activation energy throughout an entire network, and not from the actions of identifiable supervisor or planning modules. The network and competence modules themselves are static and predefined, but the control structure that emerges is dynamic and distributed.

Competence Modules and their Society

Competence modules resemble the operators of classical planning systems, and are described by the tuple (c_i, a_i, d_i, α_i): (i) c_i is the list of preconditions that have to be met before the module can become active, (ii) a_i and d_i represent the effects of a competence module's actions in terms of an add and a delete list, (iii) α_i is the activation level of the module.

The algorithm performs the following actions on each timestep:

1) It calculates the impact of the state, goal, and protected goals on the activation level of the modules.
2) It calculates the way the modules activate and inhibit related modules through their successor, predecessor, and conflicter links.
3) It normalises the activation levels of the modules to ensure that the average activation level remains constant.

The competence module that then fulfils the following three conditions is made active: (i) It has met all its preconditions (i.e. is *executable*); (ii) Its activation level is above a global threshold - θ; (iii) It has a higher activation level than all the other modules that satisfy conditions (i) and (ii).

When selected, the activation level of the chosen module is reset to 0, and the threshold level is reset to its initial value. If none of the modules are selected then the threshold is lowered by 10% for the next pass of the loop.

Figure 3.2-4 Spreading Activation Algorithm

The "society", or agent network, is defined by three sets of links that connect the competence modules: (i) A *successor link* connects module x to module y (x has y as a successor) for every proposition p that is a member of the add list of x and also a member of the precondition list of y (so more than one successor link between two modules can exist), (ii) A *predecessor link* connects module x to module y (x has y as a predecessor) for every successor link that connects y to x, (iii) A *conflicter link* connects module x to module y for every proposition *p* that is a member of the precondition list of *x* and also a member of the delete list of *y*. Activation energy is spread through the network by the algorithm shown in Figure 3.2-4.

The spreading and injection of activation energy into the network is determined by the state of the environment, the goals of the agent, and the global parameters θ, π, γ, ϕ and δ.

The global parameters are: (i) θ, the threshold for becoming active, and related to it, π the mean level of activation, (ii) ϕ, the amount of activation energy a proposition that is observed to be true (sensor-based propositions are binary, i.e., either on or off) injects into the network, (iii) γ, the amount of activation energy a global goal injects into the network (the actual activation energy injected is γ multiplied by a real number representing the strength of the goal), (iv) δ, the amount of activation energy a protected goal takes away from the network.

These parameters determine the amount of activation a module spreads backwards to its *predecessors*, forwards to its *successors*, or takes away from its *conflicters*: (i) For each false proposition in its precondition list (sub-goal) a module spreads $\alpha_{(t-1)} \cdot \frac{1}{n}$ to its predecessors (where n is the number of predecessors times the number of goals of that particular predecessor – intuitively adjusting for the number of sources of activation energy a sub-goal has). (ii) For each false proposition in its add list an executable module spreads $\alpha_{(t-1)} \frac{\phi}{\gamma} \cdot \frac{1}{n}$ to its successors. (iii) For each true proposition in its precondition list an executable module takes $\alpha_{(t-1)} \frac{\delta}{\gamma} \cdot \frac{1}{n}$ from its conflicters (only if the module is more active than the modules it is trying to inhibit).

Agent Personality

By adjusting the values of the global parameters, an agent can be made more *adaptive*, *quick* and *reactive* on the one hand, or *thoughtful* and *rational* on the other hand.

Goal-orientedness is achieved through back propagation from global goals to modules. Each goal injects γ activation energy into the modules that achieve that global goal. These modules then pass $\alpha_{(t-1)} \cdot \frac{1}{n}$ activation energy on to each of their sub-goals, which in turn pass on $\alpha_{(t-1)} \cdot \frac{1}{n}$ to their sub-goals, and so on. The back propagation of activation energy ensures that modules that contribute to more than one goal, as well as those modules that are “closest” to the current goal, are favoured. By increasing the ratio of ϕ to γ, the relative contribution of the current situation to the activation energy of a module is increased. This translates into a bias towards modules whose preconditions partially match the current situation (allowing an agent to take advantage of opportunities when they arise) and the agent becomes more *situation-oriented*.

Thoughtfulness is achieved by setting a high initial threshold θ, or mean level of activation π. With a high threshold setting it becomes more likely that the algorithm will be allowed to run through the loop a number of times “considering” more possibilities as the network has longer to settle down on an optimum solution. Lowering the threshold has the opposite effect, making it more likely that a module is selected, and thus increasing the *speed* with which the agent makes decisions.

Analysis

The Agent Network Architecture (ANA) has a number of promising credentials: (i) it is robust and adaptable; (ii) it is capable of pursuing goal-orientated behaviour, whilst still responding to the opportunities and threats encountered in a dynamic and unpredictable environment; (iii) its behaviour can be modified by adjusting a small number of global variables; and (iv) it supports partial prioritising of goal strengths.

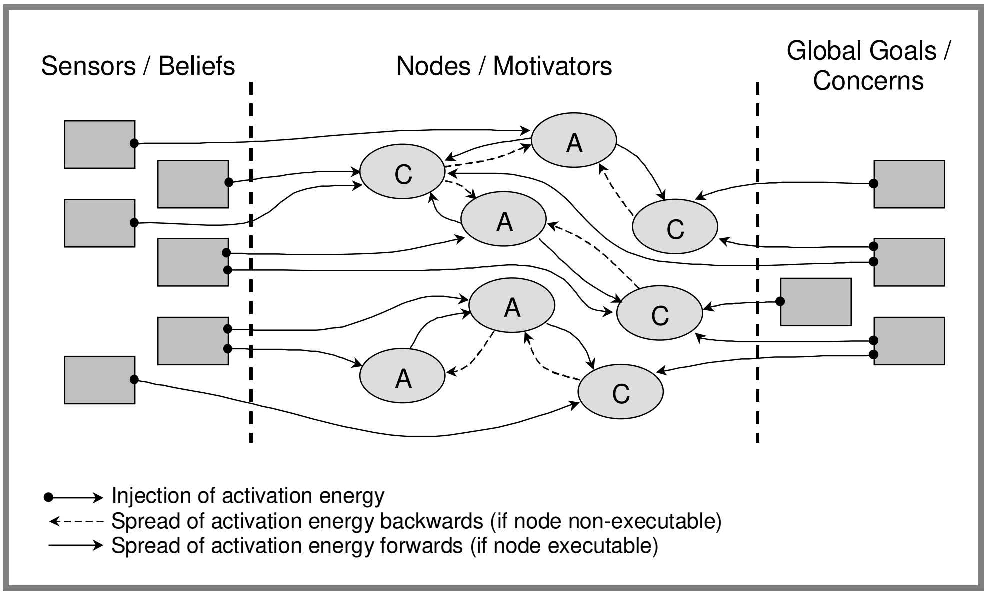


Figure 3.2-5 The Agent Network Architecture

We can think of the network of nodes of the ANA as representing the *motivators* of the agent (mechanisms that tend to produce, or modify, or select between actions in the light of beliefs [Sloman 97]), with global goals representing the *concerns* (situations the agent wants to bring about), and the sensors corresponding to agent *beliefs* (see Figure 3.2-5). We can also identify two types of node: (i) consummatory nodes (**C-nodes**) which provide an immediate direct benefit to the agent; and (ii) appetitive nodes (**A-nodes**) which require further expenditure of time and energy before any benefit is realised. Ideally we want our agent to proceed from appetitive nodes towards consummatory nodes (as any expenditure of energy due to appetitive behaviour will only be rewarded by consummatory activity). In this sense, the goal-orientated behaviour of the agent is represented by behaviour that leads to the activation of consummatory nodes. *Motivational attitude* is represented by the magnitude of activation energy associated with a node, and the *valence* by the sign.

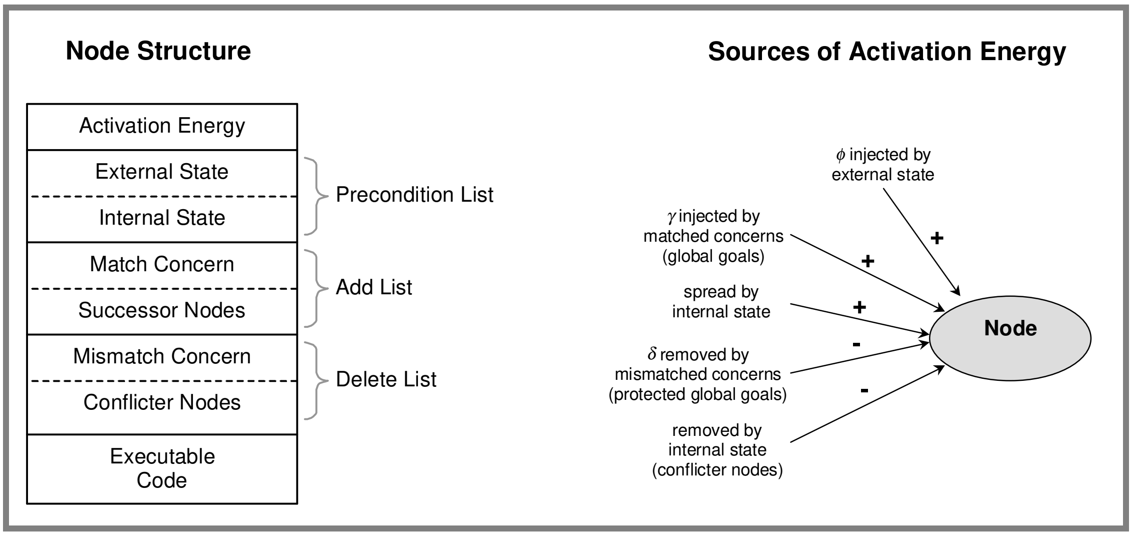


Figure 3.2-6 Node Structure and Sources of Activation Energy

Figure 3.2-6 shows the structure of an ANA node and a simplified view of its sources of activation energy. At a network level, state and goal orientated behaviour is mediated by the injection of energy from beliefs and concerns. However at the node level, the situation is complicated by the interaction between appetitive and consummatory nodes.

Forward spreading of activation energy from executable **A-Nodes** to successor **C-Nodes/A-Nodes** is intended to signify predictions of "effects that are about to become true" [Maes 89, page 10]. The intuitive idea was to prepare successor nodes for execution, and in some way mimic the injection of activation energy from external state sources. Unfortunately, by spreading and not injecting energy, it has the effect of reducing the activation energy of the executable **A-Nodes** themselves. This in turn reduces the likelihood of the executable **A-Nodes** being executed.

A second problem with the Agent Network Architecture is caused by association between preconditions and predecessor links. Figure 3.2-7 shows a simplified ANA plan structure, where the **A_x-Nodes** can be thought of as sub-goals of the **C_x-Nodes**. For our agent to be able to exhibit opportune behaviour (i.e., switching from node **A_{1b}** to **A_{2b}**, or node **A_{1d}** to **A_{1a}**) in response to a change of state, the new node must be executable before it can be activated. This means that all the *preconditions* for the new node must be satisfied, i.e., all the nodes below it in the hierarchy have been executed (not just executable) and have asserted the propositions in their add lists. The use of preconditions to determine the predecessor links as well as the conditions for execution, forces the network to activate nodes in fixed sequences. This has the undesirable effect of making the network less responsive to opportunities.

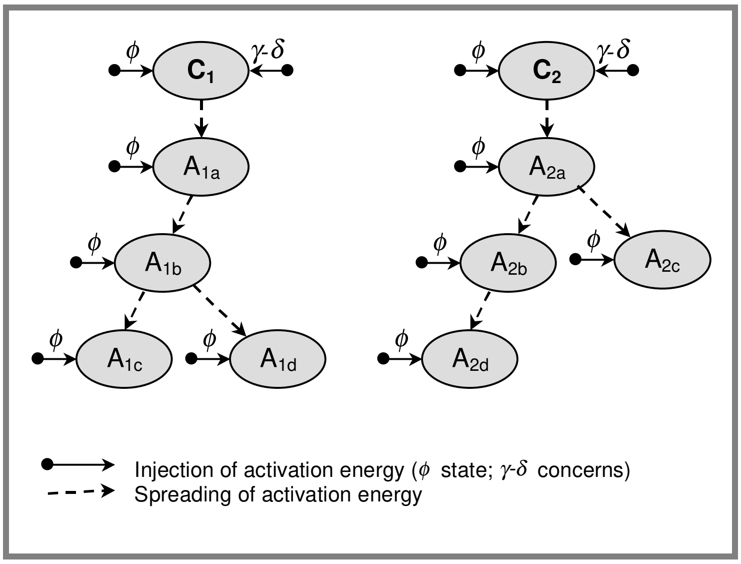


Figure 3.2-7 Agent Network Architecture Plan Structure

A related problem arises from the discharge of activation energy when a node is executed. By allowing activation energy to build up in nodes, the ANA can periodically execute low priority plans at the expense of their faster charging neighbours. However, because the execution of a single node eliminates the accumulated energy of a whole plan structure, there is little sustained goal directed behaviour towards low priority concerns.

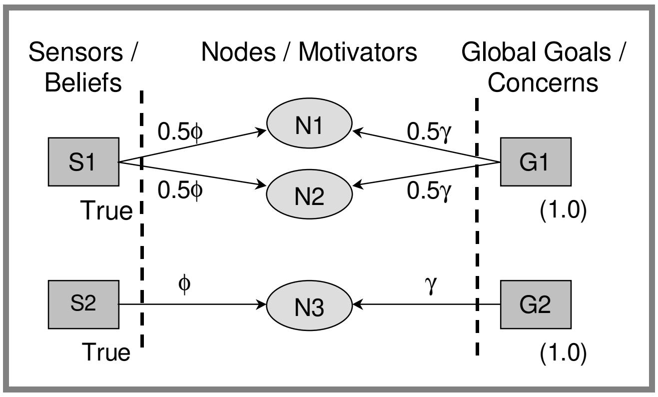


Figure 3.2-8 Unfair biasing of Nodes in ANA

A third area of concern about ANA has been investigated by Tyrrell [94]. No matter how you divide the distribution of energy, you will always end up unfairly biasing one particular arrangement of nodes or another. Figure 3.2-8 shows a simple case where a single sensor feeds two nodes. As the activation energy from sensor **S1** is divided by two to feed nodes **N1** and **N2**, nodes **N1** and **N2** are disadvantaged with respect to node **N3** (which gets a full complement of activation energy from **S2**). Similar problems occur when we look at concern activation sources.

Tyrrell performed a number of simulated experiments to test different arrangements of energy distribution. His conclusion was that no satisfactory set of division rules exists for the predecessor and conflicter connections, unless extra information is passed through the mechanism (see PHISH-Nets for a modified ANA [Rhodes 96]). The dilution of activation energy by nodes that contribute to the same goal was noted by Maes [89] in the original ANA design. However, Maes viewed this as a desirable feature of the network – favouring nodes with no competition. This does seem to somewhat contradict the desired aims of robustness and adaptability.

3.2.4 Free-Flow Hierarchy

Tyrrell proposes a free-flow hierarchical solution to the action selection problem. Although the architecture is roughly modelled on the Rosenblatt and Payton [89] (see section 3.2.2) fine-grained connectionist architecture (FGCA), there are a number of very significant differences that are important to note. Whereas the FGCA is a vertical subsumption style architecture (with each level of the architecture adding a complete layer of competence to the level below), Tyrrell's architecture adopts a hierarchical approach in which the drive-based activation energy is only injected at the top of the hierarchy – violating the subsumption philosophy as the design can no longer be partitioned such that the lower layer forms a complete working system. The Free-Flow hierarchy architecture (FFHA) also relaxes the constraint on the magnitude of activation energy that can be generated by a node (FGCA only allows values in the range of ±1.00), and confuses the terminology of nodes with that of behaviour networks. Finally, there is no attempt to use the selective update mechanism of the FGCA in Tyrrell's design, and there is no real sharing of information at a behaviour level. In the following discussion we will therefore treat the Free-Flow hierarchy as a new architecture, and not an extension to the FGCA.

A free-flow of command information is achieved by the fact that: (a) the architecture places no restrictions on the number of nodes that remain active at any one time; (b) open competition is encouraged by prohibiting inhibition and excitation of behaviours across systems (preventing one system from completely shutting down a competing system with the resultant loss of information); and (c) decisions are always deferred to the final behavioural layer (the leaves of the hierarchy), where a winner-takes-all action selection scheme is implemented. This ensures that all drive-based behavioural systems can have their say in the selection process. Figure 3.2-9 shows the "Get Food" behavioural system of Tyrrell's animat action selection architecture.

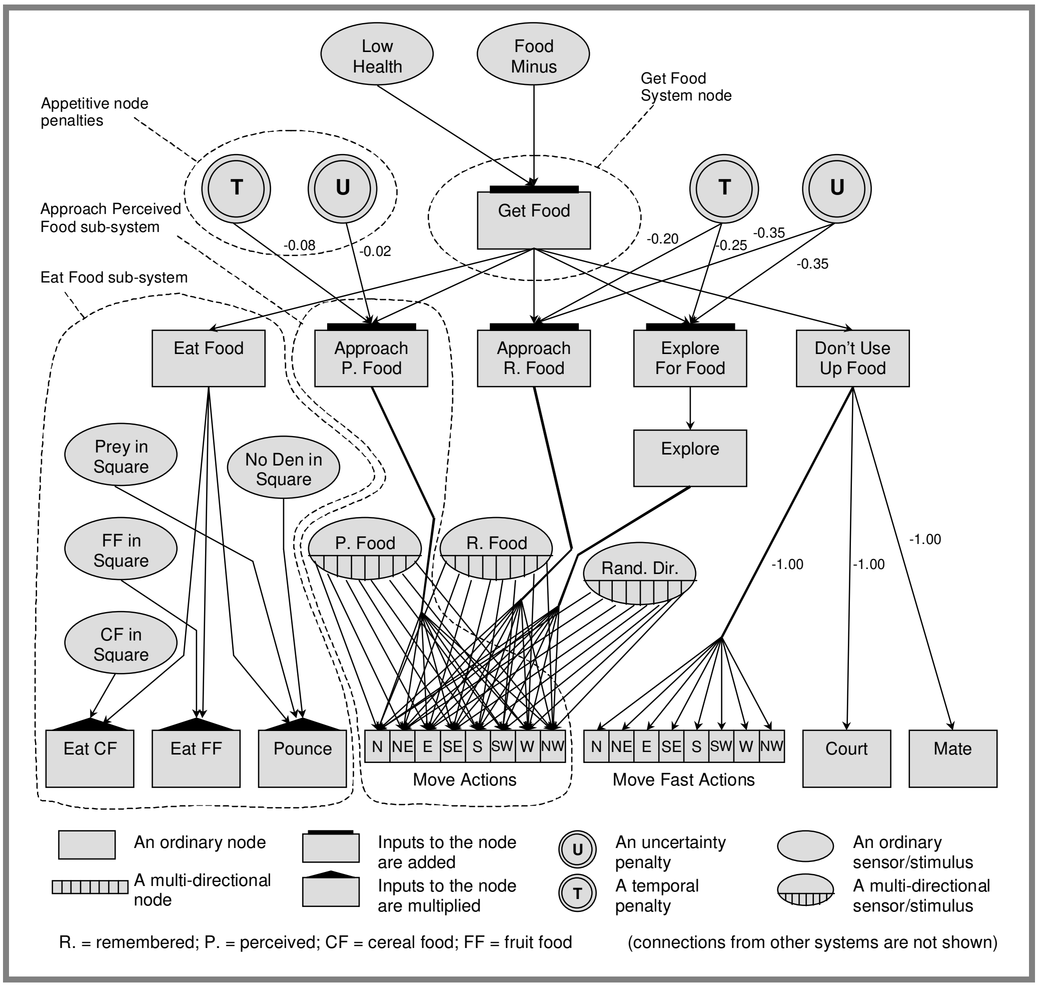


Figure 3.2-9 Get Food hierarchy [Tyrrell 93a, pages 161 and 166]

Combining Preferences

The FFHA adopts two basic forms of activation energy combination schemes: (i) addition; and (ii) multiplication. The multiplication scheme tends to be used in the final behavioural layer to allow preconditions to be taken into account before executing actions, i.e., ensuring that prey is present in the square before pouncing, or inhibiting a pouncing action when the animat is in its den. At higher levels of the hierarchy, an addition scheme is used. Simple addition was found to suffer from the same problems as those encountered by the Agent Network Architecture [Maes 89] (see section 3.2.3), i.e., unfair balancing of: (i) preferences originating from the same goal; with (ii) preferences from different appetitive nodes. The FFHA implements a scheme which combines simple summation of inputs with taking the maximum valued input. The final algorithm is shown in Figure 3.2-10. It should be noted that Tyrrell feels that “it may prove possible to develop a more principled scheme” [Tyrrell 93a, page 193].

$$A_j = \left(\frac{\max_i \left(P_{ij}^{+}\right) + \alpha \sum_{i=1}^{N+} \left(P_{ij}^{+}\right)}{1+\alpha} \right) + \left(\frac{\max_i \left(P_{ij}^{-}\right) + \beta \sum_{i=1}^{N-} \left(P_{ij}^{-}\right)}{1+\beta} \right)$$

where P_{ij}^{+} and P_{ij}^{-} are the positive and negative preferences from node i for node j

where N^{+} and N^{-} are the numbers of such preferences for node j

and α and β are constants, where $\alpha = \frac{1}{\left(N^{+}\right)^2}$ and $\beta = \frac{1}{\left(N^{-}\right)^2}$

Figure 3.2-10 Combining Preferences in FFHA [Tyrrell 93a, page 193]

Analysis

Tyrrell adopts a penalty scheme to favour consummatory nodes over appetitive nodes. Temporal penalties inhibit a node proportional to the length of time for the likely achievement of the goal, and uncertainty penalties inhibit a node proportional to the likelihood that the behaviour will eventually lead to the completion of the goal. By varying the magnitude of the penalties (temporal and uncertainty), the model can also favour appetitive nodes that are closer to consummatory nodes over those further away. The main drawback with such a scheme (as implemented) is its reliance on absolute numbers. Penalties need to reflect the relative costs of time and uncertainly to behaviours both within a system, as well as across systems. As the injection of activation energy is context sensitive, penalties can at best only represent relative costs in typical situations. If the injection of energy is large (compared to the penalties) then the penalties will have little influence, and likewise if the total injection of energy is low then the penalties could prevent all appetitive nodes from being selected.

Tyrrell's architecture performed well in its simulated environment niche, but nevertheless gives rise to serious concerns as to its scalability to more demanding environments. Tyrrell's central tenet of a free-flow of information forces action selection to occur only at the consummatory/actuator nodes. This leads to a persistence problem and an explosion in complexity as the number of checks and balances between different leaf nodes multiplies. Some attempt to constrain complexity is made by using a system of drives to add a hierarchical structure to the concern-processing mechanisms, however behaviours usually satisfy a multiplicity of drives (see section 5.2.2) which cannot be support within such a hierarchical structure. We will leave the FFHA with a list of the requirements for a behaviour selection system (Figure 3.2-11).

1) **Dealing with all types of sub-problems:** the need to be able to handle all of the various types of sub-problem successfully.
2) **Persistence:** the need to have a tendency to persist with a consummatory action beyond the time that the deficit it is reducing is strictly the most important, because of the cost of changing to another system.
3) **Activations proportional to current offsets:** in homeostatic systems, the need for the node activations or drive strengths to be in proportion to the current offsets from the optimum points.
4) **Consummatory over appetitive actions in the same system:** the need to have a tendency to prefer consummatory over appetitive actions in the same system, if both are equally relevant to the current external situation.
5) **Consummatory over appetitive actions in the other systems:** the need to have a tendency to prefer a consummatory action in one system over appetitive actions in other systems, all other things being equal.
6) **Balanced Competition:** the need for there to be no discrimination against nodes which help to achieve more than one goal, or against nodes which receive input from many different stimuli, or against nodes which receive input from only one stimulus or from none at all.
7) **Contiguous action sequences:** there is a need to have a tendency towards continuing the current sequence once started, rather than beginning a new sequence for a different system.
8) **Interrupts if necessary:** the need to be able to interrupt a sequence of actions for a relatively low-priority system if another more urgent system places a high-priority demand on the use of the animals actions.
9) **Opportunism:** the need to incorporate external stimulus, as well as deficit or motivational information, when calculating the desirability of different alternatives.
10) **No system-level winner-take-all:** because of the needs for preservation of information and choice of compromise candidates, a mechanism should not 'shut down' all but one system.
11) **Combination of preferences:** the need to be able to integrate multiple non-binary preferences from higher-level nodes when deciding the responses of lower-level nodes.
12) **Compromise candidates:** the need to be able to choose actions that, while not the best choice for any one sub-system alone, are best when all sub-problems are considered simultaneously.
13) **Real-valued sensors:** the need to extract the full amount of information from the environment and internal state.
14) **Flexible combination of stimuli:** the need to allow arbitrary functions for combining stimulus values.

Figure 3.2-11 Requirements of a Behaviour Selection System [Tyrrell 93a, pages 173-174]

3.2.5 Inhibition and Fatigue

Blumberg's [94] action selection algorithm is based on an ethological model first proposed by Ludlow [Ludlow 76; Ludlow 80]. Ludlow noted that if (i) activities are mutually inhibiting, (ii) the inhibitory gains (activity i inhibits activity j by an amount equal to activity i's value multiplied by an inhibitory gain k_{ij}) are greater than 1, and (iii) the values of competing activities are restricted to being zero or greater, then the model will result in a *winner-takes-all* algorithm.

The operation of the action-selection algorithm:

1) Activities are represented by the nodes of a tree, and are organised in loose overlapping hierarchies with more general activities at the top and the more specific activities at the leaves.
2) Nodes can have 0 or more children. All children are mutually inhibiting so that only one child can be active at a time. If the active node is a leaf node it can issue commands to the motor actuators, otherwise its children must compete for control, until a leaf node is reached.
3) Nodes compete on the basis of their value. A node's value is calculated at the start of each time step, and is a function of (i) the amount received from *endogenous variables* (internal sources of motivation), (ii) the amount added by *releasing mechanisms* (external sources of motivation), (iii) the amount removed by mutual inhibiting mechanisms, (iv) its current level of fatigue, and (v) its rate of change of fatigue (negative when inactive, and positive when active). The system then iterates until a stable solution is found, where one node has a positive value and the remaining nodes have a value within a tolerance of zero.
4) Losing nodes (in the action selection process) can post recommendations to the winning node. The winning node is then free to implement these recommendations or ignore them at its discretion.

Figure 3.2-12 Hamsterdam Action Selection Algorithm

Temporal Aspects of Behaviour

One of the problems that action-selection algorithms must address is the persistence problem: "… [the] difficult[y] to control the temporal aspects of behaviour so as to arrive at the right balance between too little persistence, resulting in dithering among activities, and too much persistence so that opportunities are missed or that the animat mindlessly pursues a given goal to the detriment of other goals." [Blumberg 94]

The Blumberg/Ludlow model addresses the persistence problem in three ways: (i) inhibitory gain factors set a general persistence level for each activity, (ii) behaviour-specific fatigue mechanisms reduce an activity's persistence level over time, giving lower priority activities a chance of interrupting high persistence activities, and (iii) releasing mechanisms allow the animat to take advantage of opportunities offered by the environment.

Inhibitory Gain Factors. Inhibitory gain factors above 1.0 add hysteresis to the action selection process, resulting in a winner-takes-all architecture. In general the higher the inhibitory gain, the higher the persistence of the activities.

Fatigue. To stop high-persistence activities dominating the action-selection process, an activity-specific fatigue mechanism is included. All activities are assigned a level of fatigue, which increases when the activity is active and decays towards zero when the activity is no longer active.

Releasing Mechanisms. Releasing mechanisms are processes that are associated with significant objects and events in the animat's world. These processes add value to an activity, allowing an animat to take advantage of opportunities offered by its environment. Opportunism works as follows: if a releasing mechanism is assigned to an object like water, an animat who happens to wander by a lake in search of food is likely to switch from the ongoing “find food” activity to the “find water” activity, taking advantage of the source of water.

Loose Hierarchical Structure

Blumberg’s model uses a hierarchical activity structure, but allows information sharing between activities. Nodes that take part in the selection competition, but fail to get selected, can post suggestions to the winning activity. The main arguments for this arrangement are that (i) it maintains the organisational advantages of a hierarchical structure, (ii) a loose hierarchy provides a focus of attention, allowing those nodes involved in the selection competition to post recommendations, and (iii) free-flow architectures, where all nodes express their preferences to the winning motor/leaf node, are very sensitive to the type of algorithm used to process the preferences.

Analysis

An animat must respond to the strength of its stimuli and not simply the stimuli’s Boolean value, i.e. whether it is there or not. This requirement makes it likely that an architecture adopts a common currency, and uses that currency to combine and compare the strengths of internal and external motivators. Blumberg’s model adopts a common currency of activation value, and uses this to express the results of endogenous variables, releasing mechanisms, and fatigue mechanisms.

A more interesting feature of the architecture is the recognition of the importance that the accounting system plays (whether to sum, subtract or multiply the input values), and the need for a flexible modelling system. Endogenous variables and releasing mechanisms can be arbitrarily complex calculations, relying on previous history if need be. Values are also subjected to maximum ranges, which can be adjusted to alter the persistence or reactivity of the agent. By increasing the possible range of values produced by releasing mechanisms, relative to the range of endogenous variables, an animat can be made more reactive to the external environment.

Blumberg’s inhibition and fatigue model shares many features with Mae’s spreading activation approach: (i) the architecture is event-driven, with events spreading activation energy between the different behaviour modules, (ii) there are no bureaucratic control modules. One of the main disadvantages of this type of architecture is the complex web of checks and balances needed between the different behaviour modules. Adding a new behaviour is no simple task, as an author must hand-code all the different situations in which the action can be used. It could be argued that nature can afford to experiment, and that the many designs that do not get the balance right are simply discarded in the course of natural selection.

3.2.6 Conclusions

As stated in section 1.3, an autonomous agent must be capable of: (i) handling multiple sources of motivation with limited resources; (ii) having and pursuing an agenda; and (iii) being robust and adaptable in the face of a hostile and uncertain environment.

The behaviour-based architectures looked at in this section tackle these problems in two ways: (a) through vertical decomposition along the lines of competencies, and (b) through biologically inspired horizontal decomposition along the lines of systems or drives. Although both of these approaches have their relative merits, they only go so far in addressing the requirements of intelligent autonomous agency.

Frameworks such as the Subsumption Architecture [Brooks 86] and Adaptive Hierarchical Control [Kaelbling and Rosenschein 91] arbitrate among behaviours by explicitly or implicitly assigning priorities to each behaviour. Such priority-based schemes are effective at choosing between incompatible commands based on local knowledge, but they do not provide an adequate means for dealing with multiple goals (sources of motivation) that can and should be satisfied simultaneously [Rosenblatt and Thorpe 95]. Whenever one behaviour overrides another in a priority-based scheme, all information contained within the overridden behaviour is lost to the system – prohibiting any chance of compromise. However, the hierarchical nature of such systems allows the agent to easily integrate very sophisticated competencies and pursue complex agendas, whilst still maintaining the robustness and adaptability of a reactive behaviour-based system.

Free-flow/Command Fusion architectures avoid the information loss problem by allowing all nodes to express their preferences to the winning motor/leaf node. However, they are very sensitive to the type of algorithm used to process the preferences, as well as being susceptible to persistence and scaling problems. Some progress towards solving the persistence and scaling problems is demonstrated by Blumberg's use of a fatigue mechanism and a primitive attention-like mechanism to select a winning system. The winning system is then free to adopt or ignore the recommendations of other systems.

3.3 Summary

In this chapter we have analysed some of the basic concern-processing mechanisms active in the deliberative and reactive levels of intelligent agent architectures. We identified a number of problems purely deliberative agent architectures have when faced with the complex real-time requirements of intelligent autonomous agency, and argued that the traditional solution of partitioning a design along functional grounds (assigning the "intelligence" to a deliberative layer that plays a supervisory/planning role in the final agent architecture) fails to take into account emergent states created by the interaction of processes operating between layers – thus leading to reduced predictability and correctness of the design.

We also looked at the behaviour-based approach to the requirements of intelligent autonomous agency. Unfortunately, although behaviour-based architectures easily meet the real-time constraints of intelligent autonomous agency (and the more general requirements of adaptability and robustness), they tend to suffer from problems of scalability and complexity of design. These problems are systematic of the fact that: (a) behaviour-based architectures focus on a single concern-processing mechanism (subsumption or spreading activation); and (b) concerns are treated as near static quantities external to the action selection solution.

In chapter 7 we will describe the design for a *Society of Mind* [Minsky 85] agent architecture that addresses many of the problems raised in this chapter – by partitioning the design along the lines of concern-processing competence levels. But first, we will expound the merits of a concern-centric design stance a little more with a closer look at the role motivational control states play in intelligent autonomous agent architectures.

4 Motivational Control States

This chapter presents a preliminary design specification for a working system to fulfil the requirements analysis laid out in section 1.3. The design is recursive in nature, replicating threads 1-5 of the design-based methodology at different levels of abstraction. This initial design process draws heavily on ideas and concepts developed by members of the Cognition and Affect Project.

In this chapter, we will extend our motivated agent framework to include a definition of affective states which covers the emergent perturbant states characteristic of certain types of *emotion* – i.e. those emotions characterised by a difficulty to control attentive thought processes such as grief, longing, and infatuation. A proto-"emotional" architecture [Wright 97] (which includes a shallow meta-management layer capable of detecting perturbant states arising out of the active management of motivators) is described, and a number of problems associated with shallow motivator management and a purely pre-attentive perceptual system are identified.

4.1 Functional Attributes of Motivators

We use the term *motivator* to refer to motivational control states that move an agent towards a desired physical/mental state in light of agent *beliefs* and *concerns* – i.e. a subclass of information structures with dispositional powers to determine action (both internal and external), and which subsumes *desires*, *goals*, *intentions*, and *wishes* [Wright 97, page 57].

In moving the agent towards a desired physical/mental state, motivational control states perform three functions [Cañamero 97; Kandel et al. 95]: (i) a *directing* function – they steer behaviour towards satisfying a particular concern; (ii) an *activating* function – they animate the agent into action, and (iii) an *organising* function – they combine individual behaviours into a coherent, goal-orientated response. The following discussion situates these functional attributes within the motivated agent framework described in the chapter 2, and Sloman's Attention Filter Penetration theory.

Attention Filter Penetration Theory

In fast changing environments, it is essential that autonomous agents are able to react quickly to unexpected and/or dangerous events. In order to achieve the necessary fast response times within a resource-limited deliberative architecture, a mechanism must be provided to allow context switching and the interruption of ongoing processing [Simon 67] (also section 5.2.2). Attention Filter Penetration (AFP) theory [Sloman and Croucher 81; Sloman 92] extends Simon's basic ideas in a number of ways: (i) the addition of a *variable* filter to protect urgent, important, or resource consuming management processes from irrelevant distraction, (ii) the requirement that insistence level (the propensity of a motivator to penetrate the filter) is assigned by relatively simple heuristics, (iii) the introduction of reactive, management and meta-management control systems, and (iv) the introduction of the term *perturbance*, or loss of control of one's own mental processes.

Directing Function of Motivators

AFP theory is primarily concerned with the architectural constraints placed on intelligent control systems attempting to support multiple motives in a rapidly changing hostile environment. Real-time performance is achieved by reactive motivator generactivators (generators/re-activators) which interrupt ongoing management processes when new events match agent concerns – see Figure 4.1-1 below. In this sense, concerns can be defined as "*motivational attitudes*, serving as *standards* against which situations are tested for compliance or non-compliance with desired norms". By matching events against concerns, motivators are able to *direct* behaviour towards satisfying those concerns.

Activating Function of Motivators

As motivators are generated by reactive processes largely ignorant of the 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 4.1-1.

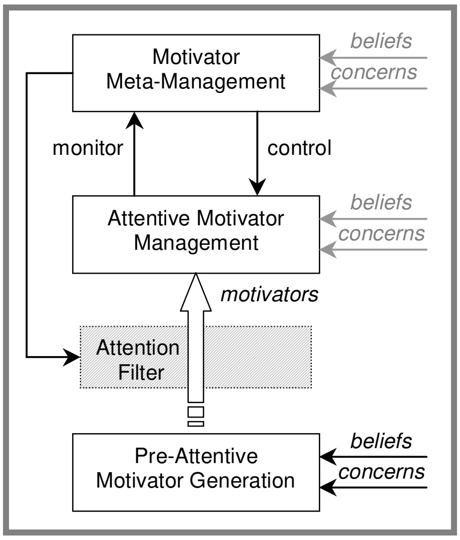


Figure 4.1-1 The role of the Attention Filter

The ability of reactive motivators to interrupt attentive processes depends on the interaction of two separate mechanisms: (i) reactive insistence assignment – representing the urgency/importance of the new reactive motivator, and (ii) 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: (i) the number of motivators already under management consideration; (ii) the assessed urgency/importance of the current management process; and (iii) the ability of management processes to cope with the current situation.

The overall "fitness" of the agent can be maximised by making both the reactive insistence and filter threshold assignments operate in intelligent context sensitive ways. It is important to view these two mechanisms as co-operative and not antagonistic – the survival value of the filter comes from its ability to harness local meta-management knowledge (in the form of the filter threshold) when reacting to life threatening situations.

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 4.1-2). 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*. It is therefore possible for a motivator to temporarily distract attention without changing ongoing plans and actions.

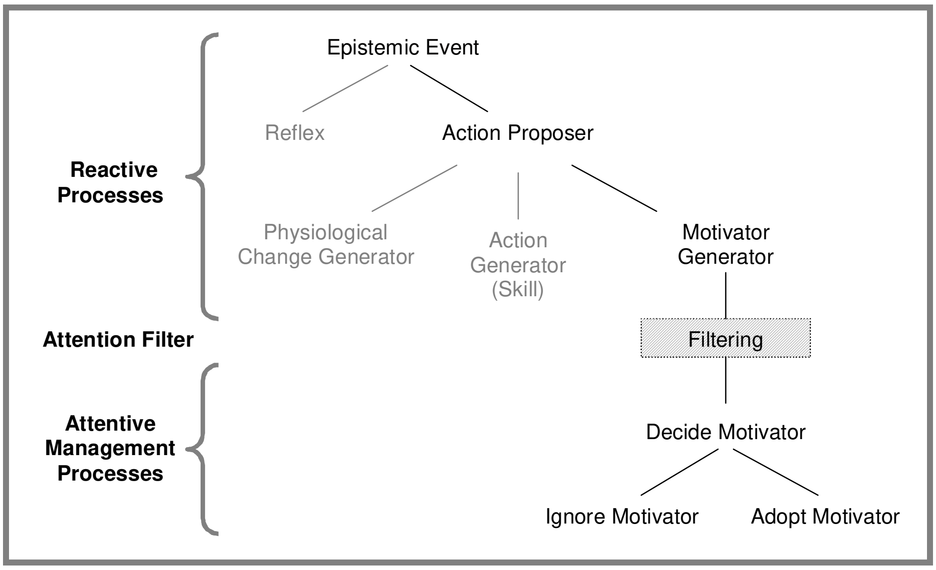


Figure 4.1-2 Motivator Adoption [extended Beaudoin 94, page 84]

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 4.1-1). By creating a profile where $W_{aff} > W_{ach} > W_{learn}$, 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_{phys}
Safety	Feeling Threatened	W_{safe}
Affiliation	Safety of Other	W_{aff}
Achievement	Achieving own Goals	W_{ach}
Learning	User Requirements	W_{learn}

Table 4.1-1 Motivational Profile

4.2 Case Study: A Motivated Agent

In section 3.1.1 we introduced the Procedural Reasoning System (PRS) [Georgeff and Lansky 86; Georgeff and Ingrand 89], and highlighted a number of problems the classic PRS possess for intelligent autonomous agency. By concentrating on the requirements of goal-processing, the NML1 architecture is able to address some of the deficiencies of PRS and give a preliminary classification of the attributes of goals – an important subclass of motivational control states.

NML1 is a broad agent architecture developed to elucidate goal-processing in autonomous agents [Beaudoin 94]. The NML1 design uses the architectural framework laid out in section 2.2, and a requirements specification similar to that of section 1.3. Early prototypes of NML1 were implemented by Luc Beaudoin and Ian Wright of the Cognition and Affect Project at Birmingham University. These partial implementations were then used to refine the design processes in the iterative tradition of the design-based methodology. A simplified view (not all the data and control paths are shown) of the NML1 architecture is shown in Figure 4.2-1.

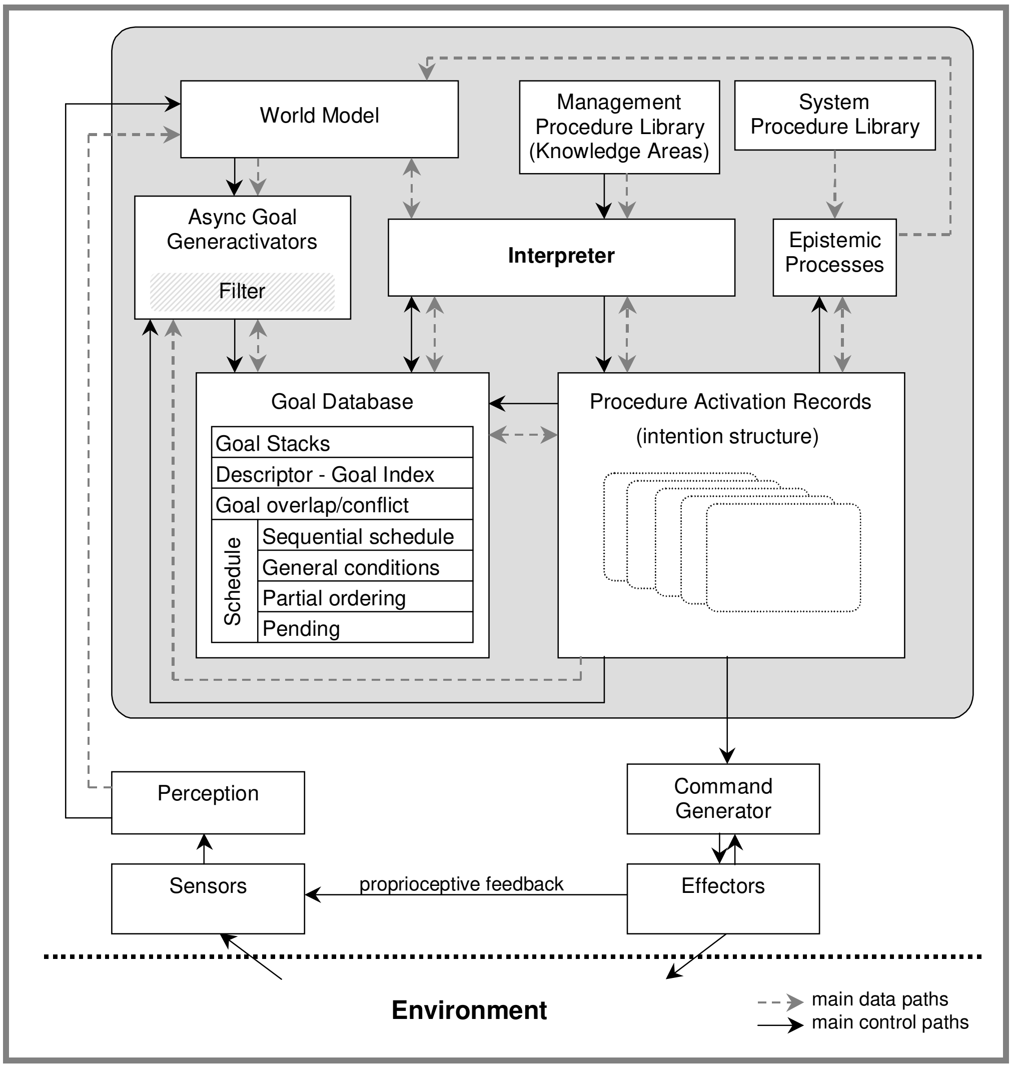


Figure 4.2-1 Simplified view of the NML1 architecture

Goal Generation

There are two sources of goal generation in NML1: (i) asynchronous goal generactivators, responding to facts in the agent database; and (ii) synchronous active management procedures (m-procedures), generating sub-goals in pursuit of an already established goal. The asynchronous goal generactivators represent a marked departure from the strategy employed by PRS. In PRS the knowledge area (KA) unification and reflection cycle of the PRS interpreter is carried out over all new events (goals and beliefs). As there are potentially many new beliefs, and many entries in the KA plan library to search, this reflection process can occupy a significant proportion of the interpreter's cycle time. NML1 reduces this overhead by asynchronously generating belief-based goals. Goal arbitration (normally carried out by meta-level KA reflection in PRS) is accomplished by filtering pre-management goals before they are inserted into the goal database. Event- and goal-driven behaviour is mediated by the type of meta-goal generated to manage the new surfaced goal (management meta-goals are generated when a goal surfaces through the filter, and placed at the top of the new goal's goal-stack in the goal database – see Goal Management below).

Insistence Assignment

Insistence is defined as the propensity for a goal to penetrate the filter. Each goal generactivator assigns an insistence value to the goal it generates using a quick and easy to compute heuristic function. For example, the insistence value of a goal to "move a baby away from a ditch" could be a function of the closeness of the baby to the ditch. It is feasible that more than one goal generactivator generates the same goal at the same time (for different reasons and therefore using different insistence assignment heuristics). In this case, the maximum suggested insistence value is taken as the insistence value for the goal. The insistence value represents an approximate measure of the importance or urgency of the goal. Over time goal insistence values decay towards zero. When a goal's insistence value reaches zero it is removed from the pre-management goal list.

Goal Filter

The NML1 goal filter applies a winner-takes-all strategy to the collection of pre-management goals, allowing at most one to surface at a time. The goal filter has three independently variable components: (i) a global threshold; (ii) a set of idiosyncratic thresholds (goal-descriptor/threshold pairs); and (iii) a management efficacy parameter. The idiosyncratic filter thresholds are used by management m-procedures to selectively increase or decrease the filter threshold for individual goals. The sensitivity of the system to such management control is determined by the management efficacy parameter. When filter threshold values are written they are multiplied by the efficacy parameter, this means that setting the efficacy parameter to zero ensures that no idiosyncratic control is possible. Idiosyncratic filters exist for a fixed number of cycles before being automatically deleted.

Although goal generation and filtering are asynchronous to deliberation, the actual goal filtering in NML1 is performed according to a three-stage synchronous algorithm: (i) all candidate goals and their filter thresholds are sampled in parallel; (ii) insistence levels are compared against filter thresholds (goal descriptors that do not unify with idiosyncratic goal-descriptors use the global threshold value); and (iii) if more than one goal has an insistence level above its filter threshold, the most insistent goal is chosen (irrespective of the relative levels of their corresponding filter thresholds).

Goal Management

The activation state of a goal that surfaces through the filter is set to "asynchronously surfacing" (see Table 4.2-1). If the goal is not already present in the goal database, a new stack is created for it and a meta-goal is pushed on to the top of the stack to "manage" the surfaced goal. If a goal-stack already exists, and its associated m-procedure is suspended, the m-procedure will be re-activated. Thus goals can be asynchronously generated, or reactivated, by the Goal Generactivators.

Goal Activation State	**Explanation of State**
G	Current agent goals
$G_{filtering\text{-}candidate}$	Unsurfaced goals that are about to be, or are being, filtered
$G_{asynchonously\text{-}surfacing}$	Surfacing goals – asynchronously generated and filtered
$G_{synchronously\text{-}surfacing}$	Surfacing goals – synchronously generated by m-procedures
$G_{suppressed}$	Goals suppressed by a filtering process
$G_{actively\text{-}managed}$	Actively managed goals – focal object of an executing m-procedure
$G_{inactively\text{-}managed}$	Inactively managed goals – focal object of a suspended m-procedure
$G_{managed}$	Managed (actively or inactively) goals
G_{off}	If none of the above apply

Table 4.2-1 NML1 Goal Activation States

Analysis

NML1 claims to extend PRS in a number of ways: (i) asynchronous goal generation; (ii) multiple goal attention filters; (iii) demon-like system procedures (s-procedures) that run independently of the PRS interpreter; and (iv) a richer representation of goals. These claims need qualifying.

Asynchronous Goal Generation: The introduction of asynchronous goal generation goes a long way to reduce the processing overhead of the interpreter cycle. Asynchronous motivator generation mechanisms can actively monitor the state of the world looking for complex triggering conditions that would be impossible to capture with simple KA descriptor unification. However, the generation of fully fledged goals does pose a problem. The original NML1 architecture (Figure 4.2-1) requires asynchronous access of the goal structure by the filter mechanism. This would potentially interfere with goal maintenance activities of the interpreter, which PRS avoids by synchronously accessing the goal structure as part of the **Execute()** step. There is also the question of goal conflicts, scheduling, and deciding – all activities that should be performed before a new goal is added to the goal database, and activities which fall within the realms of reasoning and hence should involve the interpreter and meta-knowledge areas.

An alternative solution would be to insert the surfaced motivator into the **World Model** as a motivationally charged new event. The interpreter can then process the event in the normal way, by unifying the event against KAs or m-procedures. The motivational charge may be no more than a *valenced* insistence value. This would keep the size of the procedural database to a minimum by unifying m-procedures against the *valence* and *intensity* of the motivator (translating into **Achieve(…)/Achieve(Not(…))** goal statements, and a heuristic priority measure of the goal). Much of the NML1 goal database structure should also be treated as facts about the internal state of the agent and held in the **World Model** as meta-facts. This will allow the PRS to respond to goal conflicts in the same way as it responds to any other event in its environment. The modification to the NML1 architecture (dubbed NML2) is shown in Figure 4.2-2.

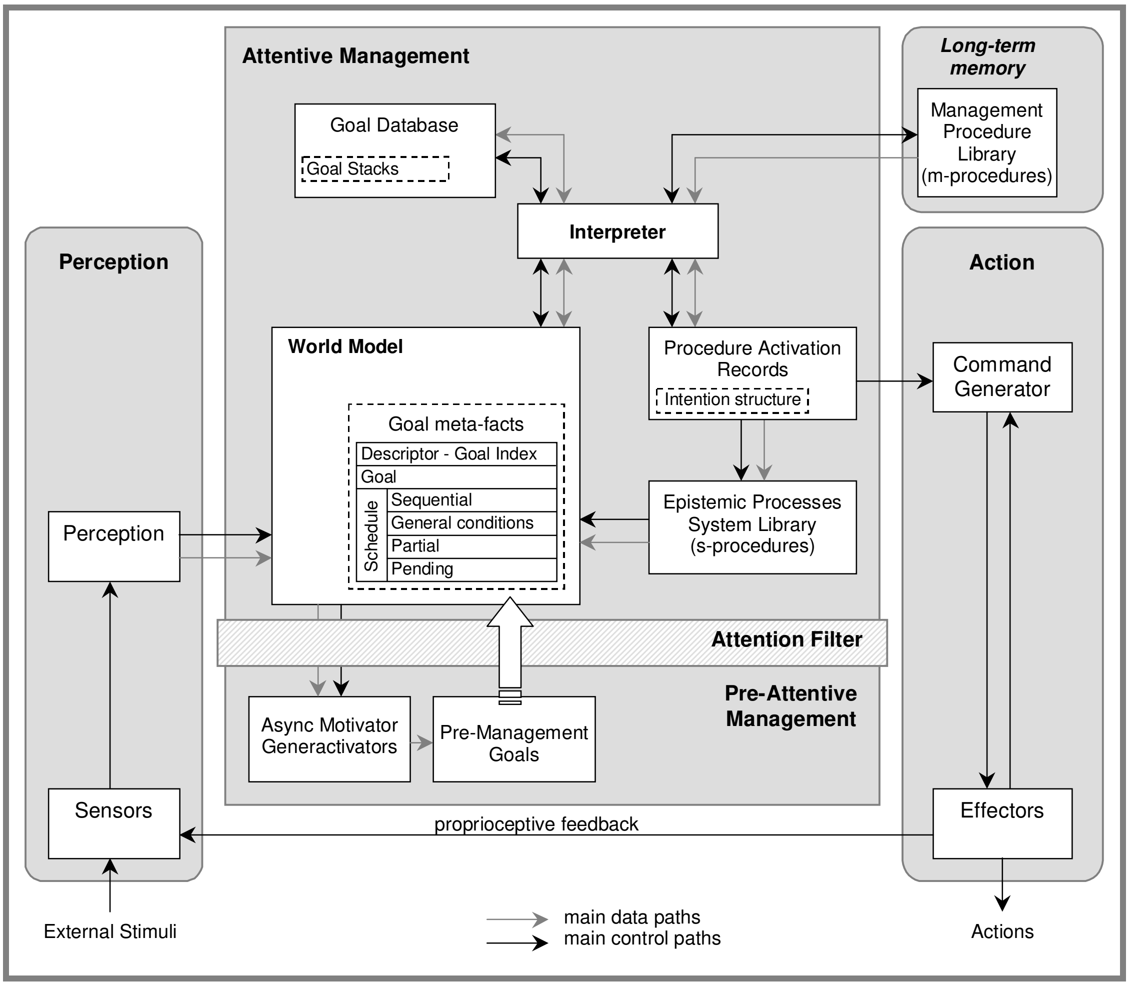


Figure 4.2-2 NML2 as a layered Procedural Reasoning System

Goal Attention Filters: The NML1 goal filter applies a winner-takes-all strategy to the collection of pre-management goals. This means that at most, only one goal will be allowed to surface at a time. Beaudoin's main argument for the winner-takes-all strategy was that it aided stability (presumably by constraining the rate at which new goals were considered by management processes). If the filter threshold remains at the same level, pre-management goals that would have penetrated the filter on the first cycle, but were prevented by the winner-takes-all arbitration, will now surface one by one over successive cycles. Deciding each goal in turn would require more management resources than allowing them all to surface together and be decided in one go. Beaudoin suggests that a refractory period might be employed, which temporarily raises the filter threshold after a goal has surfaced, to slow down the rate of generation of new goals.

MINDER1 (see section 4.4, and [Wright 97]) abandons the winner-takes-all philosophy, allowing all motivators with insistence values equal to or greater than the filter threshold to surface. Unfortunately, MINDER1 also attempts to fix the number of motivators surfaced at any one time by adapting the filter threshold to ensure an “optimum” number of motivators have surfaced. This results in oscillations of the filter threshold when two or more motivators with the same numeric insistence level surface and submerge together.

A different approach has been adopted by ALARMS [Norman 96] to allow all motivators with insistence values equal to or greater than the filter threshold to surface. But instead of using the filter to constrain the number of motivators under consideration, the filter threshold is used to regulate the rate at which motivators surface. All surfaced motivators are *decided*, and actively mitigated if they are not deemed urgent or important enough for adoption at that particular moment in time. Mitigated motivators then build-up their insistence levels at different rates (proportional to their urgency/importance) and surface again at different times in the future.

S-Procedures: NML1’s demon like s-procedures can be thought of as special purpose monitors, discharging some of the maintenance responsibility of the interpreter. S-procedures reside in their own database and perform updates to the system’s database (such as setting flags). Associated with each s-procedure is an activation condition, deactivation condition and activation mechanism. S-procedures cannot use shared resources and so cannot manipulate the goal database or intention structure directly. In this sense they are not as powerful as m-procedures in their monitoring role (PRS KAs usually establish sub-goals to monitor their own execution). S-procedures are, however, very responsive and can achieve similar results by setting flags that unify with m-procedures during the reflective interpreter cycle.

Attributes of Goals

Richer Representation of Goals: The final area in which NML1 contributes to the PRS architecture is through an analysis of goal-processing. This has led to a richer representation of goals in NML1, and the proposal of a structured goal database (retained as meta-facts in NML2). Beaudoin’s contribution should be seen as a framework for using PRS in complex goal-processing environments. The attributes of goals are summarised in Table 4.2-2.

Attribute Type	Attribute Name and Description
Essence	*Goal Descriptor*: The propositional aspect of the goal, often written in predicate calculus notation – i.e. charged(babyA)
Support	*Belief*: beliefs about components of the goal's descriptors, along with probabilistic statements about the certainty of the beliefs.
Assessment	*Importance*: represents the costs and benefits of satisfying or failing to satisfy the goal. *Urgency*: represents the temporal information about the costs, benefits and probability of achieving the goal. *Insistence*: a heuristic measure of the importance and urgency of the goal. *Dynamic state*: such as "being considered" or "plan suspended".
Decision	*Commitment status*: such as "adopted" or "rejected". *Plan*: a set of plans for achieving the goal. *Schedule*: denoting when the goal is to be executed or considered. *Intensity*: a measure of the strength of the disposition to act on the goal.

Table 4.2-2 Attributes of Goals

Conclusions

In this section we have described a design-level specification for an intelligent agent architecture (NML2) that meets the functional requirements of motivational control states – i.e. to *direct*, *activate*, and *organise* behaviour towards satisfying agent *concerns* – and the broader requirements of intelligent autonomous agency (see section 1.3). By adopting an information-level design stance we have identified the need for a richer set of *goal* attributes, to support intelligent agency, than is normally considered necessary for rational agency. Finally, by adopting the motivated agent framework of section 2.2 – rather than a functional partitioning on the specific attributes we wish to bestow on our agent – we are able to relate architectural features to information-level descriptions of the internal processing of control states.

The NML2 architecture highlights the different ways in which concerns are processed within the motivated agent framework. Event- and goal-driven behaviour can be modified through a combination of filter threshold settings and arbitration schemes for meta-level m-procedure unification. This will allow the agent to adapt its behaviour to the requirements of the environment, based on flags reflecting the internal state of the agent. For example, an s-procedure that detects a low success rate of goal achievement could set a flag to cause the filter threshold to be raised and a change in the deciding and scheduling algorithms of the agent. This could reflect an agent's concern to achieve its own goals. Setting a high insistence level for motivator generactivators that respond to situations where babies are in danger would reflect an "Affiliation" (well-being for others) concern.

As this discussion shows, an agent can express many different types of control state without requiring specific *control state* mechanisms – concern-processing is implicit in the reactive generation of motivators as well as in the setting of the attention filter threshold. The following section will build on this idea by looking at the emergence of perturbant control states which have many of the characteristics we normally associate with *emotions*.

4.3 Perturbance and Affective States

Most of the work on Attention Filter Penetration (AFP) theory has been concerned with tertiary emotional states (see section 5.1 for a classification of emotional states) – those involving a temporary loss of control of thought processes. The controversy surrounding the term "emotion" can be avoided by introducing a new term to refer to a "state of temporary loss of control" – that of *perturbance*. Perturbance is a by-product of functional mechanisms (motive management, interruption and filtering), and so is not in itself intrinsically functional or dysfunctional. In this sense *perturbant states* are best thought of as naturally occurring emergent states, of resource-limited intelligent agents responding to rapid changes in hostile environments. This afunctional definition offers a warning against attempting to assign "intrinsic function" to the *temporary loss of control* often associated with emotions (i.e. the blindness of love or anger) over and above that of attention switching.

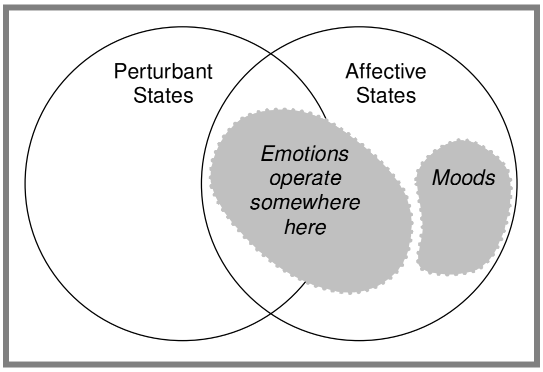


Figure 4.3-1 The Relationship Between Emotions and Perturbant States

Within the confines of AFP theory, affective states are defined as (a) dispositional states, (b) at various levels in the control system (reactive, deliberative, meta-management), (c) that include positive or negative evaluations of something and (d) have at least a tendency to produce motivational states, which (e) in turn have a tendency to produce behaviour. Affective states include not only valenced affects such as *moods*, *emotions* and *feelings*, but also valenced reactive evaluations that never reach management consideration. For example, an affective state of hunger or pain will still exist, if all the eliciting conditions are present and an ongoing activity occupies our attention to the extent that we do not notice our hunger or pain. Affective states may also lie dormant, as is the case of *standards* or moral beliefs that we hold to be true, but only act on occasionally.

By noting that the propensity of a motivator to interrupt ongoing management processing is a function of relatively simple reactive heuristics, AFP theory argues that certain *perturbant affective states*, such as grief, can be explained as an inappropriate interruption of ongoing management processes due to reactive heuristics that have not yet had time to adapt to a change in the agent's environment (i.e. the death or absence of a loved one). The grieving process may very well have a catharsis effect on the agent. But this effect should be viewed as emergent within the context of a reorganisation of the heuristics associated with the object of grief, and not necessarily some social control selected for during evolution – we return to this point in section 7.1.2.

The theory makes a number of predictions about resource-limited autonomous agents, such as the emergence of *perturbant states* (partial loss of control of attention) and the sometimes inappropriate switching of attention due to heuristic insistence assignment. These predictions are based on the natural side-effects of an interrupt mechanism and an attention filter, and are not explicitly built into the theory. By postulating a concurrent control structure with an attention filter, the theory can also allow attention to be diverted without actually causing action plans to be interrupted or disturbed.

4.4 Case Study: A Proto-"Emotional" Agent

MINDER1 [Wright 97] takes the first tentative steps towards a proto-"emotional" agent by including a limited self-monitoring layer capable of detecting the *temporary loss of control of thought processes* associated with tertiary emotions. As MINDER1 does not have specific goals directed towards controlling its own management processes, it cannot be said to exhibit true perturbance – hence the use of the term proto-"emotional" agent. The MINDER1 architecture is shown in Figure 4.4-1.

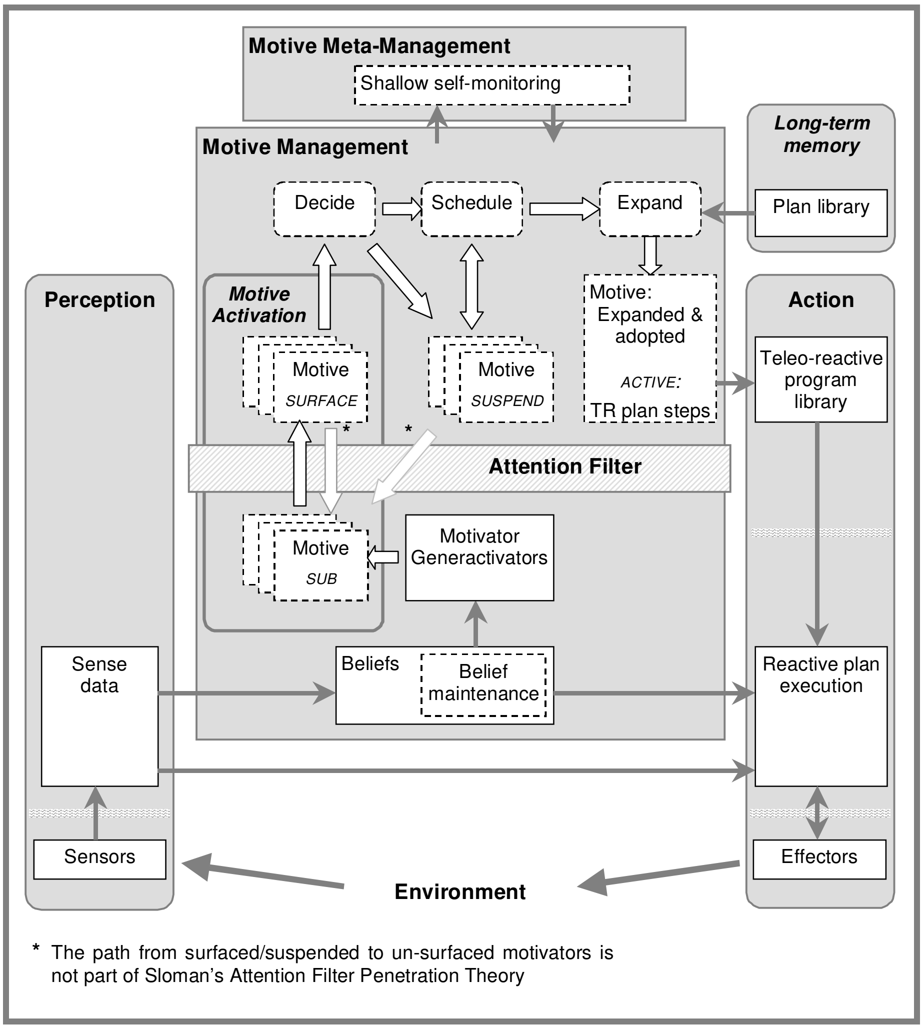


Figure 4.4-1 Simplified view of the MINDER1 architecture

Perturbance

Perturbance arises out of a mismatch between the heuristic *insistence* assignment of the reactive generactivators and the *importance* and *urgency* assigned to the same motivator during the deciding phase of motive management. High insistence motivators might be deemed unimportant or non-urgent by management processes and rejected, only to resurface again and distract attention. MINDER1 determines the presence of a proto-perturbant state by monitoring the rate of rejection of motivators – a high rate of rejection is indicative of a proto-perturbant state. The detection of proto-perturbance could in principle be used to modify the heuristics of the reactive generactivators, allowing adaptation of the agent to the appraised state of the environment.

Motivator Management

The MINDER1 architecture makes explicit the different phases of motive processing identified by Beaudoin [94]. This has led to a refinement of the possible activation states of motivators (see Table 4.4-1) in acknowledgement of the sometimes circular nature of motivator management. In order to decide a motivator, it may be necessary to schedule and partially expand it before all the salient features can be extracted – this will require a meta-plan (a plan executed by the management system as opposed to the plan executor). MINDER1 activation states therefore include suspension during meta-planning as well as during execution.

Activation State	Explanation of State
M	Current agent motive
M_{sub}	Unsurfaced motives
$M_{surfacing}$	Surfacing motives
$M_{surfaced}$	Surfaced motives
$M_{suspended}$	Surfaced but suspended motives
$M_{suspended,\ meta}$	Surfaced and suspended during meta-planning
$M_{suspended,\ execute}$	Surfaced and suspended during execution
M_{active}	Active motives (surfaced and adopted)
$M_{active,\ meta}$	Surfaced and adopted for meta-planning
$M_{active,\ execute}$	Surfaced and adopted for execution

Table 4.4-1 MINDER1 Motive Activation States

Attention Filter

MINDER1 attempts to actively fix (as opposed to manage) the number of surfaced motivators by raising and lowering the filter. The original rationale behind the filter was to protect management resources from excessive interruption by non-urgent events, which is subtly different to fixing the number of surfaced motivators. The MINDER1 architecture therefore deviates from the Attention Filter Penetration (AFP) theory, but should still be considered a useful exploration in the design-space of possible filter designs. There are however a number of problems with the MINDER1 approach that are not necessarily present in the original AFP theory: (i) *filter oscillation* – when two or more motivators with similar insistence values repeatedly surface, and then submerge through the filter together; (ii) *indecision* – when two similarly insistent motivators repeatedly supplant each other as the adopted motivator; and (iii) *rumination* – when the three most insistent motivators are continually rejected and management time is wasted dealing with their decision and scheduling phases.

The above problems associated with the filter are in part due to the shallow design of the architecture: (i) perception is restricted to the reactive layer and hence new motivators are only generated by the reactive generactivators – creating an incentive to use the filter to limit the absolute number of motivators; and (ii) shallow motivator management means that the initial heuristic insistence value carries a large weighting in the final assessment of motivator importance and urgency. But they can also be attributed to the decision to include a path from surfaced/suspended to un-surfaced motivators which does not form part of the AFP architecture – once a motivator has surfaced it should be attended to by deliberative processes, and either accepted or rejected.

4.5 Summary

In this chapter we have extended the motivated agent framework by analysing Beaudoin's [94] case study on the requirements of goal-processing in autonomous agents, and Wright's [97] case study on proto-emotional states. The short-term motivational control states (*beliefs*, *desires*, and *intentions*) introduced in section 2.1 can to a greater extent be captured in the symbolic programming paradigm of the PRS (Procedural Reasoning System) and NML1. Unfortunately, the same cannot easily be said for the more diffuse cluster concept of *emotion*. Some progress towards an intelligent autonomous agent architecture capable of supporting emotional control states has been made [Wright 97], but there is still much more work to be done before we are even close to creating an architecture that can support but a small fraction of the full range of human emotions.

As the above discussions show, we can move forward in elucidating the structural attributes of *motivators* and *emotions* by referring them to information-level descriptions of the underlying control architecture. In the next chapter, we will provide more supporting evidence for the generality of the motivated agent framework by mapping it on to some leading psychological and neurological theories of emotion [Frijda 86; Damasio 94; LeDoux 96].

Part III

“Emotional” Agents

5 Emotional Control States

"[M]any concerns consist of representations of states of affairs that evoke pleasant or unpleasant affect. … Affect elicited by objects or events defining such concerns cannot be said to "serve" these concerns; it merely expresses them. Emotions (affects plus some mode of action-readiness change) elicited by such objects do serve these concerns, by involving signals to the action system, and subsequent changes in action readiness."

– Frijda, *The Emotions* (pages 118-119)

The human emotion process can be viewed as a classic example of an information-processing system primarily geared towards "serving" concerns at all levels of an agent architecture. In this chapter we will provide a broad requirements specification for such an emotion process and, using recent theories from psychology and neurology [Frijda 86; Damasio 94; LeDoux 96], explain the mechanisms inherent in the different classes of emotional states (*primary*, *secondary*, and *tertiary*) from an information-level design-based perspective.

In our quest for a better understanding of the attributes (*functional, dimensional*, and *structural*) of emotions, we will first spend a little time examining some recent (and not so recent) theories of emotion from the fields of cognitive science, psychology and neurology. We will start with Silvan Tomkins' *Affect Theory* (1954), and use his observations as our initial requirements specification for an emotion theory – although we feel his conclusions were a little wide of the mark. Then, in keeping with this historical perspective, we look at Herbert Simons' thoughts on the relationship between cognition and affect (1967) – the inspiration for both Aaron Sloman's and Nico Frijda's interrupt theories of emotion. We then examine Frijda's *Emotion Process* (1986), and by mapping it on to our motivated agent framework start to make explicit the types of information-level concern processes involved in the generation of the different classes (and sub-classes) of emotional state. Finally, we discuss Antonio Damasio's (1994) and Joseph LeDoux's (1996) neurologically grounded theories of emotion, and integrate these theories into our motivated agent framework.

5.1 Classification of Emotions

Emotions form a powerful, but ill-defined class of motivational control states that have spawned a wealth of competing definitions and theories. The architectural framework introduced in section 2.2 (expanded in Figure 5.1-1) allows us to take some tentative steps towards untangling this web of conflicting ideas.

Sloman [97] notes that by referring the different definitions and theories of emotion to the different layers of the motivated agent framework, we can identify three main classes of emotional state – *primary*, *secondary*, and *tertiary*. Those theorists who: (a) stress emotions centred 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 self-reflective, or meta-management layer.

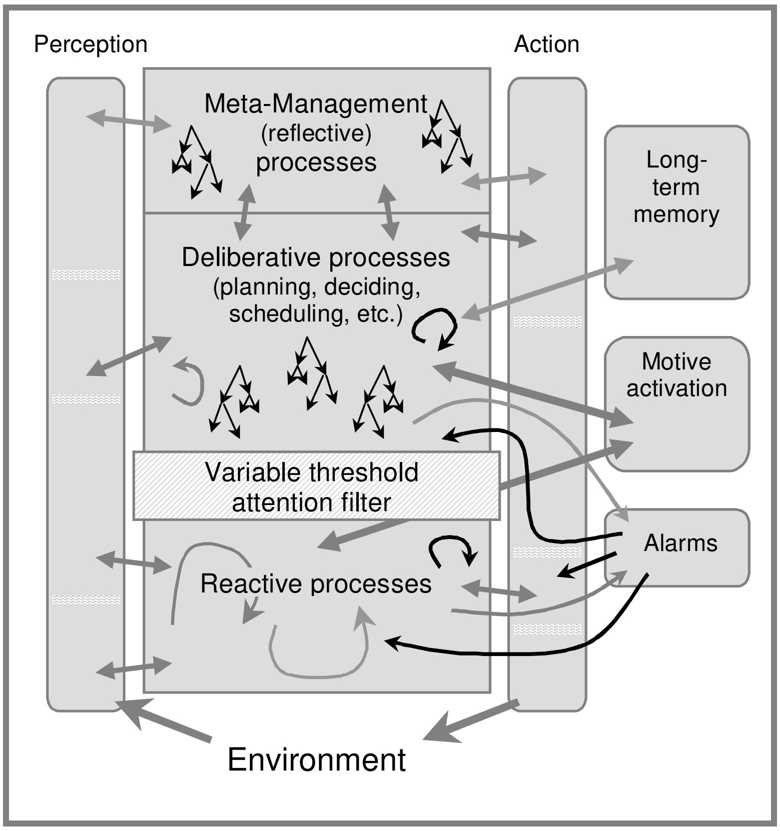


Figure 5.1-1 Motivated Agent Framework [Sloman 99]

The three different classes of emotional state are:

1) *Primary* emotional states: such as being startled, terrified, or sexually stimulated, 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 94; 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. An alarm system capable of detecting features in theses 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 [94] terms 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 "losing 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 not lose what you do not have [Sloman 99]. Tertiary emotions correspond to secondary emotions which reduce self-control.

Our emotion classification scheme does not attempt to explain the rationale of folk-psychology emotion label types (for this we must look to the specific concern-processing mechanisms of the architecture itself), but it does provide an architectural framework within which to understand the mechanisms active in the emotion process better. Although our emotion classes are orthogonal to the classification of emotion types (happiness, sadness, joy, anger, etc.), some emotion types are more closely associated with one particular class of emotional state than another (i.e. relief is generally associated with the non-happening of an unwanted *deliberatively* imagined future event, and therefore normally takes the form of a secondary emotion). However, we would generally expect emotion types to exhibit different characteristics (i.e. varying degrees of cognitive richness) dependent on the layers of the architecture involved in the emotion process. For example, fear can be generated: (a) as an innate response to a situation/event in the external environment – a *primary* emotion; (b) by cognitively identifying a potential future threat – *secondary* emotion; or (c) as a perturbant state in which we repeatedly attempt to reassure ourselves that the threat is not real – a *tertiary* emotion. Each class of fear has its own physiological characteristics (with primary emotional states eliciting the strongest physiological response), and hedonistic tone (with tertiary emotional states being the most cognitive in nature).

5.2 Cognitive Theories of Emotion

> "Just what do the terms cognition and emotion refer to? Do they refer to real functions that are represented in the brain or only to labels that we use as shorthand descriptions of real brain functions? It seems to me that labels is the answer. … Cognition itself has no specific representation in the brain because it is nothing more than a word we use to describe a group of related but diverse information-processing functions, including sensory processing, perception, imagery, attention, memory, reasoning and problem-solving. … Similarly 'emotion' is best viewed not as a function of the brain but as a label that refers to a closely related set of brain functions. The brain has systems that mediate fear, anger, and pleasure, but not a system that mediates 'emotion'." – [LeDoux 94, pages 216-217]

Emotions are a label we apply to a range of concern-processing mechanisms that play a vital role in the survival of biological autonomous agents. However, as the wealth of competing definitions and theories of emotion serve to underline, the term *emotion* itself does not refer to a well-defined class of phenomena clearly distinguishable from other mental and behavioural events [Wright 97]. It is therefore not surprising that the mechanism of *emotion* is notoriously hard to pin down and generally absent from the design of most artificial autonomous agent architectures.

In this section we will attempt to elucidate the concern-processing mechanisms active in emotional states by mapping leading cognitive theories of emotion on to our motivated agent framework.

5.2.1 Affect Theory

> "The affect system is … the primary innate biological motivating mechanism because without its amplification, nothing else matters, and with its amplification anything else can matter." – [Tomkins 84, page 164]

The view that affects are our primary motivational mechanisms was first put forward by Silvan Tomkins in 1954 (these initial ideas were later developed into *Affect Theory*). Tomkins proposed that any theory of affect must address at least three issues [Tomkins 84, page 168] – see Figure 5.2-1

1) The affect has to be activated by some general characteristic of neural stimulation, common to both internal and external stimuli and not too stimulus-specific like a releaser.
2) The activator has to be correlated with biologically useful information.
3) Some of the activators have to be capable of habituation, and some capable of non-habituation.

Figure 5.2-1 Requirements for a Theory of Affect

Affect as Amplification

The basic power of the affect system is seen as a consequence of its ability to combine with – and amplify – a variety of other components in what Tomkins refers to as the central assembly (analogous to the mechanisms of attention and working memory):

> "This is an executive mechanism on which messages converge from all sources, competing from moment to moment for inclusion in this governing central assembly. The affect system can be evoked by central and peripheral messages from any source and, in turn, it can control the disposition of such messages and their sources. Thus it enjoys generality of dependence, independence, and interdependence … Affect can determine cognition at one time, be determined by cognition at another time, and be interdependent under other circumstances." – [Tomkins 95, page 56]

In *Affect Theory*, a single principle – that of the rate and density of neural firing (see Figure 5.2-2) – accounts for all variants of innate affect activation: (i) stimulation increase for startle, fear, and interest; (ii) stimulation decrease for laughter and joy; and (iii) stimulation level for distress and anger. Affect amplifies – increasing the urgency of the thing with which it is co-assembled – through "separate mechanisms, involving bodily responses quite distinct from other bodily responses they are presumed to amplify" [Tomkins 84, page 185]. Amplification is therefore achieved not in a strict linear sense, but through a process of analogy whereby separate mechanisms mimic the neural activity of the activator. "Therefore, enjoyment amplifies by simulating decreasing gradients of neural stimulation. Interest, fear, and surprise amplify by simulating increasing gradients of neural stimulation. Distress and anger amplify by simulating maintained [high] level of simulation." [Tomkins 95, page 89]

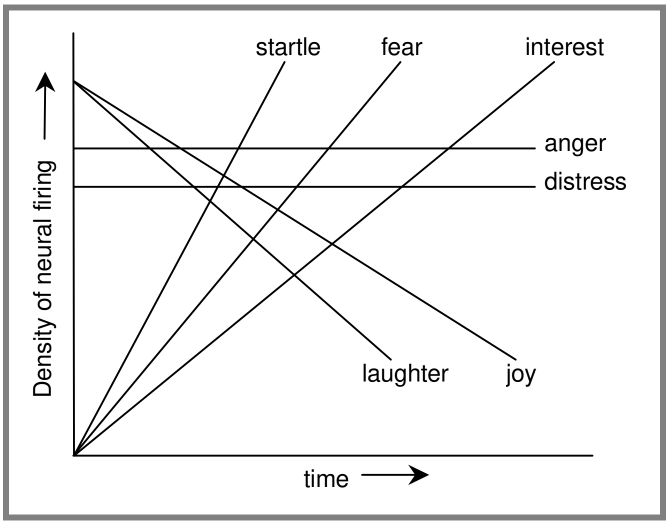


Figure 5.2-2 Model of the Innate Activators of Affect [Tomkins 95, page 46]

Tomkins believed that although each affect was mediated by specific sensory receptors in the skin of the face (see also [Ekman et al. 82]), his somatic centric view did not exclude appraisal from playing a critical role in determining the actual affect experienced. Thus, "the difference between the terror of a specific phobia and the objectless terror that Freud distinguished as 'anxiety' is not a difference in the cold sweat and sensitized, erect hair follicles. It is rather a difference in the consciousness of what information has entered and been co-assembled with affect in the central assembly." [Tomkins 95, page 59]

Further, Tomkins noted that what we "inherit in the affect mechanism is not only an amplifier of its activator but also an amplifier of the response that it evokes." These responses are not restricted to observable motor responses, it may also be in terms of retrieved memories or constructed thoughts, which might vary in acceleration if amplified by fear or interest, in quantity if amplified by distress or anger, or in deceleration of rate of information-processing if amplified by enjoyment.

Conclusions

Although the central principle of affect amplifying by mimicking the neural rate of the stimulator finds little experimental support (it is just as likely that the change in neural activity is an effect of, rather than a trigger for, the mechanisms of primary emotions – i.e. the affect process starts with heuristic relevance evaluations in the early sensory cortex of the brain), there are other aspects of *Affect Theory* that do find some congruence with recent neurological theories of emotion.

Affect Theory makes an implicit assumption that the mechanisms of affect are distinct processes, physically separated from the mechanisms of drives and general cognition. By identifying the limbic system around the amygdala (and not the sensory receptors in the face) as the seat of our emotions, recent neurological research has in a sense confirmed this view – but, as always, the picture is not quite that black and white (see Frijda's *Emotion Process* in section 5.2.3 and LeDoux's *The Emotional Brain* in section 5.2.4). The face can indeed provide some affective trigger/amplification [Ekman et al. 82], however, this can also be explained by the mechanisms of *somatic markers* (see below).

Tomkins' notion of amplification within a central assembly finds analogy with Damasio's mechanisms of *somatic markers*. "In the full somatic-marker hypothesis, I propose that a somatic state, negative or positive, caused by the appearance of a given representation, operates not only as a *marker for the value of what is represented, but also as a booster for continued working memory and attention*. The proceedings are "energized" by signs that the process is actually being evaluated, positively or negatively, in terms of the individual's preferences and goals" [Damasio 96, pages 197-8, emphasis in original]. *Somatic markers* do not amplify through analogy, but rather bias a representation's value/importance/order within working memory and attention – in the terminology of our motivated agent framework, we could say that *somatic markers* increase the *insistence value* of certain motivators.

Two other observations Tomkins made about the affect system deserve special mention: (i) the generality of time, object, intensity, and density (product of intensity times duration) of the affect system forms the structural, innate features that make learning possible; and (ii) "that natural selection has operated on man to heighten three distinct classes of affect – affect for the preservation of life, affect for people and affect for novelty." These are not unlike the high-level classes of motivation identified by Maslow [54] – which can be used to form the basic motivational profile of an agent.

Although we disagree with the conclusions Tomkins draws on the mechanisms through which affect operates (density of neural firing), his observations provide a useful starting point for a requirements specification for the affect system. In the next section we will look at how the mechanisms of affect and cognition interact from an information-processing perspective.

5.2.2 Motivational and Emotional Control of Cognition

> "The environment places important, and sometimes severe real-time demands upon the [organism] … If real-time needs are to be met, then provision must be made for an interrupt system." – [Simon 67]

Like Damasio (see section 5.2.5), Simon acknowledges that human thinking begins in an intimate association with emotions and feelings which is never entirely lost. However, whereas Damasio's theory has a neurological basis and concentrates on the predictive nature of affective states (through *somatic markers*), Simon's interrupt theory starts from the premise that almost all human activity, including thinking, serves not one but a multiplicity of motives at the same time. It is also important to note that Simon's ideas were formulated over thirty years ago – long before recent advances in the neurosciences.

Serial Nature of Central Nervous System and its Control Hierarchy

Simon's theory starts with two basic assumptions: (i) that the Central Nervous System (CNS) is essentially serial in composition and (ii) that behaviour is regulated by a tightly organised hierarchy of goals.

In order to clarify the term serial, Simon identified the basic unit of time of an elementary process as 100ms (the time for a simple reflex) and the basic unit of data as a chunk (i.e. a single familiar symbol such as a syllable, word, phrase, or digit). The serial nature of the CNS was then be inferred from the fact that: (i) the processes that operate during 100ms affect only a few chunks (at most seven) among all those in short- and long-term memory, (ii) during this period not much else can, or does happen, and (iii) "It is difficult to specify how to organize a highly parallel information-processing system that would behave coherently" [Simon 67].

The assumption that behaviour is regulated by a hierarchy of goals is drawn in part from ethological modelling of animal behaviour, but also from the belief that the "obvious" way to organise the behaviour of a serial processor, is as a hierarchy of subroutines. Simon proposed two ways in which a hierarchical serial system could be adapted to cope with multiple goals and avoid the single-minded behaviour of a strict hierarchy: (i) by queuing, and (ii) by generalising a goal to include multifaceted criteria against which possible solutions to the problem can be tested.

Goal Queuing and Multifaceted Criteria

The simplest way in which multiple motives can be attended to in a serial system is by a process of queuing. If an organism is already processing one goal when a new goal is generated, the new goal can be postponed until the first goal has been completed. Here Simon makes an important distinction between goal "completion" and goal "achievement". Goal completion can be decided by a number of criteria: (i) *Aspiration achievement* – A subroutine terminates when its subgoal has been achieved, (ii) *Satisficing* – A subroutine terminates when it has achieved "well enough" its subgoal, (iii) *Impatience* – A subgoal terminates after a certain period of time has been used in trying to achieve it, and (iv) *Discouragement* – A subroutine terminates after a certain number of processes have been tried and failed to achieve it. A goal that "completes" before its sub-goals have been "achieved" can either be abandoned or rescheduled.

If a queuing scheme is to be responsive to an organisms needs, the total time required to complete a goal must remain a fraction of the overall time available to the organism. It is also necessary to generate goals a sufficient time in advance of their "achievement becoming necessary for survival" to allow for the time a goal can sit in a queue before being processed. If goals are more or less periodic (such as the need for sleep), the queuing system can be supplemented, or replaced, by a time-allocation system which processes goals on fixed phases of a cycle.

Simon's second approach to reducing the single-minded behaviour of serial systems was to recognise that goals are rarely unitary entities. The achievement of a goal often calls for a behaviour that meets a range of different criteria. Thus the goal "deliver a speech" could include criteria for "impress the audience", "improve my position within the company", or "enjoy a few days talking with peers". There is no need to single out any one of these criteria as the "goal". It is just as meaningful to say that associated with a behaviour will be a hierarchy of programs responsive to a whole set of criteria.

Real-Time Requirements

A simple queuing system may work well in benign environments, but the bottleneck associated with attention, or short-term memory, soon becomes a crippling factor in adverse environments where real-time needs have to be met. If an organism is to survive in such a competitive environment, provision must be made for an *interrupt system*.

The interrupt system places two requirements on an organism: (i) A certain amount of processing must go on continuously, or almost continuously, to enable the system to *notice* when conditions have arisen that require ongoing programs to be interrupted, and (ii) the noticing program must be capable of *interrupting* and setting aside ongoing programs when real-time needs of high priority are encountered.

Cognition and Affect

Formally, Simon defines the action of *interruption* as the setting aside of the current focus of attention by an activated subset of long-term memory. The activation of long-term memory can either occur as part of the process of cognition (i.e. suddenly remembering something) or as a result of stimulation of the autonomic nervous system or endocrine system (as an emotional response). This definition leads nicely into the area of Simon's theory which deals with the close association between human cognitive thinking and affect.

Simon notes the striking differences between cognition and affect: (i) affect is diffuse, hard to describe and harder to differentiate and classify, whereas cognition is highly specific, mostly representable by strings or structures of symbols, (ii) affect is analogue in nature and susceptible to continuous graduation in degree, whereas cognition is digital in character with symbol structures being discriminated by yes-no tests from other symbol structures, and (iii) affective states change not only continuously, but usually relatively gradually, whereas cognitive structures succeed one another in short-term memory in rapid succession. Given these differences between cognition and affect, Simon argues that the most plausible answer to the question of "how can two such radically different languages communicate?" is by postulating mechanisms of *interruption* and *arousal.*

The mechanisms of *interruption* and *arousal* allow Simon to define the familiar folk-psychology expressions of "affect", "emotion", "mood", "valuation", and "arousal", as states and processes of the CNS (long-term memory, the attentional system and its interrupter, the autonomic nervous system, and the endocrine system): (i) *affect* is used as a generic term, (ii) *emotion* refers to affect that interrupts and redirects attention (usually with accompanying arousal), (iii) *mood* refers to affect that provides context for ongoing thought processes without noticeably interrupting them, (iv) *valuation* refers to association of cognitive "labels" attributing positive or negative valence to objects or events, and (v) *arousal* refers to the stimulation of the autonomic nervous system and the endocrine system [Simon 82].

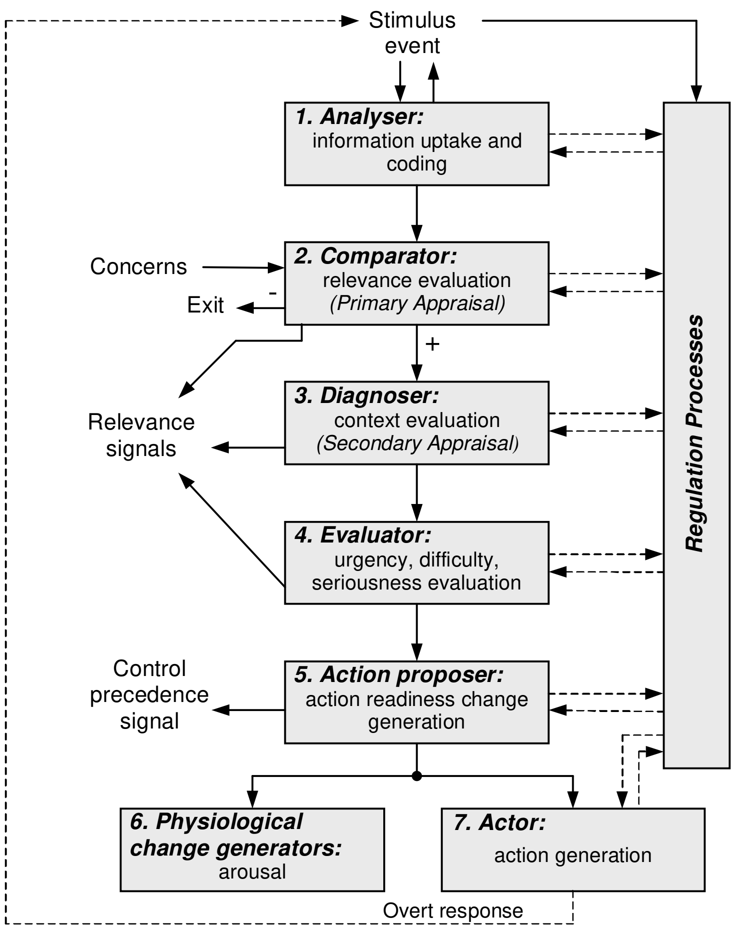
Conclusions

Simon views "emotion" as a control state of the CNS characterised by the diversion of the current focus of attention by some form of affective action (via a change in activation of a subset of long-term memory). The theory leaves open the cause of the affective action (the immediate cause could just as likely be the result of a deliberate appraisal, an automatic defence response, or the result of listening to a soothing piece of music). What is important however, is that the processes of interruption and arousal provide mechanisms through which affective systems can communicate with and/or control cognition – i.e. "Affect can determine cognition at one time, be determined by cognition at another time, and be interdependent under other circumstances." [Tomkins 95, page 56]

Simon's interrupt theory forms the starting point for both Sloman's *Attention Filter Penetration Theory* (discussed in section 4.1), and Frijda's *Emotion Process* (discussed in the next section).

5.2.3 The Emotion Process

Frijda defines an "emotion" as an ongoing process that starts with an affective stimuli and ends with an action in the form of cognitive/overt behaviour and physiological manifestations of *action readiness change*. More specifically, emotions are defined as "*modes of relational action readiness, either in the form of tendencies to establish, maintain, or disrupt a relationship with the environment or in the form of mode of relational readiness as such*" [Frijda 86, page 71]. This process is shown schematically in Figure 5.2-3.



Analyser: The subject is confronted with a stimulus event. The Analyser codes the event, if it can, in terms of known event types and what they might imply with respect to cause or consequence.

Comparator: The stimulus event is appraised as to its relevance for one or more of the subject's concerns: relevance evaluation, or *primary appraisal*. Relevance evaluation results from comparing the event with the satisfaction conditions or sensitivities of the various concerns, for many, or all, concerns in parallel. Outputs of the Comparator are the four relevance signals: pleasure, pain, curiosity, and desire; or, by default, irrelevance, whereupon the process exits.

Diagnoser: The stimulus situation as a whole is appraised in terms of what the subject can or cannot do about it. Context evaluation or *secondary appraisal* diagnoses possibilities or impossibility for coping, and can be regarded as a series of diagnostic tests. Output is a patterned diagnosis or *situational meaning structure* – how the situation appears to the subject, i.e. the valence of the situation as a whole. The *situation meaning structure* comprises of three kinds of elements: (i) cognitions of what the situation does or offers the subject; (ii) cognitions of what the situation allows the subject to do; and (iii) evaluations as to whether the various outcomes are desirable or not. In addition, the Diagnoser, together with the Comparator, provides the Evaluator with information as to how difficult, urgent, or serious events really are.

Evaluator: Urgency, difficulty, and seriousness are computed, and combined in a signal of control precedence (in the *action proposer*) for dealing with the current event. They thus cause action interruption if need be, or else they cause distraction when previous action happens to continue.

Action Proposer: Action readiness change is generated and presses for, or occupies, control precedence. Action readiness change consists of: a plan for action – action tendency (tendencies to execute expressive behaviour, which are present prior to, and independent of actual execution) – and/or for mode of activation. In the case where action programs are fixed and rigid, the concept of action readiness change loses much of its meaning – it only exists to the extent that inhibition can be used to block actions. However where action programs are flexible and alternative courses of action are possible, intentions and goals become independent of the particular actions and the term action readiness becomes more relevant (i.e. an action tendency could be readiness for attacking, spitting, insulting, or turning one's back, whichever appears appropriate at a given moment).

Physiological Change Generator: Physiological change is effected, in accordance with the action readiness mode generated by the action proposer.

Actor: Action – overt or cognitive – is selected, as determined by the action readiness mode and by other aspects of the situation.

Figure 5.2-3 The Emotion Process [Frijda 86, pages 454-456]

An On-going Process

Emotion processes are neither discrete events, nor linear processes. Most of the time information flow is not only from the top down, but bi-directional – stimuli are often actively acquired and context evaluations made prior to information coding. Different stages of the emotion process can be skipped, and the process as a whole interrupted – leading to the myriad of variants of emotional phenomena. But Frijda's view of an emotion in its typical form embodies the process (outlined in Figure 5.2-3) in its entirety. We can make some significant progress towards elucidating this *Emotion Process* by mapping it on to our motivated agent framework.

Primary Emotions are typically triggered by patterns in the early sensory input (sensory thalamus) and detected by a dedicated global alarm system (centred on the limbic system). Figure 5.2-4 shows a simplified graphical representation of the information flow that leads to a primary emotion state within our three-layered model. In earlier diagrams (Figure 5.1-1) we kept the global alarm system as a separate entity, here we attempt to place it within the confines of the three pillars (*perception*, *cognition*, and *action*). Although we would like to depict a clean boundary between perception and cognition (i.e. both deliberation and the reactive *concern-processing* substrate), the border between the two is in reality very fuzzy (this is not all that surprising when you acknowledge that both perception and cognition simply refer to labels that help us carve up the functionality of the brain). It can be argued that relevance and context evaluation belong to both cognition and perception – as physically the sensory thalamus is also responsible for the early sensory processing of such information in our brain (see *'Amygdala Pathways and Fear Conditioning' in the Emotional Brain* in section 5.2.4).

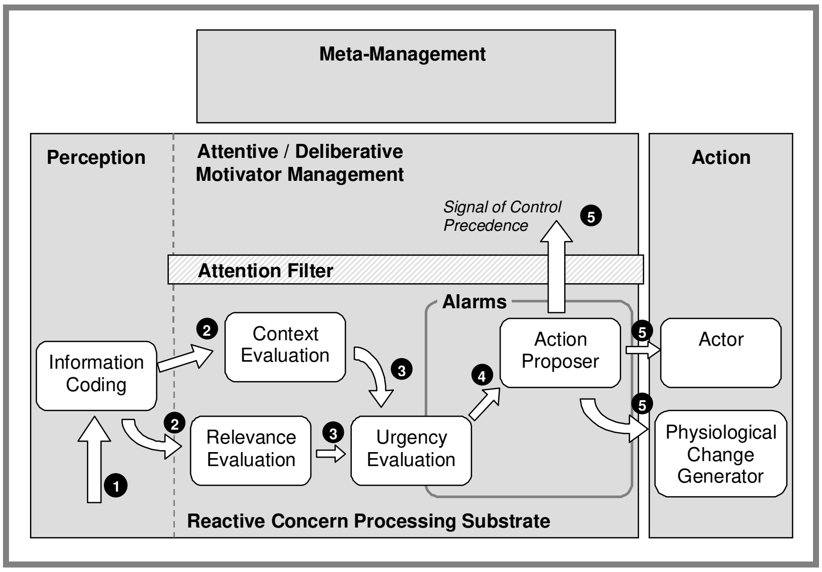


Figure 5.2-4 Information Flow leading to a Primary Emotion State

The information flow that leads to a primary emotion can be summarised as: (**1**) external percepts are detected and encoded into known event types; (**2**) in parallel, the event is evaluated relative to the agent's concerns and the context of the current situation (i.e. the agent's coping strategies) – relevance and context evaluation must rely on simple heuristics such as speed, intonation, size, habituation or familiarity; (**3**) the urgency of the event is evaluated – as a simple function of the current level of arousal, context and relevance evaluations; (**4**) action readiness change is generated and presses for control precedence; and finally (**5**) attentive cognition is interrupted as the motivator gains control precedence, an involuntary action *might be* performed, and some form of physiological change *might be* instigated according to the action readiness mode generated by the action proposer.

Agents that have a deliberative layer (or at least an active attention mechanism and working memory) are in theory better able to evaluate the true seriousness of a situation – and therefore produce a more measured emotional response. Figure 5.2-5 shows the flow of information for a *Secondary Emotion* that relies on a deliberative evaluation of coping strategies.

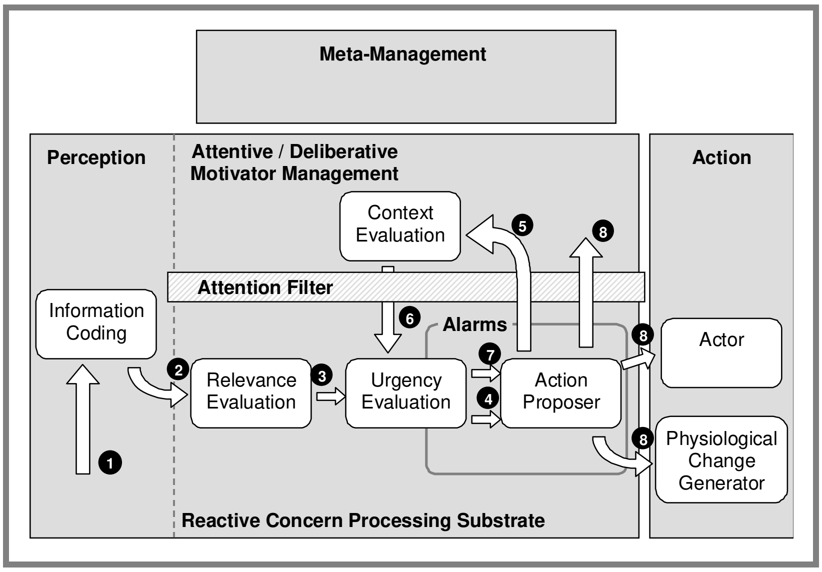


Figure 5.2-5 Secondary Emotion featuring Deliberative Context Evaluation

In humans the difference between the generation of a primary or a secondary emotion can simply be a question of the initial urgency attached to the stimulus. In Figure 5.2-5 the emotion process proceeds from (**1**) through (**4**) as per a typical primary emotion. However, instead of triggering a full emotional response, only attention is captured before the context is deliberately evaluated at (**5**). At this point we could simply be reacting to a loud noise (a startle response), without actually assessing the context of the situation (if we were alone in a dark house a reactive context evaluation in the form of heightened arousal could already be enough to trigger a physiological emotional response). Having evaluated the context as serious (**6**), our alarm system kicks in and generates an emotion proper (**7**) and (**8**).

Certain secondary emotions such as being anxious, apprehensive, or relieved, depend on the existence of a deliberative layer in which plans can be created and executed with relevant risks noticed, progress assessed, and success detected. Figure 5.2-6 shows the information flow for a secondary emotion triggered by such a deliberative thought process. Although secondary emotions might not actually generate physiological change (which varies from individual to individual), they still utilise much of the machinery of primary emotions (global alarm system) when capturing and diverting attention.

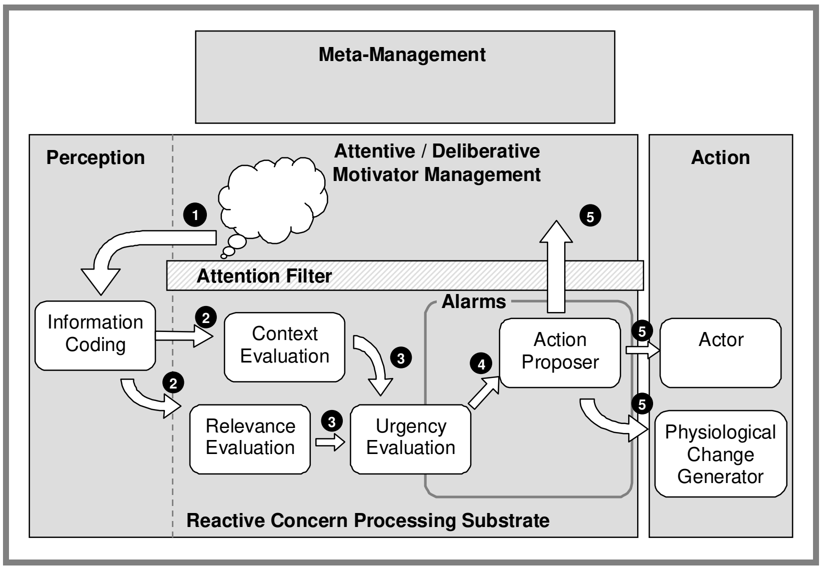


Figure 5.2-6 Secondary Emotion triggered by Deliberative Thought Processes

Finally, we can have the special case of secondary emotions which reduce control of the attention mechanism – the machinery of *Tertiary Emotions*. Figure 5.2-7 shows the information flow within a tertiary emotion.

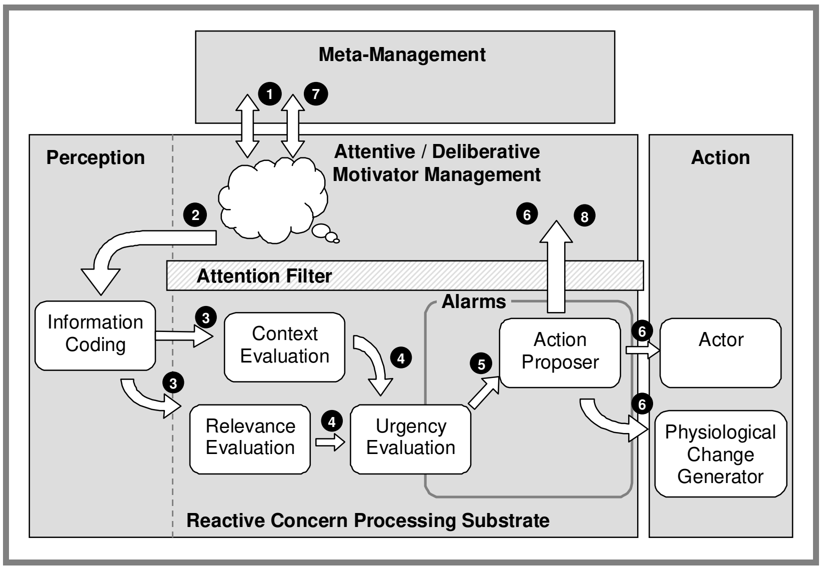


Figure 5.2-7 Tertiary Emotion featuring loss of control of Deliberative Management

Tertiary emotions – such as grief and longing – are characterised by a difficulty to focus attention on urgent or important tasks. The architecture must therefore support a meta-management layer that attempts to manage the deliberative thought process (**1**). Normal deliberative thought processes trigger the mechanisms of primary emotions resulting in a signal of control precedence which presses for and temporarily gains control of the attention mechanism (**2**) through (**6**). Meta-management processes are still able to detect this change and re-evaluate the situation as less important than the current task and so regain control (**7**).

However, thoughts keep returning to the object of concern (as the reactive context evaluation has yet to adjust to the new situation), and ongoing deliberative processes trigger further interruptions (**8**) – resulting in an emergent perturbant state.

Conclusions

In this section we have shown how the different classes of emotional states utilise different information-processing pathways in the brain, and established the *Emotion Process* firmly in the terminology of the motivated agent framework. In our discussion, we have remained a little vague about the actual physical form an emotional response must take (aside from generating a signal of control precedence). This vagueness is a reflection of the multiple pathways through which the mechanisms to which we attach the label "emotion" operate – emotions are *emergent* states, and can be a little fuzzy around the edges.

Our common "folk psychology" understanding of what an emotion is, usually includes the idea of a feeling state in the form of some kind of physiological arousal. This is certainly true of intense emotional "passions", but it is not always so obvious in more "cognitive" manifestations of emotions. For example, the emotional state of guilt can refer to a perturbant state of repeated re-direction of attention (as thoughts switch between whether detection will occur, whether to confess, likely punishment, how to atone, how to avoid detection, etc.) without resulting in measurable levels of physiological arousal.

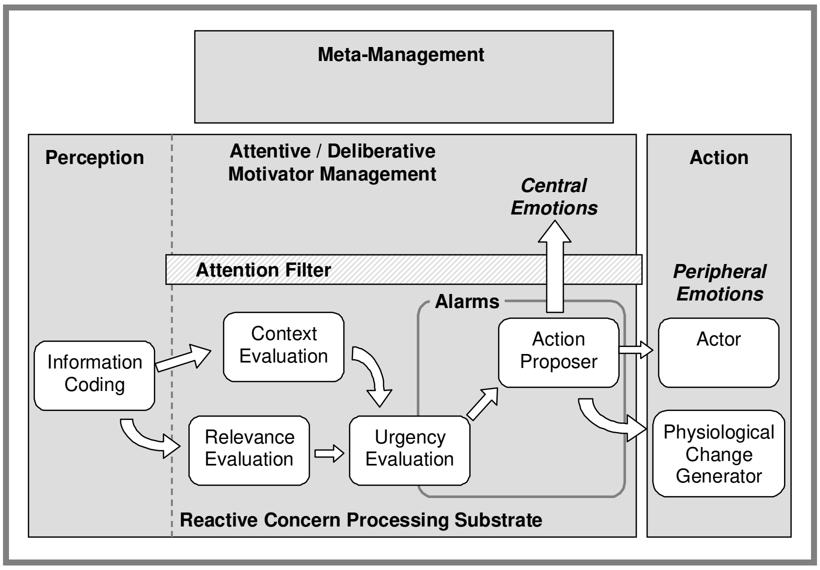


Figure 5.2-8 Central and Peripheral Emotional Sub-Classes

Based on the path(s) taken from the Action Proposer (see Figure 5.2-8), we can identify two sub-classes of emotional state (orthogonal to our main classes of *primary*, *secondary*, and *tertiary* emotions): those of (i) *central*; and (ii) *peripheral* emotions [Sloman 99, section 5.6]. *Central emotions* refer to the core aspects of the *Emotion Process* – namely re-direction of attentive processing in service of agent concerns. *Peripheral emotions* refer to those emotions that *also* trigger a change in body state (a change in body state *without* interruption of attentive processing is classified as affect, and not emotion). Both mechanisms utilise the machinery of a global alarm mechanism (centred on the limbic system in humans).

We have started to map aspects of the emotion process on to the actual physical structures of the brain (sensory thalamus, sensory/polymodal/supramodal cortex, and the amygdala). In the next section we will expand this investigation by looking at a neurological model of fear processing within the terminology of our extended motivated agent framework.

5.2.4 The Emotional Brain

> "[E]motion is best viewed not as a function of the brain but as a label that refers to a closely related set of brain functions. The brain has systems that mediate fear, anger, and pleasure, but not a system that mediates 'emotion'. At one time, it seemed that the limbic system might fill the role of a general-purpose emotion system [MacLean 52], but we now know that this is not the case [Brodal 82; Swanson 83; LeDoux 87, 92]. The limbic system points us in the direction of some relevant parts of the brain for emotion but tells us very little about the brain mechanisms of any given emotional process." – [LeDoux 94, Page 217]

Emotions do not refer to a nice self-contained system of the brain, with a well-defined functional role, and a clear physical boundary. Emotions emerge from the interaction of many different systems, performing many different roles, and operating at many different levels within a biological agent architecture. LeDoux [94] believes that the best way to unravel the underpinnings of emotional life is to systematically study the neural pathways of the individual emotion systems – starting with the system that mediates fear.

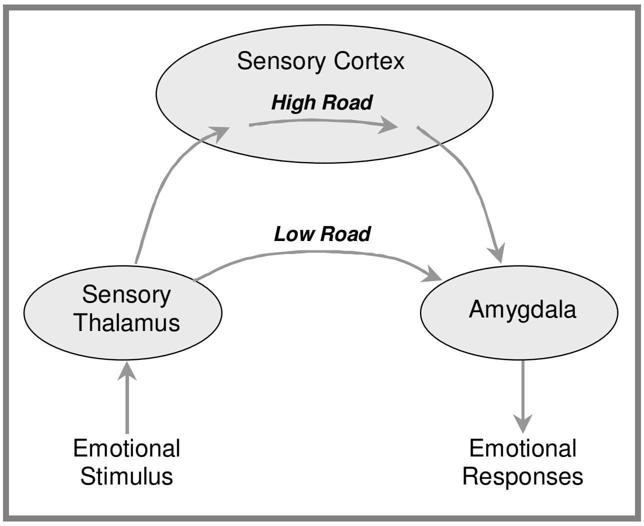


Figure 5.2-9 The Low and High Roads to the Amygdala [LeDoux 96, page 164]

Figure 5.2-9 shows the high and low information-processing pathways of the brain that lead from emotional stimulus to emotional response. The low road provides the quick and dirty pathway for our immediate reactions. The high road leads through the sensory cortex and provides a more accurate representation of the stimulus, but takes a little longer to reach the amygdala. Whereas the sensory thalamus is biased towards evoking a response, the role of the sensory cortex can be viewed as that of preventing an inappropriate response (rather than producing an appropriate one). In the terminology of the *Emotion Process*, the sensory thalamus performs the initial relevance evaluation of the stimulus, and the sensory cortex performs part of the context evaluation process (relevance and context evaluation are labels used to describe operations that occur within the emotion process, that may or may not map on to discrete physical structures of the brain).

The low road from sensory thalamus to amygdala gives the amygdala a head start in responding to an emotional stimulus (the thalamus pathway takes 12 milliseconds to process an auditory stimulus in rats, whereas the cortical pathway takes almost twice as long [LeDoux 96, page 163]). These extra few milliseconds serve a number of useful functions: (a) they trigger the body's defence mechanism in preparation for action; and (b) through heightened arousal and attention, provide a focus for sensory cortex perception and create associations and memories of the event. Figure 5.2-10 shows the amygdala pathways active during fear conditioning – the amygdala provides input to the sensory cortex and supramodal cortex, biasing perception and memory retrieval.

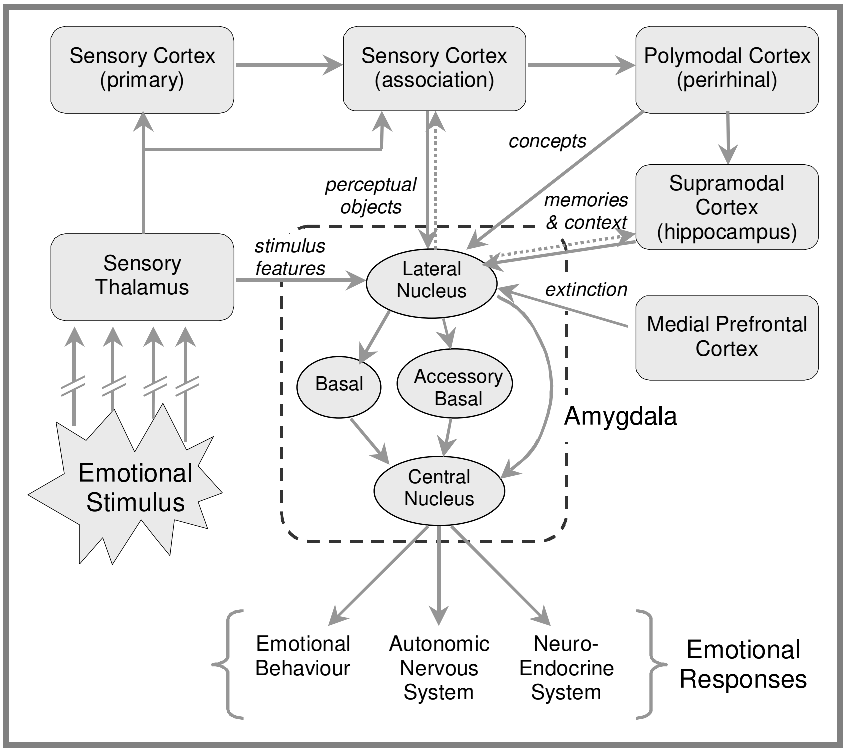


Figure 5.2-10 Amygdala Pathways in Fear Conditioning [modified LeDoux 95, 96]

The amygdala lies squarely at the centre of the fear emotion complex: (i) the sensory thalamus provides a fast pathway to the amygdala, by responding to low-level features of the stimulus; (ii) the sensory cortex provides a path for more complex aspects of the stimulus (event/object) to reach the amygdala; (iii) the polymodal cortex creates concepts/associations between the different sensory modes (visual, auditory, and somatic); which then feed into (iv) the supramodal cortex (hippocampus) to allow explicit past memories of similar situations to affect the emotion process; and finally (v) the medial prefrontal cortex allows extinction of previously conditioned responses through habituation.

Emotions and Memory

As we hinted at above, there are two different memory systems involved in the emotion process: *implicit* memory of the current event; and *explicit* memory of past emotions. Both memory systems are affected by the arousal of the cortex during an emotional episode.

> "While much of the cortex is potentially hypersensitive to inputs during arousal, the systems that are processing information are able to make the most of this effect. For example, if arousal is triggered by the sight of a snake, the neurons that are actively involved in processing the snake, retrieving long-term memories about snakes, and creating working memory representations of the snake are going to be especially affected by arousal. Other neurons are inactive at this point and don't reap the benefits. In this way, a very specific information-processing result is achieved by a very nonspecific mechanism." – [LeDoux 96, pages 287-288]

There are a number of different systems which contribute to arousal – all of which interact with the amygdala in some way. Four systems are located in regions of the brain stem, and operate by releasing different neurotransmitters: acetylcholine; noradrenaline; dopamine; serotonin. A fifth system is located in the nucleus basalis (near the amygdala), and is likely to be the principle player in emotional arousal – releasing acetylcholine in response to a novel or otherwise significant emotional stimulus.

Arousal occurs in response to any novel stimulus as part of the normal attention mechanism. Arousal mediated by direct inputs from the sensory system to the arousal networks is quickly habituated, and therefore temporary. However, novel stimuli which are emotionally significant also trigger the amygdala and the nucleus basalis arousal system. These two systems together amplify the arousal effect (through the release of acetylcholine, and not amplification by analogy à la Tomkins (see section 5.2.1) – although somatic feedback can still allow facial muscles to influence affect), placing the cortical network in a state of hypersensitivity. The hypersensitive cortical network combined with a direct connection from the nucleus basalis, drives the amygdala to form a feedback loop, and maintain the state of hypersensitivity.

> "These representations converge in working memory with the representations from specialized short-term memory buffers and with representations from long-term memory triggered by current stimuli and by amygdala processing. The continued driving of the amygdala by the dangerous stimulus keeps the arousal systems active, which keeps the amygdala and cortical networks engaged in the situation as well. Cognitive inference and decision making processes controlled by the working memory executive become actively focused on the emotionally arousing situation, trying to figure out what is going on and what should be done about it. All other

inputs that are vying for the attention of working memory are blocked out." [LeDoux 96, page 291]

The information content provided by the arousal system is very weak, and so arousal tends to lock you into whatever emotional state you are in when the arousal occurs. This state is then maintained until something else happens that is significant enough to shift the focus of arousal. "While arousal is nonspecific and tends to lock you into the state you are in when the arousal occurs, unique patterns of visceral, especially chemical, feedback have the potential for altering which brain systems are active and thus may contribute to transitions from one emotion to another within a given emotional event." [LeDoux 96, page 293]

Conclusions

By focussing on the emotion of fear, LeDoux has been able to identify at least one set of distinct information-processing systems and pathways that make up the *emotion process*. These regions of the brain perform a number of different roles aside from processing affective stimuli – fitting with our hypothesis of emotions as emergent motivational control states. Further, although the amygdala has been identified as the main player in the global alarm system for fear, it would be premature to assume that it has such an active role in all emotion systems within the brain. "[T]he amygdala, which sits in the depth of each temporal lobe, is indispensable to recognizing fear in facial expressions, to being conditioned to fear, and even to expressing fear. […] The amygdala, however, has little interest in recognizing or learning about disgust or happiness. Importantly, other structures, just as specifically, are interested in those other emotions and not fear." [Damasio 99, pages 61-62]

Although there is no such thing as a single emotion system within the brain, we can still usefully map the individual emotion mediating systems on to our motivated framework and start to ask questions about how these different control states interact – especially in terms of attention and working memory. The classic fear circuit utilises the full armoury of the amygdala to trigger arousal and a *peripheral* emotional response (i.e. one involving interruption and the generation of a physiological change). However, it is also possible for an affective stimulus to activate many of the pathways of the fear system without interrupting attention (i.e. when the motivational attitude is not great enough to gain control precedence) – in which case we would still be able to detect some physiological change, but without a specific focus, would describe ourselves as '*anxious*'. In the next section we will examine how somatic markers, and "as if" loops, can be used to generate *central* emotions (without significant physiological change) and the cognitive feelings that aid complex decision making by adding *valence* and *motivational attitude* to current/future possible events.

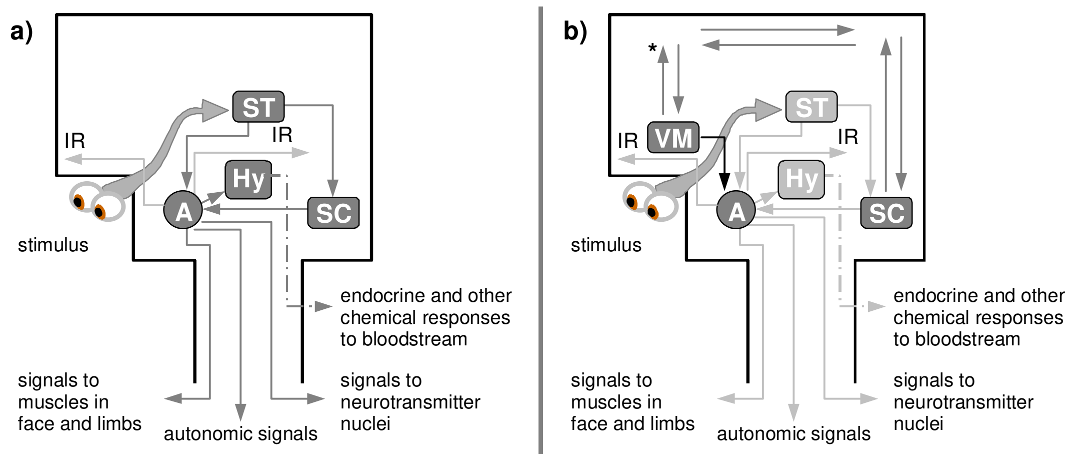
5.2.5 The Somatic Marker Hypothesis

> "[E]motion is the combination of a *mental evaluative process*, simple and complex, with *dispositional responses to that process*, mostly *toward the body proper*, resulting in an emotional body state, but also *toward the brain itself* (neurotransmitter nuclei in brain stem), resulting in additional mental changes." [Damasio 96, page 139]

We are born with certain innate neural machinery capable of generating somatic states (both visceral and non-visceral) in relation to certain classes of stimuli – Damasio's machinery of *primary emotions*. In addition to these innate capabilities, we also possess the ability to form systematic connections between categories of objects and situations on the one hand, and primary emotions, on the other. These learned associations and feelings – which Damasio has termed the mechanisms of *secondary emotions* [Damasio 96, page 134] – are the somatic markers of the *somatic marker hypothesis*.

The somatic marker hypothesis is more than a simple classification scheme for primary and secondary emotional states based along the lines of *innate* versus *acquired* associations – this would certainly fail to take into account the emergent nature of emotion and the true complexity of the *emotion process* as we have outlined in the preceding sections. Primary emotions can easily involve quite complex acquired associations between different sights, sounds, and smells (as demonstrated by animal fear conditioning experiments). We would still classify fear as a *primary* emotion if the *acquired* associations *only* expressed themselves through the reactive concern-processing machinery of the *limbic system*. In other words, although the generation of both *primary* and *secondary* emotions can involve *acquired* associations, secondary emotions are distinguished from primary emotions by the fact that the attentive/deliberative layer plays some part in the emotion process – "Structures of the limbic system are not sufficient to support the process of secondary emotions." [Damasio 96, page 134]

Damasio's view of *secondary emotions* begins with the deliberative consideration of the event, object, or situation, expressed through mental images organised in the thought process (as shown in Figure 5.2-11). The images/representations themselves are constructed under the guidance of dispositional representations held in higher-order association cortices, but the actual neural substrate for the images is the collection of separate topographically organised representations, occurring in various early sensory cortices (the polymodal and supramodal cortices are also required for the processing of concepts and memories). Networks in the prefrontal cortex automatically and involuntarily respond to signals arising from the processing of these mental images (the prefrontal response comes from dispositional representations that embody knowledge pertaining to how certain types of situation have been paired with certain emotional responses, i.e. from *acquired* rather than *innate* dispositional representations). Finally, the automatic responses are signalled to the *amygdala* and the *anterior cingulate* – *utilising the machinery of primary emotions* – resulting in: (i) activation of the autonomic nervous system; (ii) activation of the motor system to give the external picture of the emotion; (iii) activation of the endocrine and peptide systems, resulting in changes in body and brain state; and (iv) activation of the non-specific neurotransmitter nuclei in the brain stem and basal forebrain (arousal).



Primary Emotions. The black perimeter stands for the brain and brain stem. After an appropriate stimulus activates the amygdala (A) – via sensory thalamus (ST) and sensory cortex (SC) –, a number of responses ensue: internal responses (marked IR); muscular responses; visceral responses (autonomic signals); and responses to neurotransmitter nuclei and hypothalamus (Hy). The hypothalamus gives rise to endocrine and other chemical responses which use the blood stream route.

Secondary Emotions. The stimulus may still be processed directly via the amygdala but is now also analysed in the thought process, and may activate frontal cortices and the ventromedial prefrontal cortex (VM). VM acts via the amygdala (A). In other words, secondary emotions utilise the machinery of Primary Emotions.

* The information pathways are schematic, and do not represent the actual neural pathways of the brain.

Figure 5.2-11 Emotion Mechanisms [modified Damasio 96, pages 132 and 137]

Mapping Damasio's emotion mechanisms on to our motivated agent framework (see Figure 5.2-12) allows us to clarify a little better what exactly is meant when talking about "secondary emotions utilising the machinery of primary emotions". Things get a little complicated in biological systems as attention and interruption are in part mediated by arousal (a physiological change) – we will circumscribe this problem by defining physiological change to exclude effects on the attention system.

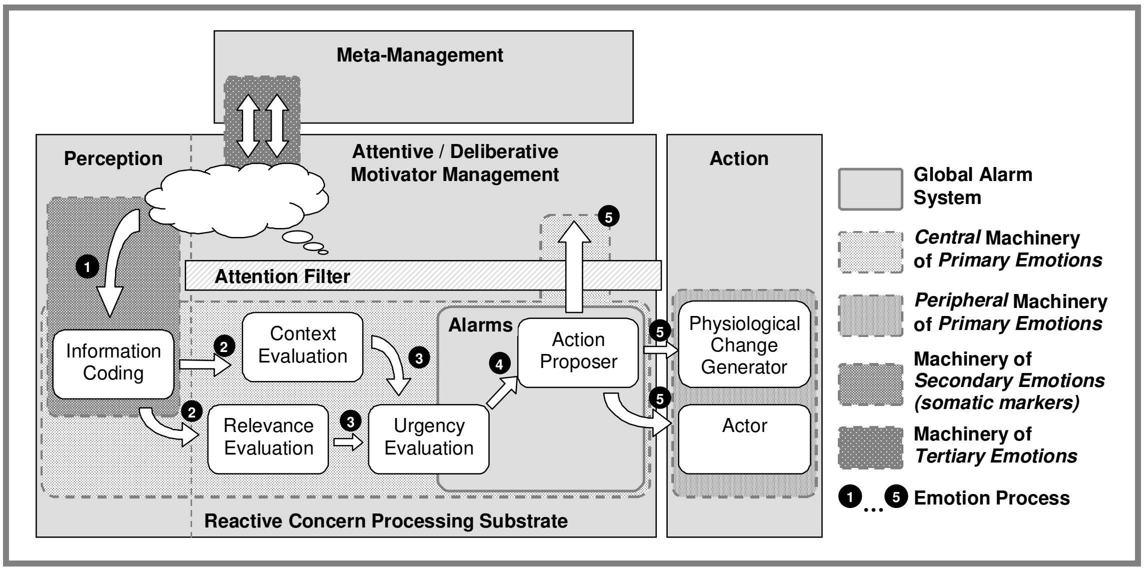


Figure 5.2-12 Emotion Mechanisms and our Three-Layered Architecture

Conceptually, there are two different parts to the machinery of primary emotions: (i) the *central* machinery is responsible for detecting the relevance of the event and interrupting attentive processing; whereas (ii) the *peripheral* machinery generates overt action and physiological change. An emotion can utilise the *central* machinery of primary emotions, without triggering the *peripheral* machinery (this is especially true of the more cognitive *tertiary* emotions such as guilt). If we are generous and allow the *central* machinery of primary emotions to include most of the limbic system (i.e. the *anterior cingulate* and *amygdala*), then most theorists are likely to agree that "secondary emotions utilise the machinery of primary emotions" – especially as the *anterior cingulate* plays an important role in attention. However, not all parts of the machinery of primary emotions are utilised by all secondary emotions (not all parts of the machinery are even utilised by all primary emotions), and parts of the machinery play a critical role in non-emotional processes as well – the brain is a product of evolution and "*nature's tinkering style of engineering*." [Damasio 96, page 137, emphasis in original]

Having described the main players in the emotion plot, we can now take a closer look at Damasio's somatic marker hypothesis, and the role of somatic markers in the decision making process.

Somatic Markers and "as if" Loops

> "Whether we conceive of reason as based on automated selection, or on logical deduction mediated by a symbolic system, or – preferably – both, we cannot ignore the problem of order. I propose the following solution: (1) If order is to be created among available possibilities, then they must be ranked. (2) If they are to be ranked, then criteria are needed (values or preferences are equivalent terms). (3) Criteria are provided by somatic markers, which express, at any given time, the cumulative preferences we have both received and acquired." – [Damasio 96, page 199]

In its strongest form, the somatic marker hypothesis claims that: (i) the valence associated with emotional episodes is generated by the somatosensory cortex (either through real physiological change involving the body loop, or triggered directly through the “as if” loop); and (ii) that the resultant somatic state can be used to rank choices in deliberative reasoning. As Rolls [99, page 63] points out, the beauty of this stance is that both these predications can be tested in patients with somatosensory cortex damage – interestingly, Damasio does offer some evidence for emotional and reasoning impairments in people with stroke damage to the dominant right somatosensory cortex [Damasio 96, pages 67-69] (also appendix C).

We do not need to adopt this extreme view to provide a useful information-level analysis of the somatic marker hypothesis. In the following analysis, we will adopt a weaker form of the theory – still allowing the somatosensory cortex to play a central role, but not insisting that it is solely responsible for the valence of emotional events or an ability to rank situations on the basis of somatic state. In our emergent view of emotions, we are also quite happy to accept that mechanisms, other than somatic states, are at play in attaching valence to situations and events – our brains have had plenty of time to chance upon useful information-processing short-cuts (just as Damasio’s “as if” loop bypasses the body proper – see following discussion).

The somatic markers are the “gut-reactions” that allow us to make instinctive decisions as to whether something is right or not. More specifically, they refer to “*a special instance of feelings generated from secondary emotions* … [*which*] *have been connected, by learning, to predicted future outcomes of certain scenarios*. When a negative somatic marker is juxtaposed to a particular future outcome the combination functions as an alarm bell. When a positive somatic marker is juxtaposed instead, it becomes a beacon of incentive.” [Damasio 96, page 174]

Aside from influencing decision making through the establishment of a particular body state (the *peripheral* machinery of primary emotions), a second mechanism is also prevalent in the somatic marker process – that of the “as if” loop. “In the alternative mechanism the body is bypassed and the prefrontal cortices and amygdala merely tell the somatosensory cortex to organize itself in the explicit activity pattern that it would have assumed had the body been placed in the desired state and signaled upward accordingly.” [Damasio 96, page 184] Damasio sees these “as if” loops as the symbolic representations of somatic states. They play an increasingly important role in the decision making process as we mature and start to categorise repeated situations and experiences. The extent to which people depend on real versus “as if” body states is left as an open question, but probably varies considerably from person to person.

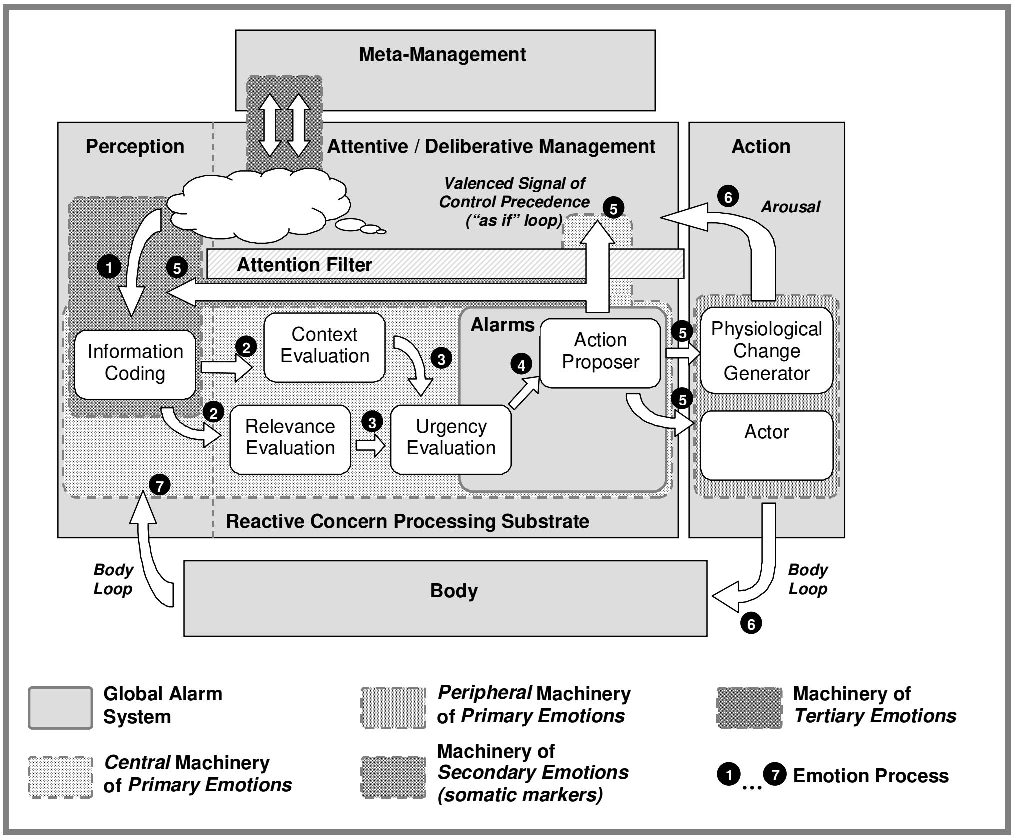


Figure 5.2-13 Somatic Markers and the Body Loop (real and "as if")

Figure 5.2-13 shows the information-processing pathways active in emotional states that utilise the real, or "as if", body loop to generate the "feeling" state of an emotional experience. Although many of our daily decisions proceed without feelings, it does not mean that evaluations that would normally lead to change in body state and associated feeling state have not taken place. It could simply be that the body state signal (real or "as if") may have been activated, but not been made the focus of attention (awareness). Equally, a change in body state could have been activated by neurotransmitter nuclei (arousal), thus biasing cognition in a covert manner and influencing the decision making process indirectly without actually being consciously "felt".

Conclusions

Somatic markers perform a similar function to Frijda's relevance signals (pleasure, pain, desire, and curiosity) – they provide a heuristic *anticipated* measure of the future value (relevance) of an event or situation to an organism. This measure does not come in the form of a precise number that falls out of some neurological equivalent to a fitness function, but rather a non-specific feeling that dynamically changes as the focus of attention shifts to accommodate different aspects of the situation or event (i.e. as the particular coalition of concepts, memories, and objects change – see the *Amygdala Pathways and Fear Conditioning* in Figure 5.2-10).

Our ability to discriminate the feeling state associated with somatic markers is poor (due in part to the cumulative nature of somatic markers, but also due to the "snap-shot" nature of somatic markers, which is so far removed from the dynamic emotion process to which they are paired). However, we can still generally reduce this feeling state to one of valence and intensity. Having a "gut-feeling" does aid decision making to some degree, but the overall valency and intensity of a somatic marker does not represent a concrete value with which we can reliably order the different possibilities generated during the deliberative decision making process (cognitive comparisons require highly specific representations, whereas the somatic markers provided by the affect system are very diffuse – see discussion by Simon [67]; also section 5.2.2).

> "What does the *somatic marker* achieve? It forces attention on the negative outcome to which a given action may lead, and functions as an automated alarm signal which says: Beware of danger ahead if you choose the option which leads to this outcome. The signal may lead you to reject, *immediately*, the negative course of action and thus make you choose among fewer alternatives." – [Damasio 96, page 173]

Damasio's claim that somatic markers allow you to choose between fewer alternatives needs further clarification. It is not simply the case that certain situations carry a valence (in the form of a somatic marker) that allows us to make a good choice, but the fact that some mechanism in our brain responds to this valenced evaluation to *reduce our choice automatically*.

Through a number of ingenious "gambling" experiments [Damasio 96, pages 212-222], Damasio was able to show how patients with ventromedial prefrontal cortex lesions were unable to change their initial behaviour, even if that behaviour was known to be counter productive (i.e. the patients went broke during the game and needed a loan from the "bank"). These patients were still able to attach valence to the situation, but seemed unable to act on this knowledge. "One of the hallmarks of frontal lobe damage in humans is perseveration, the inability to stop doing something once it is no longer appropriate." [LeDoux 96, page 249; Rolls 95, page 1100] It is almost as if the affective feeling state was present, but the sense of urgency, or importance, was missing – the somatic markers were devoid of *motivational* meaning and did not act as alarm signals.

> "Damage to this region [the ventromedial prefrontal cortex] in animals interferes with short-term memory about reward information, about what is good and bad at the moment, [Gaffan et al. 93] and cells in this region are sensitive to whether a stimulus has just led to a reward or punishment. [Thorpe et al. 83; Rolls 92; Ono and Nishijo 92] Humans with orbital frontal damage become oblivious to social and emotional cues and some exhibit sociopathic behavior. [Damasio 94] This area receives inputs from sensory processing systems (including their temporal buffers) and is also intimately connected with the amygdala and the interior cingulate region. The orbital cortex provides a link through which emotional processing by the amygdala might be related in working memory to information being processed in sensory or other regions of the neocortex." – [LeDoux 96, page 278]

What cannot be said for certain is if the ventromedial prefrontal cortex primarily operates via the *amygdala*, or directly through the *interior cingulate*. The medial prefrontal cortex "receives signals from the sensory regions of the cortex and from the amygdala, and sends connections back to the amygdala, as well as many areas to which the amygdala projects. The medial prefrontal cortex is thus nicely situated to be able to regulate the outputs of the amygdala on the basis of events in the outside world as well as on the basis of the amygdala's interpretation of those events." [LeDoux 96, page 248]. The ventromedial prefrontal cortex can almost be said to perform a meta-management role in the emotion process.

Damasio's patient with calcified amygdala (patient "S") showed no signs of the sort of social and life problems that accompany patients with ventromedial prefrontal cortex damage. Further, Damasio's patients who suffer from ventromedial prefrontal cortex lesions can still make rational decisions as to what is good or bad, but suffer from an infinite regress – they have no measure of sufficiency or threshold above which good is good enough. There is obviously something else happening in the somatic marker process apart from simply attaching valence to situations and events. It was almost as if patient "S" did not *have* the negative somatic markers that would have been associated with the "fear" emotion, whereas patients with ventromedial prefrontal cortex damage have somatic markers, but those markers carry no *motivational* meaning.

To summarise, *somatic markers* come in two varieties: (i) *explicit* valenced memories of the situation which utilise the hippocampal regions (with somatic input from the emotion process); and (ii) *implicit* valenced memories which utilise the amygdala and the machinery of primary emotion systems. Additionally, there are mechanisms at play which add *motivational* meaning to somatic markers – with the most likely candidate brain regions being the ventromedial prefrontal cortex and anterior cingulate. Finally, somatic markers express themselves through two different pathways: (i) using the machinery of primary emotions through the body loop; and (ii) by directly projecting into the somatosensory and attention networks through the "as if" loop. The extent to which people use the real or "as if" body loop varies considerably from person to person, and emotion to emotion.

5.2.6 Conclusions

The first conclusion we can draw from this brief overview of research into emotions, is that there is *no single system in the brain that mediates emotion*. There are systems that mediate arousal, attention, perception, concepts, memories, and physiologic change, but no single system to which we can point and say "here lies emotion." This does not mean that our common folk psychology understanding of emotion is totally misguided – emotions are real phenomena that play a critical role in the concern-processing requirements of biological organisms. However, the fact that our emotions feel unified to us is in reality "nothing but an illusion" (in the same sense that our common folk psychology understanding of a unified memory system is an illusion).

In an attempt to understand the mechanisms of emotions better, we returned to first principles and adopted Tomkins' three requirements for a theory of affect: (i) affect is activated by some general characteristic of neural stimulation, common to both internal and external stimuli and not too stimulus-specific like a releaser; (ii) the activator is correlated with biologically useful information; and (iii) some of the activators are capable of habituation, and some capable of non-habituation. To this picture we added Herbert Simon's information-processing view of how two systems as different as cognition and affect could interact, giving us a general set of requirements for the machinery of emotions.

Simon's interrupt theory of emotions provided the conceptual starting point for both Sloman's and Frijda's design-based theories of emotion. Although there are important differences in the details of these two theories (see Wright [97] for a critical review), their common heritage provides enough synergy to easily transfer the language of Frijda's *Emotion Process* on to Sloman's motivated agent framework introduced in section 2.2. Armed with this tool, we were then able to describe the information-processing pathways of the different classes of emotional state within our extended motivated agent framework.

Having identified the basic emotion circuits active in the different emotional classes, we turned to the neurologically-based theories of Damasio and LeDoux to add depth to, and support for, the generality of our approach. The picture that emerges from these neurological models is one of a number of different emotion processing circuits innately "primed" to respond to different types of situations and events. These circuits have common points of entry and exit – receiving their input from the sensory cortex and sensory thalamus, and expressing themselves through the amygdala, hypothalamus and anterior cingulate –, however in between the generality of these two points, lie structures whose activity can be correlated to specific emotional states. Associated with this *machinery of primary emotions* are a number of other brain circuits that respond to and control deliberatively generated (mental) images and events – forming the machinery of secondary (and the sub-class – tertiary) emotions.

Returning to Tomkins' initial requirements we can now explain: (i) how generality of affect activation is achieved by networks in the prefrontal cortex automatically and involuntarily responding to signals arising from the processing of mental images in the early sensory cortex; (ii) the activator is correlated to biologically useful information by utilising the innate primary emotion mechanisms; and how (iii) some of the activators are capable of habituation, and some capable of non-habituation by virtue of the fact that different emotion circuits are active in different emotions – the same also applies to extinction (activators of the primary fear circuit are notoriously hard to extinguish).

The actual phenomena we label as an "emotional" state emerges out of the interaction of a *variable* number of intricately connected cognitive systems, and so it is hardly surprising that the actual attributes of an emotional episode vary greatly from person to person and occasion to occasion – dependent on the suddenness of the initial stimulus, the preparedness of the organism, the importance of the concern threatened or promoted, cultural conditioning, source of stimulus, class of emotional response, and background affective state. However, from an information-level perspective, all emotional states have two things in common: (i) *valence* – which can take the form of *somatic markers*; and (ii) *motivational attitude*.

5.3 Summary

In **parts I** and **II** we argued that the requirements of complex human environments/ scenarios demand a concern-centric approach to autonomous agent design. We also described how the constraints imposed on the real-time concern-processing mechanisms of such multi-layered designs lead to the natural emergence of perturbant states – partial loss of control of attention. From an information-level perspective, these perturbant states are characteristic of the class of mental states we refer to as *tertiary emotions* (as distinct from say the emergent state of "thrashing" in computer systems – *thrashing* has no information-level equivalent to a *loss of control of attentive processes*).

In this chapter we have extended this analysis by presenting an information-level design-based analysis of the phenomena we commonly call emotion. We started by arguing that a lot of the confusion surrounding the term emotion can be attributed to the fact that different theorists focus on different concern-processing mechanisms (*reactive*, *deliberative*, or *self-reflective*) active in the emotion process – this is related to our argument that emotions are emergent mental states. We then extended our analysis by mapping leading cognitive theories of emotions [Frijda 86; Damasio 94; LeDoux 96] on to our motivated agent framework, and identified the different mechanisms active in *primary*, *secondary* and *tertiary* emotional states. Finally, we established the information-level relationship between perturbant states and the machinery of tertiary emotions (Figure 5.2-7), and prepared the groundwork for information-level representations of the machinery of primary and secondary emotions – to be presented in chapter 7.

6 “Emotional” Agents

> *“It has been difficult to define emotions, and this difficulty continues. We will be rash and start this chapter with a working definition of a kind that has been gaining acceptance. It goes something like this.*
>
> *1 An emotion is usually caused by a person consciously or unconsciously evaluating an event as relevant to a concern (a goal) that is important; the emotion is felt as positive when a concern is advanced and negative when a concern is impeded.*
>
> *2 The core of an emotion is readiness to act and the prompting of plans; an emotion gives priority for one or a few kinds of action to which it gives a sense of urgency – so it can interrupt, or compete with, alternative mental processes or actions. Different types of readiness create different outline relationships with others.*
>
> *3 An emotion is usually experienced as a distinctive type of mental state, sometimes accompanied or followed by bodily changes, expressions, actions.”*
>
> – Oatley and Jenkins, *Understanding Emotions* (page 96)

Having introduced emotion theory in the last chapter, we will now look at some of the gains made by “emotional” architectures over the designs described in chapter 3. One argument for introducing “emotions” into autonomous agents is to facilitate the development of life-like characters or more intuitive user/command interfaces. However, as we will show below, there are also more practical reasons for introducing affect into our agent designs.

6.1 Related Work

Before launching into our own design for an “emotional” agent (see chapter 7), there is still more to be learnt from an analysis of related work that addresses the emotion process within the confines of the requirements of autonomous agency and/or “human-like” agent architectures. In this section we will look at a number of designs that attempt to capture at least some of the attributes of the emotion process within an autonomous agent architecture.

6.1.1 Will

Traditional AI has concentrated on designing agent architectures along functional lines based on the internal workings of the solution – i.e. with discrete units for perception, planning, acting, reacting, and meta-reasoning. This leaves the designer with a number of non-trivial integration problems to solve when it finally comes down to combining the disparate units into a contiguous architecture. Will [Moffat and Frijda 95; Moffat 97] takes a novel approach to integration by using relevance evaluation (*primary appraisal*) as a common thread to tie the discrete AI modules together. The resultant architecture is a simple compact design, that also happens to model the emotion process [Frijda 86] described in section 5.2.3 – the message being that emotions are important for both traditional and behaviour-based AI.

Cognitive Elements of Will

Will uses a blackboard architecture to provide global connectivity and allow standard AI modules and techniques to be used in the design. This greatly speeds up the design processes and bases the theory in readily acceptable terminology. The complete Will architecture is shown in Figure 6.1-1.

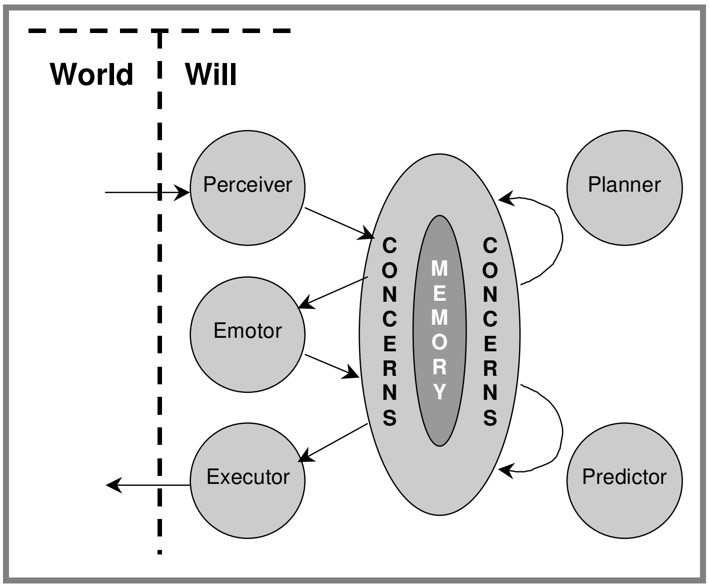


Figure 6.1-1 The Will Architecture [Moffat 97]

Perceiver. Events in the outside world are assigned symbolic meaning and written to the memory as new percepts. Perception could take the form of complex vision processing or simple keyboard entry transcription.

Executor. The executor converts intentions into actions. Both the executor and perceiver are seen as self contained modules, although Moffat points out that this is an extreme simplification.

Concerns. Frijda's information-processing theory of emotion is based around the concept of concerns as dispositions to desire occurrence or non-occurrence of a given kind of situation. The first stage of Frijda's emotion process is primary appraisal, or relevance matching of events with respect to concerns. All events written to memory, whether new, or modified old events, are first matched against existing concerns and assigned a control precedence value, or charge, by the concerns module. The event with the highest charge is then processed by the other modules, in this way charge acts as an attention mechanism and control arbitrator.

Predictor. The second stage of Frijda's emotion process is secondary appraisal or context evaluation. As the context of the situation often depends on what might follow from a stimulus event (i.e. is a threat likely to result in serious injury), a predictor module has been added to the design to make future predictions and inferences from events in memory. The predictor can also be used to predict the consequences of an agent's actions, allowing unpleasant consequences to gain motivational visibility (when written to memory) and thus attract the attention of the planner – see operation below.

Planner. The context of the situation also depends on the coping ability of the agent. If a threat can be easily dealt with then it does not pose a serious problem. Problem solving is the role of the planner.

Emotor. The emotor embodies those parts of the emotion process not covered by the other modules – namely secondary appraisal and action tendency. Secondary appraisal attaches attributes to events such as, valence, un-/expectedness, control, agency, urgency, morality and probability. Action tendencies are the final part of the emotion process and include the general behaviours normally associated with emotional reactions, i.e. approach/avoid and fight/flight, and some goal directed behaviours i.e. try_harder/give_up.

Memory. The memory contains the world model, in the form of a blackboard, and is used to hold factual beliefs and semantic causal rules.

Operation

Events arrive at the perceiver where they are symbolically labelled and written to the blackboard memory, via the concerns module. All events written to memory pass through the concerns module where they are checked for relevance to concerns and assigned a control precedence or charge value. In this way not only current concerns, but future predictions and past events can be evaluated with respect to concerns (which represent the motivators of the agent) – all events therefore have what Moffat calls *motivational visibility*. After processing, results are written back into memory, again via the concerns module. As only the event with the highest charge (called the focus item) is passed to each module as the problem to solve, the value of charge also acts as an attention control mechanism within the memory module.

Events in memory are subjected to an *autoboredom* mechanism which ensures that processed events are not attended to ad infinitum. When a module fails to process an event, the charge is reduced by a fixed percentage. Events that have been processed therefore fall down the importance hierarchy, allowing new events to rise to the top of the stack and be attended to.

Action tendencies are generated by the emotor in response to all the appraisals made on the focus item by the other modules. The action tendency is programmed into the emotor as an appropriate response for the perceived emotion, and written to the memory as an intended action. Unless this action is then modified by another module it becomes the next intention for the executor.

Analysis

Although Frijda and Moffat proposed their architecture as a solution to the integration problems of traditional AI, the theoretical background for Will lies in the more unusual direction of Frijda's [86] *emotion process*. Will is at heart an "emotional" agent architecture – remaining fairly true to Frijda's model of emotion process. However, Will does deviate subtly from the theory in a number of key areas – with not insignificant repercussions on the nature of the supported "emotional" states.

In Frijda's original model, the control precedence signal was seen as a function of relevance evaluation, context evaluation and urgency/difficulty evaluation. In Will the control precedence signal is assigned solely on the basis of relevance evaluation within the *concerns* module – with context and difficulty evaluation being performed separately on the most relevant problem (motivator or focus item) at the start of each cycle. Although events undergo the same evaluations as in the emotion process, the context and urgency/difficulty evaluations do not contribute to a motivator's charge (autoboredom only comes into effect when a module fails to process the focus item) – i.e. there is no active rejection of a motivator on the basis of context and/or urgency/difficulty evaluation.

Will achieves interruption of attentive processing by assigning motivational charge to events in accordance to their match/mismatch against agent concerns – such that an event can become the new focus item of the cognitive modules. Will can therefore be said to partially support core *primary* and *secondary* (as events generated from deliberative modules are also checked for relevance) "emotional" states. However, action tendencies are only generated by the *emotor* module once an event has attained motivational visibility (control precedence), and it is here that Will deviates from the dynamics of the emotion process – hence our qualification of 'partially' in the previous sentence. Certain "emotional" states clearly do require attentive/deliberative context and urgency/difficulty evaluation, but this does not mean that the action tendencies are then generated on the same level as planning or predicting. Will is certainly not attempting to claim that such action tendencies are generated in response to a deliberative appraisal of the type "x has happened, therefore I am in a *happy* state, therefore I generate this action tendency" – unfortunately, this is the message that is likely to be received. One possible solution would be to move the action tendency generation into the *concerns* module, and leave the *emotor* to take care of secondary appraisal.

The decision to simplify the design to a single focus item also has a number of other consequences on Will's ability to generate emergent affective states. As the *emotor* module is only capable of reflecting on appraisals made on the focus item, diffuse states such as moods cannot be supported, and the use of a single focus item makes it hard for the architecture to make comparisons between different events or construct plans that might satisfy more than one concern. Finally, it is hard to see how the architecture can easily be extended to include learning of the type that bootstraps the *secondary* emotion process (see Damasio's somatic marker hypothesis in section 5.2.5). The *concerns* module could of course be extended, to allow learning of emotionally significant events.

6.1.2 Cathexis

Velásquez has proposed a general-purpose emotion-based control framework for autonomous agent architectures called Cathexis [Velásquez 96; 97; 98]. This framework extends the work of Maes and Blumberg (outlined in sections 3.2.3 and 3.2.5) into the affective domain, integrating ideas from [Damasio 94; LeDoux 96; Ekman 92; Izard 92, 93; Johnson-Laird and Oatley 92]. As Velásquez and Breazeal are also members of the "Cog Shop" at MIT (Rodney Brooks' push into humanoid robotics), these ideas may also have relevance to Cog's brand of human-like intelligence.

The basic Cathexis framework is shown in Figure 6.1-2. The sensor, behaviour, and drive systems operate along similar lines to Maes' and Blumberg's spreading activation model. However, the addition of an emotion generation system is unique to Cathexis.

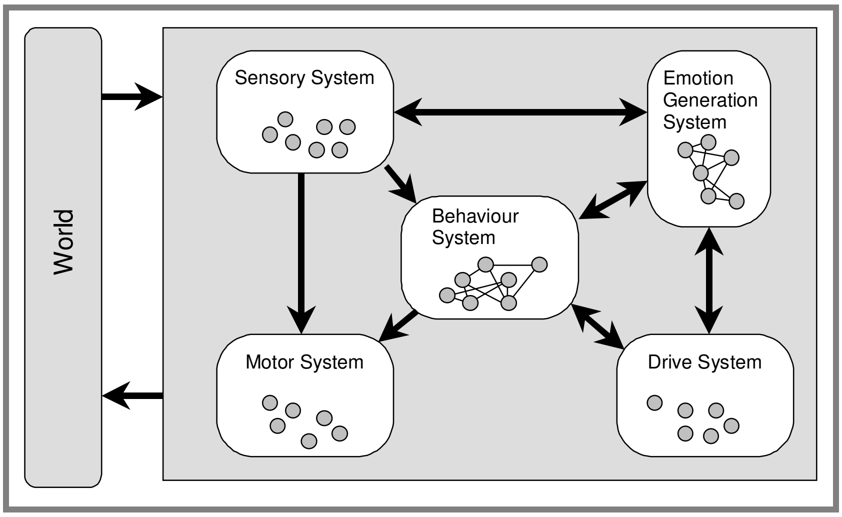


Figure 6.1-2 Cathexis Emotion-Based control Framework [Velásquez 98]

Computational Units

The Cathexis architecture is composed of five different classes of computational units: (i) *Sensory Systems* provide Cathexis with information about its environment; (ii) *Emotional Systems* represent the different families of affective response that provide the main motivational source for the agent; (iii) *Behaviour Systems* represent interconnected self-interested behaviours of the agent – behaviours are non-exclusive, and so non-conflicting behaviours can be executed at the same time; (iv) *Drive Systems* represent urges that impel the agent into action and bias behaviours – "it is both the error signal produced by the *Hunger* drive and the *Distress* caused by it, that motivates the agent to obtain food" [Velásquez 98]; and (v) *Motor Systems* provide the agent with the means to interact with the external environment.

Emotional System

The Cathexis model defines affective states (emotions, moods, and temperaments) as the instantaneous state of a network of specialist emotion systems. The nodes within each system represent the agent's basic needs, and in this sense are similar to Minsky's concept of a "proto-specialist" [Minsky 87, page 165]. Emotion systems are used to represent both the background affective state (mood) and the current emotional state of the agent. Moods are explained as low tonic levels of arousal within the emotion system, whereas emotions are explained by the high arousal levels of a few proto-specialists (inhibition between proto-specialists acts to select clearly defined emotional states).

Initially, secondary emotions were seen as blends or mixtures of these basic proto-specialists – "more than one emotion proto-specialist can be active at the same time, which means that two or more basic emotions may co-occur, representing, as a whole, emotion *blends* and *mixed* emotions. The intensity level and the influences (both in expression and in experience) of each of the active emotions, give rise to these secondary emotions." [Velásquez 97]. However, as the Cathexis framework has evolved, so too has the model of secondary emotions – falling more in line with the cognitive theories described in chapter 5. Damasio's [94] somatic marker hypothesis is used to add learning to the basic Cathexis framework (an approach Cañamero [97] also mentioned as the obvious next step for her Abbott architecture) – "secondary emotions have been modeled with an associative network comparable to Minsky's K-lines [Minsky 86], in which primary emotions are connected to the specific stimuli (e.g., executed behavior, objects or agents) that have elicited them during the robot's interaction with the world." [Breazeal and Velásquez 98] "[T]he purpose of emotional memories is twofold. First they allow for the learning of secondary emotions as generalizations of primary ones. And second, they serve as markers or biasing mechanisms that influence what decisions are made and how the agent behaves." [Velásquez 98]

Analysis

By allowing inhibition to operate between emotion proto-specialists, Cathexis is able to provide a focusing mechanism that acts as a sort of motivational attention. This simple form of motivational attention is then used to amplify the motivational attitude provided by Cathexis' drives – and it is in this sense (à la Tomkins [84] in section 5.2.1; see also Cañamero's [97] Abbott architecture in section 6.1.4) that emotions can be said to provide the main motivational source for the agent. As we have argued in chapter 5, emotions *are* important emergent motivational control states (performing a wide range of functions: global alarms; meta-management; motivational sharpening; communication; learning; etc.), but they rarely account for the everyday motivations of an agent (except in unusual cases involving certain classes of mental disorder).

Unfortunately, the emotion systems add to the complexity of the accounting task of the spreading activation network by introducing a new variable into the equation (even if the role of drives is much reduced in Cathexis). Although not explicit within the framework, this complexity can be reduced by thinking of emotions as non-homeostatic drives that mediate emotion behaviours – as distinct from homeostatic drive behaviours. However, there is still the innate complexity associated with balancing the activation energy injected by the emotion systems, behaviour releasers, and drives, along side the problem of modelling urgency using the single dimension of activation energy (it is probably necessary to assume that the mean activation energy of the emotion system is higher than the behaviour network). These problems are inherent in the spreading activation model itself (see section 3.2 for a critique) and not unique to Cathexis.

The Cathexis framework takes a number of important steps towards the instantiation of the emotion theories discussed in chapter 5: (a) the emotion system provides a form of motivational sharpening by modifying the drive system; (b) the injection of activation energy from non-homeostatic drives can act as a global alarm system; (c) K-lines are effective at capturing Damasio's [94] secondary emotions; and (d) communication through emotive expressions is used to drive social interaction – see Kismet [Breazeal and Velásquez 98].

Although Cathexis uses emotion type labels for the individual emotion system nodes, it also acknowledges that many of the attributes of emotions are themselves emergent (as is the case with social interaction in Kismet). Assigning emotion type labels to the discrete cognitive systems that respond to certain emotion eliciting conditions is clearly an over simplification of the real emotion process (we can classify fear as a *primary*, *secondary*, or *tertiary* emotion depending on the particular interaction of cognitive systems in the emotion process – i.e. there is no single discrete fear system that accounts for all the attributes of the phenomena we call *fear*). However, it does provide a useful first step towards clarifying the different role each cognitive system plays in the emotion process – we must wait and see how Velásquez integrates the different nodes in the next instantiation of the framework.

Finally, one potential draw-back of the Cathexis framework is that it forces the designer to capture all emotion types (and classes) within the same concern-processing mechanism – thereby relying on vague concepts such as emotion mixes and blends to account for the more obtuse (and/or cognitive) emotion types. It will be very interesting to see how Cathexis evolves in the future.

6.1.3 CMattie

McCauley and Franklin have proposed a cognitively inspired autonomous agent architecture that has the ability to display adaptive emergent "emotional" states of varying types and intensities [McCauley and Franklin 98]. The core agent architecture is based on pandemonium theory – which was initially proposed by Selfridge [59], but later extended and made more concrete by Jackson [87] (see also [Baars 88; Franklin 95, pages 234-244; and Śmieja 96]) – and is part of the larger ongoing CMattie project (Conscious Mattie).

The Playing Field, The Spotlight and Consciousness

The sports arena analogy provides a powerful metaphor for describing the global workspace theory of consciousness [Baars 88] on which CMattie is based (see Figure 6.1-3). Inactive codelets (or demons in pandemonium theory terminology) sit in the stands watching the game being played on the playing field below. When something happens on the field that is relevant to the competence of a codelet, it gets excited, and if it shouts loud enough is allowed to go down on to the field and take part in the action. When a codelet joins the field it forms an association with the other codelets already on the playing field, and in this way coalitions of codelets are gradually established. These coalitions represent the concepts/tasks currently being processed by CMattie, the action on the playing field represents the unconscious cognition process, and the coalition currently under the spotlight represents the concept/task that has reached consciousness.

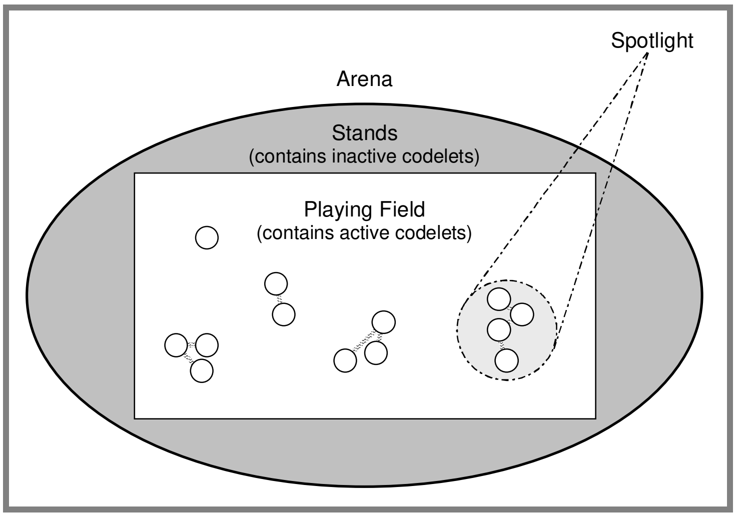


Figure 6.1-3 CMattie's Playing Field [Bogner 98, page 36]

The CMattie architecture (see Figure 6.1-4) contains a number of different types of codelet generators, a behaviour network, a drive mechanism, a perceptual learning mechanism, numerous types of memory, and a meta-cognition mechanism – all in all a very sophisticated autonomous agent. The behaviour network is based on the Maes/Blumberg spreading activation architecture (see sections 3.2.3 and 3.2.5), and is used to select actions to perform in the agent's environment (through email). CMattie's homeostatic concerns are represented as drives in the conventional way, whereas non-homeostatic goal-based concerns are supplied by the coalitions of codelets on the playing field.

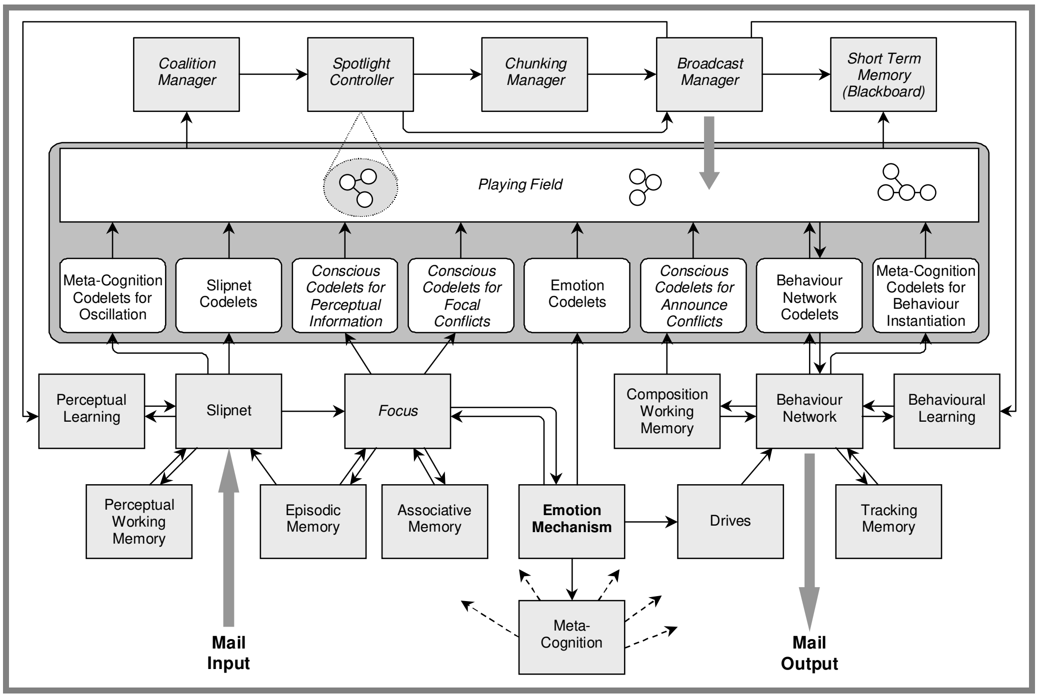


Figure 6.1-4 CMattie's Architecture [Bogner 99, page 58]

CMattie's emotion mechanism plays many different roles in the architecture. Emotion can: (a) alter the current level of the drives that feed into the behaviour network; (b) affect the focus of episodic and associative memory; (c) influence the suggested actions of sparse distributed memory; (d) affect meta-cognition making CMattie more/less reactive; and (e) determine the strength of associations between codelets on the playing field.

Emotion Codelets

Within the goal-based cognition process, emotion codelets are used to provide an assessment of the desirability of the current situation and set the gain that determines how the coalition link strengths are updated. CMattie's gain is a vector of four real numbers roughly analogous to the four emotions: anger; sadness; happiness; and fear. The agent's emotional state is therefore considered a combination of these *basic* emotions. When an emotion codelet's preconditions are met it fires, creating an instantiation of itself with any necessary arguments unified for that particular moment. The instantiated codelet then modifies the element in the gain vector associated with its emotional class. This is a two step process:

1) The intensity of the emotion codelet is calculated to include valence, saturation, and repetition according to the formula in Figure 6.1-5a.
2) Each emotion codelet that fires creates an instantiation of itself with the current value of adjusted intensity. This new codelet enters the playing field and performs actions by adding its adjusted intensity value to the element in the gain vector that represents its particular emotional class (using the formula in Figure 6.1-5b).

a)

$$a = v\frac{1}{1+e^{(-vx+x_0)/1.5}}$$

where:
a = adjusted intensity at creation time
x = initial intensity of the emotion
v = valence {1,-1}
x_0 = shifts the function left or right

b)

$$y = ae^{-b(t-t_0)}$$

where:
y = contribution to global gain vector
a = adjusted intensity at creation time
b = decay rate of the emotion
t = current time
t_0 = time at creation of the codelet

Figure 6.1-5 Emotion Intensity Calculation [McCauley 99, page 30]

Analysis

CMattie's "emotional" state is seen as emergent in the sense that "[n]o combination of emotions are preprogrammed; therefore, any recognizable complex emotions that occur will be emergent" [McCauley and Franklin 98]. Emotions emerge from the action of the four basic emotion types represented by emotion codelets – which in itself is not dissimilar to the Cathexis [Velásquez 96; 97; 98] approach involving emotion mixes and blends (described in section 6.1.2). The emotion codelets themselves are *not* instantiated emotions, they are instantiated emotion circuits. This is an important distinction that can easily be overlooked – *emotion* is an emergent property of the system, and can not be reduced to the presence (or absence) of a collection of active emotion codelets. Emotion is above all a process (see chapter 5).

Although it is a little tricky to map CMattie on to our motivated agent framework, we can still provide some useful analogies that help to clarify the emotion process within the architecture.

a) Emotion codelet generation represents the *relevance evaluation* of the emotion process within the agent (where situations and events are matched against agent concerns). We can think of CMattie as having four relevance evaluation circuits.

b) The gain represents CMattie's background affective state (as opposed to emotional state – i.e. there is no interruption of attentive processing). We could almost say that gain is instantiated "mood" and emotion codelets instantiated short-term affective states.

c) Emotion codelets themselves represent a mix of motivational control states analogous to the actions of hormones or neuromodulators on the one hand, and somatic markers on the other.

Emotion is affect that captures attentive processing, and so a fair approximation of "emotion" within CMattie would be affect (i.e. emotion codelets) that captures the spotlight of consciousness. This does indeed happen when emotion codelets join coalitions and inject their activation energy into the coalition in order to attract the spotlight of consciousness – with the resultant interruption of attentive processing representing an emergent "emotional" state.

In the present arrangement, CMattie's emotion codelets are generated in response to events in the perception register (the current perceptual focus). CMattie therefore already supports the machinery of *primary* emotions (see section 5.2.3). Emotion codelets can also be triggered by remembered past experiences, thus providing the machinery for Damasio's [94] somatic marker *secondary* emotions. If we were to extend the architecture to allow relevance evaluation (through emotion codelets) of the cognition process itself (a meta-cognition task), then we could also provide the machinery for the full class of *secondary* and possibly *tertiary* emotions.

As with any spreading activation model, there are a number of accounting problems that would need to be addressed. As the coalition's activation level is deemed the average of all the codelets in the coalition, large coalitions are immediately disadvantaged as the contribution from an emotion codelet would effectively be diluted amongst the coalition members – this is not a problem in general cognition as each codelet brings its own activation energy into the equation. There are also likely to be problems in determining how much energy a new emotion codelet injects into each coalition already on the playing field, and how emotion codelets are combined.

McCauley and Franklin [98, page 4] ask themselves "does the spotlight of consciousness ever shine on the emotion mechanism or on an emotion codelet? Here the answer is no, not because it would be difficult to make it happen, but because we've found no justification for doing so." One possible justification would be that emotion codelets provide a source of both *valence* (the output of the four relevance evaluation circuits) and *motivational attitude* (activation energy) to the system. Allowing the presence of emotion codelets to be detected within coalitions would place CMattie in a better position to respond to urgent requests and/or situations that were known to be relevant – even if the relevance could not be represented by specific codelets within the coalition. Affect adds generality and urgency to problems – knowing that something is good, bad, or important, is a useful first step to problem solving.

6.1.4 Motivated Society of Mind

Cañamero has proposed a *Society of Mind* (SoM) [Minsky 85] approach to the action selection problem using *Motivation* agents to co-ordinate and organise the behaviour of the society [Cañamero 97]. The SoM is collectively called Abbott[1] (or Abbotts as a species). Abbotts inhabit a dynamic and unpredictable two-dimensional environment called Gridland – similar in nature to Tyrrell's Simulated Agent Environment (see section 3.2.4). Details of the environment are discussed in section 6.2, where we describe two implementations of the Abbott architecture.

Abbotts are constructed within a subsumption style framework, where more complex behaviours are implemented by adding agents to the society without modifying existing society members. Abbotts are endowed with primitive motivational states – *impulses to action based on bodily needs* – and "emotions" – *peripheral and cognitive responses triggered by the recognition of a significant event*. "Emotions" perform an alarm/meta-management function in Abbott, releasing chemicals (hormones, neuromodulators, and neurotransmitters) to alter both the perception of external stimuli and the activation levels of Abbott's internal motivational drives.

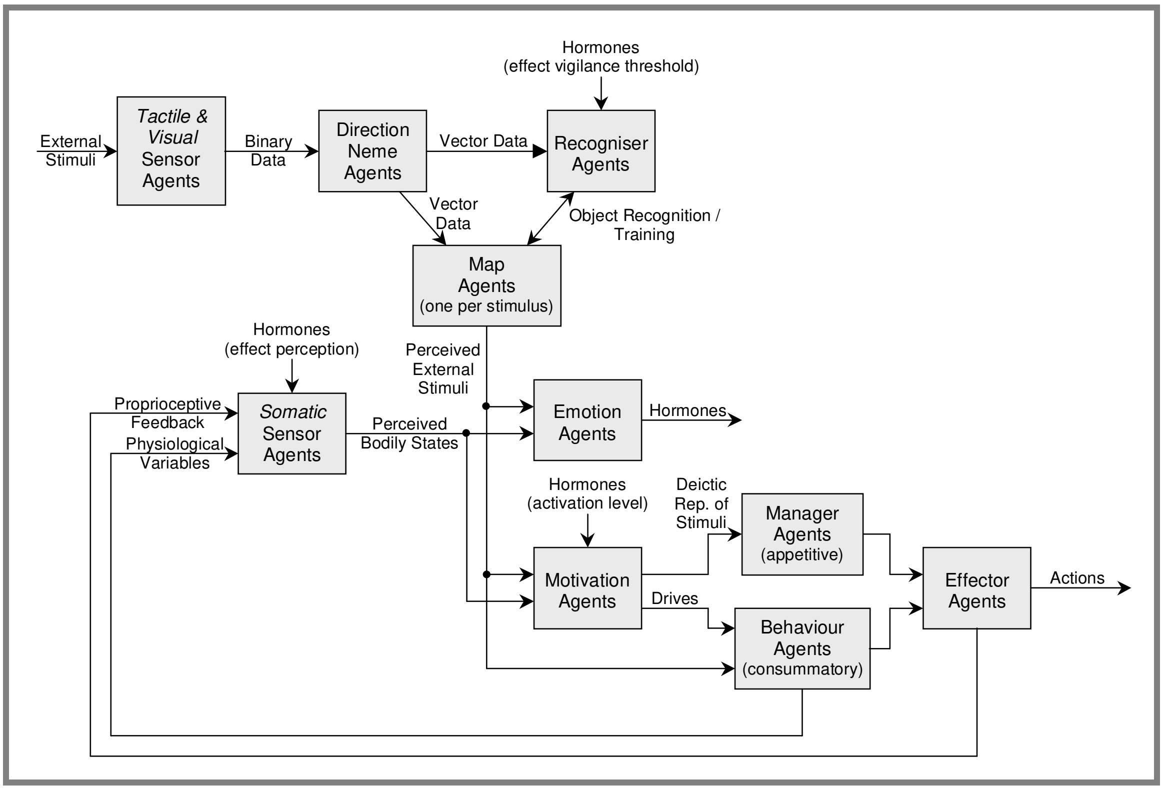


Figure 6.1-6 The Abbott Architecture

An Abbott is comprised of a society of nine different types of agent (see Figure 6.1-6): (i) *Sensor* agents provide Abbott with information about its environment – somatic sensor agents provide information about aspects of Abbott's internal environment (i.e., its physiological variables), tactile and visual sensor agents provide information about the external

[1] Named after Edwin E. Abbott, whose novel *Flatland* [Abbott 1884] inspired some of the features of the Gridland Scenario [Cañamero 97, page 1].

environment; (ii) *Direction neme* agents take the output of sensor agents and transform it into vector data about a particular direction or region in space; (iii) *Recogniser* agents use the vector output of direction neme agents to identify objects in the environment; (iv) *Map* agents communicate with direction neme and recogniser agents to produce stimuli for motivations and behaviours; (v) *Behaviour* agents implement the consummatory behaviours of Abbotts; (vi) *Manager* agents implement very simple skills which represent the appetitive behaviours of Abbotts (Manager agents – such as finder, look-for and go-towards – respond to the stimuli that other agents tell them to attend to in a form of deictic representation [Agre and Chapman 91]); (vii) *Motivation* agents organise an Abbott's behaviours so as to satisfy its motivational drive (hunger, thirst, etc.); (viii) *Emotion* agents amplify or modify the motivational state of an Abbott and its perceived bodily state (i.e., the "happiness" agent releases endorphine which reduces the perception of pain); and (ix) *Effector* agents perform actions in the outside world.

Action Selection

Abbott represents an infant whose *purpose in life* is to ensure that a number of physiological variables (including its chemical control signals) are maintained within a desired range. Intelligent action selection is defined as the process of selecting the best behaviour with which to achieve this task at any given moment in time.

The operation of the action-selection algorithm:

1) The activation level of all agents and their activation state are (re)set to zero.
2) Both the internal variables and the environment are sensed, objects "subliminally" recognised, maps built. Not all this information will be attended to by Abbott, but only those pieces that are relevant to its motivational states.
3) Motivations are assessed and the effects of the creature's emotional state computed. The motivation with the highest activation is selected.
4) The active motivation selects the behaviour(s) that can best satisfy its drive – a consummatory behaviour if the incentive stimulus is present, an appetitive one otherwise.

Figure 6.1-7 Abbott's Action Selection Algorithm

Cañamero has adopted a biologically inspired "drives" (see Tyrrell [93a] for a classification, and comparison, of different action selection models) approach to action selection. Behaviours are selected that best match the current motivational state (i.e., contribute to satisfying the drive with the greatest activation level). Figure 6.1-7 shows the action selection algorithm, and Table 6.1-1 shows Abbott's physiological variables – used to define its body state.

Parameter	Initial Val.	Set Point	Range
Blood pressure	12	12	±4
Blood sugar	30	30	±10
Energy	120	100	±50
Heart Rate	75	75	±25
Pain	0	0	±2
Respiration Rate	8	8	±7
Temperature	37	37	±3
Vascular Volume	25	20	±10
Adrenaline	10	10	±5
Dopamine	10	10	±5
Endorphine	20	20	±10

Table 6.1-1 Physiological variables used to define Abbott's Body State [Cañamero 97, Table 2]

Motivational States

Motivation	Drive
Aggression	Decrease adrenaline
Cold	Increase temperature
Warmth	Decrease temperature
Curiosity	Increase endorphine[2]
Fatigue	Increase energy
Hunger	Increase blood sugar
Thirst	Increase vascular volume
Self-protection	Decrease pain

Table 6.1-2 Abbott's Motivations and their Corresponding Drives [Cañamero 97, Table 4]

In general, motivations can be thought of as inferred internal states postulated to explain the variability of behavioural responses that cannot be exclusively accounted for by observable stimuli. In Abbott, motivations are explicitly modelled by agents. Motivation

[2] Strictly speaking endorphine is a pain inhibitor, but it can also be associated with happy/euphoric states. The implication seems to be that when Abbott is happy it goes off searching for novel stimuli.

agents are characterised by: (i) a controlled variable (bodily state); (ii) a set point and a nominal variability range; (iii) an external stimulus (i.e., the presence of food) that can increase/decrease the motivation's activation level, but cannot trigger it; (iv) an error signal or drive; and (v) a satiation criterion.

Emotion Agents

An emotion is an agent that amplifies (or modifies) the motivational state of the creature and its perceived bodily state. An emotion agent is characterised by: (i) an incentive stimulus; (ii) an intensity proportion to its level of activation; (iii) a list of chemicals it releases when activated; (iv) a list of physiological symptoms; and (v) a list of physiological variables it can affect.

As Abbott is defined as an infant, it is allowed to always be in a clear emotional state – which is selected by a winner-takes-all strategy. *Emotion* agents are selected using the following activation/discrimination criteria: (i) External events – an object or the outcome of a behaviour (these events can either be innate as in Table 6.1-3, or memorised); (ii) General patterns of stimulation which provoke different types of changes in physiological variables (i.e., a sustained abnormal high level of any drive activates the *anger* agent) – this corresponds to Tomkins' view of affect activation; and (iii) Particular patterns of physiological variable values. External events have priority in Cañamero's implementation, and can decide the emotional state on their own. If no compatible external events are perceived then the general patterns of stimulation criteria are used. Since general patterns can often result in the activation of more than one *emotion* agent, particular patterns of stimulation are then used to discriminate between *emotion* agents activated by the same general mechanisms (i.e., high heart rate for fear vs. low heart rate for interest).

Emotion	Triggering Event
Fear	Presence of enemy
Anger	Accomplishment of a goal menaced or undone
Happiness	Achievement of a goal
Sadness	Inability to achieve goal
Boredom	Repetitive activity
Interest	Presence of a novel object or event

Table 6.1-3 Innate External Stimuli Triggering Emotions [Cañamero 97]

The selected *emotion* agent influences the action selection mechanism in two main ways: (i) it can increase/decrease the intensity of the current motivation, through the release of chemical control signals; and (ii) it modifies the reading of the sensors that monitor the variables the emotion can affect, therefore altering the perceived body state.

Behaviours

Behaviour agents resemble the competence modules of Maes' Agent Network Architecture [Maes 89] (see section 3.2.3). The intensity with which a behaviour is executed determines the way in which the behaviour contributes to the satiation of the drive. For motor activities the intensity determines the strength of the motor action, for other behaviours, the intensity determines the duration of the behaviour, provided that no other event makes another motivation more urgent.

A *behaviour* agent can only be executed if: (i) it has been selected by the motivational/ emotional state of the creature; and (ii) its incentive stimulus is present. If the stimulus is not present, then the motivational system will either look for another behaviour that can satisfy the current need, or call an appetitive (*manager*) agent that will move the agent towards making the stimulus active.

Stimulus	Motivation	Behaviour	Main Effect
Living being	Aggression	Attack	Decrease adrenaline
Water	Thirst	Drink	Increase vascular volume
Food	Hunger	Eat	Increase blood sugar
Abbott, Block	Curiosity	Play	Increase endorphine
Top flat block	Fatigue	Rest	Increase energy
Free Space	Cold	Walk	Increase temperature
Pain	Self-protection	Withdraw	Decrease pain

Table 6.1-4 Selected Behaviour and Main Effect

Analysis

Abbott offers an interesting and promising approach to the problem of concern-processing in autonomous agents. Abbott exhibits both goal orientated and opportunistic behaviour, and the use of a subsumption-style philosophy (behaviours are added without modifying existing behaviours) in combination with a *Society of Mind* approach allows Abbott to be readily expanded. The inclusion of affect also provides a reward/punishment mechanism to explore learning.

Opportunistic behaviour is achieved by allowing external stimuli to increase/decrease a motivation's activation level, allowing a hungry Abbott to switch from a Hunger to a Thirst motivation as it passes water whilst searching for food. The problem of behaviour persistence is also addressed: (i) *emotion* agents can increase/decrease the current motivation, adding a form of hysteresis; and (ii) *emotion* agents can influence the perception of sensor data making the activation of other motivators easier/harder.

The Abbott architecture also provides a useful framework for integrating the five classes of concern-processing identified in section 4.1 (*safety*, *physiological*, *achievement*, *affiliation*, and *learning*): (i) safety and physiological concerns are handled as drives by *motivation* agents; (ii) achievement and learning concerns are handled partly through the amplification mechanism of affect (or *emotion* agents), and partly through the use of k-line *memory* agents (left by Cañamero for future research); (iii) affiliation concerns are not really addressed in the Gridland scenario, but could easily be integrated using the existing *emotion* agents.

One problem not addressed by the Abbott architecture is the issue of simultaneously pursuing multiple goals. The use of a winner-takes-all attention filter mechanism limits the behaviour selection criteria to the requirements of a single motivator – when in many cases alternative behaviours could be used that satisfy a number of motivators at a time. The use of a *Society of Mind* model allows the architecture to easily be expanded, and a more sophisticated motivator deciding algorithm implemented. We will address this issue in chapter 8.

Finally, a couple of other aspects of the Abbott architecture will be investigated in a modified Abbott architecture to be introduced in the next section. Firstly, Cañamero allows Abbott's affective state to affect perception by altering the vigilance thresholds of the neural-net used by the *recogniser* agent to classify objects. Although this goes some way to model the inaccuracies that might occur when Abbott is highly aroused, it does not address the fact that in such situations Abbott would be more likely to mistake blocks for enemies, than enemies for blocks [Cañamero 97, Behaviours]. Perception is biased not simply by arousal, but by affordance [Gibson 79] – i.e. how the perceived object contributes to the concerns of the agent. Secondly, the functional effect of Abbott's chemical control signals need clarification: (i) they are not only released by *emotion* agents, but also by *behaviour* agents; and (ii) they are not restricted to influencing perception and motivation, but also form part of Abbott's body state and can thus generate drive-based motivations directly.

6.1.5 Conclusions

In this section we have looked at four architectures that address the issue of concern-processing in "emotional" autonomous agents from slightly different angles. Will [Moffat 97] is perhaps the closest to an "emotional" agent (in the sense of being an instantiated theory of emotion), but all the architectures have their merits and make significant contributions to the emotion debate. In our analysis of the different approaches adopted, the main focus has been on the use of the term "emotion" in relation to the emotion process described in chapter 5 (this is one area in which we aim to provide a better explanatory framework for intelligent autonomous agency). With the exception of Will, all the architectures treat "emotions" as the product of discrete "emotion" systems within the agent architecture – and yet paradoxically acknowledge the emergent nature of the emotional phenomena. We have argued that much of this confusion can be attributed to the opportune use of emotion type labels in the architectures – which gives the false impression of *basic* “emotion” nodes/codelets/agents. In the next section we will build on this argument by analysing the role of the “emotion” agents in Cañamero’s [97] motivated *Society of Mind* architecture (within our extended motivated agent framework) – the methodology and arguments can also be applied to Cathexis [Velásquez 96; 97; 98] and CMattie [McCauley and Franklin 98].

6.2 Case Study: Extended Motivated Society of Mind

When exploring the design space of any agent architecture, it is always tempting to set-up a series of experiments in such a way as to allow the performance of the different architectures (see section 3.2.4, and Tyrrell [93a]) to be numerically defined (i.e. survival time). However, the utility of this approach is questionable – at best it will tell us how well a design matches a very artificial niche, and at worst it may lead us down blind allays with little understanding of the underlying mechanisms and their relation to the vast space of possible designs. Our experimental setting will be empirical – based on a series of design implementations and subsequent *information-level* analysis within a simulated agent environment. By analysing designs at the information-level we gain a greater understanding of the interactions of the multiple concern-processing mechanisms. This is the essence of the design-based approach, and is ideally suited to the broad objectives of this research to *elucidate concern-processing in autonomous agents*.

Cañamero’s Motivated Society of Mind (MSoM) architecture (see section 6.1.1) and its Gridland Scenario provide the initial starting point in our investigation of reactive and deliberative mechanisms of concern mediation. The Abbott architecture already captures a number of different concern processes in a unified framework – using the twin approach of homeostatic drives, and non-homeostatic affect amplifiers.

Modified Abbott Architecture

A modified Abbott architecture, along with the Gridland scenario, were implemented in Pop11 using the SIM_AGENT toolkit [Sloman and Poli 96] – extended as part of this research. This new architecture is shown in the style of our motivated agent framework in Figure 6.2-1.

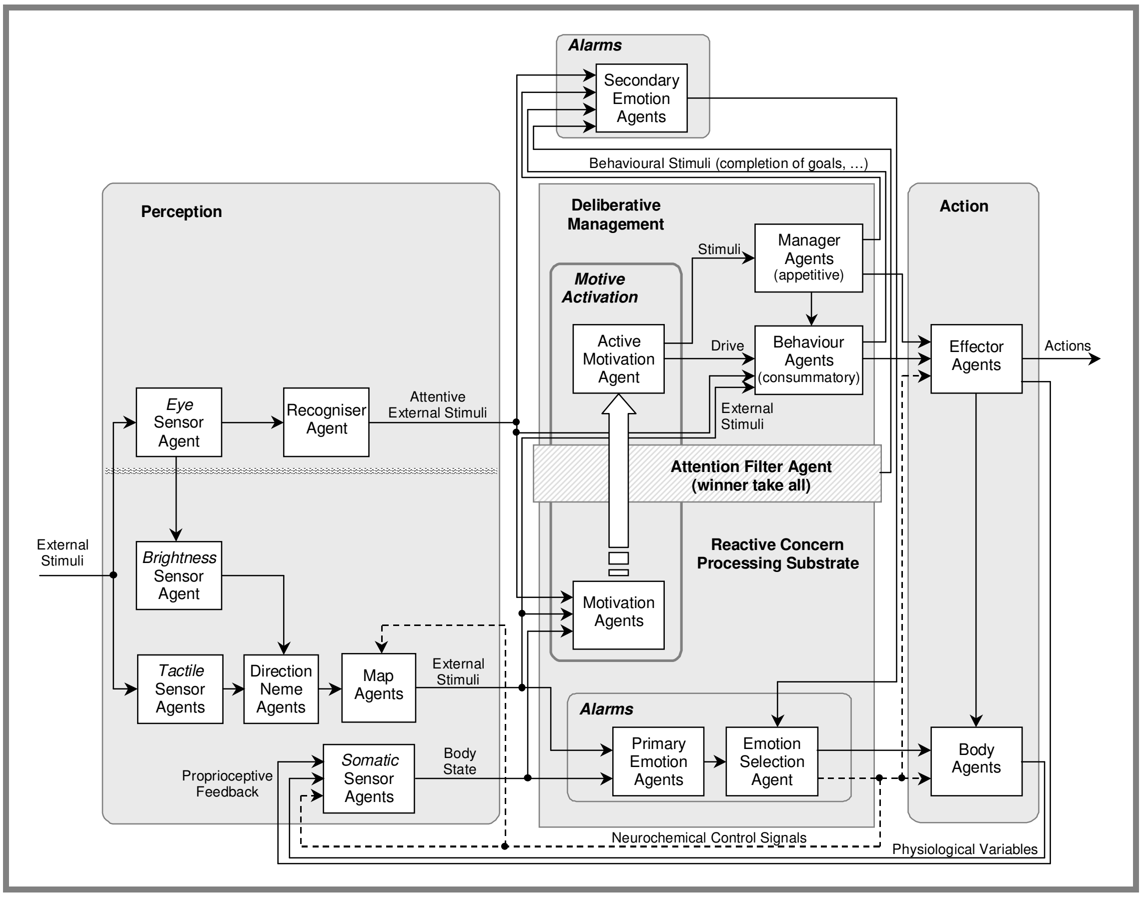


Figure 6.2-1 Abbott2 Architecture

The main modifications to Cañamero's original architecture (see Figure 6.1-6) 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.

A further simplification was made to the architecture by removing the neural-net associated with the *recogniser* agents. As it stood, the neural-net only used arousal to bias the granularity of perception and not for affect-based learning – i.e. somatic markers. Although perception is an important part of concern-processing [Gibson 79] it was felt that the inclusion of the neural-net based *recogniser* agents added little to the overall architecture and abilities of Abbott. The effects of arousal on the perceptual abilities of Abbott are mimicked through an attentive *recogniser* agent which remains distinct from the pre-attentive *map* agents. An aroused Abbott selectively tunes-in to the attended-to object at the expense of the background *map* agents. The results of the *recogniser* agent are made available to the pre-attentive *motivation*, and *secondary emotion* agents.

Motivational Control States

The motivational state of Abbott is a function of three factors [Balkenius 93]: (i) internal drives which tell Abbott about its current needs; (ii) external incentives which tell it about concern objects which are directly accessible; and (iii) internal incentives which tell it about more distant possibilities.

1) *Internal Drives:* An agent has a number of primary concerns that vary dynamically over time. When a concern is not fulfilled, an internal drive signal is generated that increases the probability of the agent selecting actions that serve to fulfil that concern. For instance, one drive could correspond to the need to eat while another drive could make the animal look for predators at regular intervals. Drives can be further partitioned into homeostatic and non-homeostatic drives [Prem 96].

    a) *Homeostatic drives* depend on the deviation of certain values (controlled variables) from specific optimal values, and usually lie between exact boundaries. They are less dependent on environmental conditions or learned features of the agent-environment interaction. Examples are temperature regulation, hunger, sleep, etc.

    b) *Non-homeostatic drives* possess variable optimal values which often strongly depend on learning and environmental variations like triggers or availability. Examples are sexuality, exploratory drive, and emotions.

2) *External Incentives:* At all times, an agent receives sensory input that tells it about the possibility of fulfilling a concern. For instance, viewing or smelling food would constitute an incentive to eat. An external incentive is therefore a representation of the possibilities of the immediate environment given by perceivable concern objects such as food.

3) *Internal Incentives:* Internal incentives play the same role as external incentives, except that they do not directly depend on the currently perceived situation. Instead they are generated by some internal process as a result of prior learning. For example, an expected food situation would make the agent more likely to search for food even if the food is not in sight.

Emotional Control States

In keeping with Cañamero's original scheme, we have associated Abbott's different emotional states with the actions of unique *emotion* agents. This is in stark contrast to the emergent nature of emotions for which we have consistently argued throughout this thesis. However, it serves our immediate purpose of simplifying the emotion process in order to clarify the role emotions can play in intelligent autonomous agent architectures. We will discuss the weaknesses of this approach in the conclusions, and present a new design (which treats emotions as emergent phenomena) in the next chapter.

6.2.1 Implementation Details

Abbott and the agents (enemies, blocks, food, and water) within the virtual environment of the Gridland scenario have been created using the SIM_AGENT toolkit [Sloman and Poli 96] developed within the Cognition and Affect Project. This toolkit allows us to achieve rapid prototyping of our agent architectures in a flexible condition-action rule-based environment, with extensive trace and debug facilities for the isolation and investigation of predicted and unpredicted emergent properties.

Within the confines of this research, we have extended the SIM_AGENT toolkit to provide a graphical user interface and simulated environment capable of supporting both the Gridland and Nursemaid scenarios – the extensions to the standard toolkit are described in appendix A. Figure 6.2-2 shows a screen shot of the virtual environment of the Gridland scenario (see also appendix B).

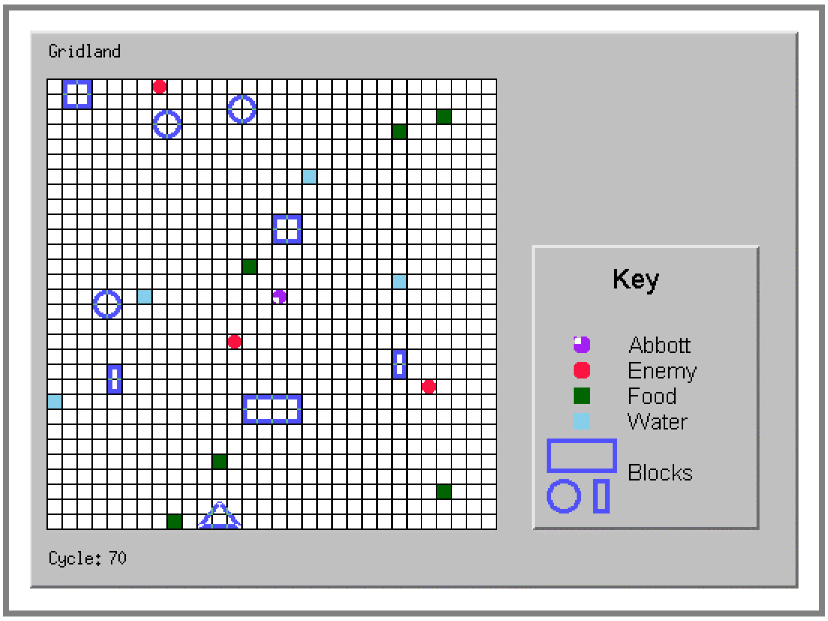


Figure 6.2-2 Gridland Scenario

Society of Mind Model

Abbott is implemented as a *Society of Mind* (SoM), with a single parent, and a number of child agents all of whom share a global blackboard. The parent agent is identified in the Gridland world as “Abbott”, and has a physical presence that can be detected by the other agents that inhabit Gridland. The child agents form the Abbott architecture shown in Figure 6.2-1. The separation into parent and child agents allows us: (i) to separate housekeeping tasks from those of cognition; and (ii) present a unified agent to the outside world.

Multiple Concerns

Abbott is equipped with eight *action tendencies* with which to maintain its body state (see Table 6.1-2): Aggression; Cold; Warmth; Curiosity; Fatigue; Hunger; Thirst; and Self-protection. Each *action tendency* is represented in the architecture by a *motivation* agent which: (i) monitors the status of a single controlled variable (energy, temperature, …); and (ii) selects different behaviours to bring the variable back into range. Life is complicated by the fact that 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. Abbott must therefore balance many competing *concerns*, eventually learning to select actions that address not one but multiple sources of motivation. But before Abbott starts to run, it must first learn to walk and attend to his most urgent needs one at a time.

Proto-Specialists

Minsky uses the term proto-specialist to refer to a “separate agency for each of several basic needs” [Minsky 87, page 165]. The Abbott architecture uses two classes of proto-specialists: *motivation* and *emotion* agents. *Motivation* agents are responsible for monitoring Abbott’s internal body state and responding to specific body needs. However, simply responding to the most urgent motivation at any moment in time inevitably leads to dithering as motivations with similar activation levels compete for control precedence. In Abbott this problem is partially addressed by including an artificial sharpening mechanism that operates on the current active motivation. *Emotion* agents provide this general mechanism to sharpen and enhance Abbott’s sources of motivation: (i) they modify the activation level of the current motivation (through amplification or dampening); and (ii) they change the perception of certain internal variables. In a sense, *emotion* agents take on a motivator management role, identifying urgent and important situations (fear and anger), marking the successful completion of goals (happiness), or detecting situations in which the current strategy is failing and a new approach needs to be adopted (sadness).

The chemical control signals released by Abbott’s *emotion* agents form part of Abbott’s internal body state, and can trigger *motivation* agents directly (aggression and curiosity agents monitor the levels of adrenaline and endorphine respectively) – an angry Abbott can therefore strike out aggressively, or a curious Abbott start searching for novel stimuli. This is obviously an oversimplification of the link between neurochemicals and motivations, but provides an interesting starting point. By relaxing the constraint of a single chemical for a single motivation, it should prove possible for the presence of an enemy to elicit an emotion of fear or anger and, in both cases, lead to motivations of aggression or self-protection –depending on Abbott's previous history and current state. The background chemical mood can therefore represent a snap-shot of Abbott's internal state, and so act as an automatic context evaluation mechanism – if Abbott has been successful in achieving goals it will be in a happy state and therefore more likely to persist with new motivations, if on the other hand things have not been working out then motivations are likely to be less persistent as Abbott becomes depressed.

Like an infant, Abbott's emotional repertoire is very simple, and it is always in a clearly defined emotional state. The emotional state – which can vary in intensity – is selected by the aptly named *emotion selection* agent. Although not implemented in the Gridland scenario, a second role *emotion* agents play is that of communicating Abbott's needs to other Abbotts or caretakers. This communication role is made all the more effective if the emotional states are polarised and easy to identify (a potential justification for simple infant emotions – although it is probably more realistic to speculate that infants have yet to develop the cognitive abilities for more complex emotions). Somatic markers could facilitate this communication need by connecting emotional expression with an internal representation of that expression's concern satisfaction requirements (a role for facial feedback and sound in the somatic marker process).

> "To help their offspring grow, most animals evolve two matching schemes: communication is a two-way street. On one side, babies are equipped with cries that can arouse parents far away, out of sight, or sound asleep – for along with sharpening those signs, cross-exclusion also amplifies their intensity. And on the other side, adults are made to find those signs irresistible: there must be special systems in our brains that give such messages a high priority. To what might those baby-watching agents be connected? My guess is that they're wired to the remnants of the same proto-specialists that, when aroused, caused us as infants to cry in the first place. This leads adults to respond to babies' cries by attributing to them the same degrees of urgency that we ourselves would have to feel to make us shriek with the same intensity. This drives the babies' caretakers to respond to their needs with urgent sympathy." [Minsky 87, page 171]

Attention Mechanism

Abbott uses a simple attention mechanism to direct resources to the most urgent source of motivation. This attention mechanism is implemented as an *attention filter* agent within the *Society of Mind* model. Although the attention filter has a nominal threshold, there is a need to distinguish between the attention filter setting and modifying the activation level of the current motivation. Repetitive activity leading to boredom should ideally only affect the motivation that generated the repetitive activity (or better still the current *manager* agent). Whereas attempting to solve a difficult problem or responding to an urgent motivator should ideally raise the filter threshold to prevent interruption of the current motivation. The attention filter therefore plays a part in two distinct processes: (i) motivation selection; and (ii) motivation management – separating managed and pre-management motivators.

In Abbott's winner-takes-all attention filter strategy, it becomes meaningless to talk about raising or lowering the filter threshold setting – this is done implicitly when the *motivation* agent is selected. As Abbott has no specific mechanism to take on an active motive/behaviour management role, it must remain content with simply modifying the current active motivation. Further, without the ability to distinguish between actively managed and pre-management motivators, all transactions must remain in the common currency of activation energy. The *attention filter* agent can however give a boost to the activation energy of the selected motivation and thus provide Abbott with a motivator persistence mechanism.

Aside from the inability to target specific *manager* and *behaviour* agents, relying solely on neurochemicals for motivator management makes Abbott prone to affective contagion due to the slow decay rate of chemical messengers. This is analogous to the background affective state more commonly referred to as *moods*.

Blackboard Architecture

Abbott is implemented using a global blackboard architecture, although each agent is allowed its own private work area on the blackboard. Agents communicate by posting messages on the blackboard, and can respond to messages directly addressed to them or by eaves-dropping on messages posted between other agents. This communication transparency allows *emotion* agents to easily monitor the progress of the current motivation, detect *manager* agent failures, identify repetitive activity, etc.

There are no restrictions within the architecture on the number of active agents (agents actively under attentive management), but in general the set of active agents will only consist of a single *motivation* and *behaviour* agent with possibly multiple *manager* agents. Agents are activated/deactivated by explicit activation messages posted on the blackboard. These messages originate from the *attention filter* agent. Once activated, an agent can enlist the help of other agents by propagating further activation messages (with an activation energy setting equal to, or less than, its own level – activation energy is not preserved in the transaction). Activation energy levels also act as a common currency to allow *effector* agents to arbitrate between competing commands from *manager* and *behaviour* agents.

Asynchronous Design

Abbott SoM agents run asynchronously to each other and to their parent "Abbott" agent. However, as agents in Gridland can only move at the end of a World time-step (currently five clock cycles), external stimuli and actions are synchronised to the environment and World time. The distinction between internal cycle time and external World time allows us to vary the relative speed with which agents in Gridland process information. On each cycle, all the child agents are processed in a fixed order – *sensor* agents, *direction neme* agents, *map* agents, *motivation* agents, *emotion* agents, *attention filter* agents, *manager* agents, *behaviour* agents and finally *effector* agents. The SoM model also allows us to run different child agents at different rates: (i) by duplicating them in the Abbott processing order (i.e. calling the *sensor* agents again before the *manager* agents are called); and (ii) by specifying a cycle limit for each agent (see [Sloman and Poli 96] – and appendix A).

6.2.2 Experiment 1: Motivational Control of Behaviour

In this first experiment, the role of Abbott's homeostatic motivational control states are investigated in isolation from its *emotion* agents and their associated chemical control signals. The flexibility of the *Society of Mind* approach allows us to remove the *emotion* agents without making changes to the rest of architecture. This implementation of Abbott consists of the following agents:

Sensor Agents

Abbott is equipped with a number of sensors to detect and monitor both its internal state and characteristics of objects in its immediate surroundings. Each *sensor* agent monitors a particular variable/attribute and posts the perceived value on the blackboard at the start of each world cycle. *Somatic sensor* agents report the deviation of their monitored body-state variable from a set-point as an *error signal* for *motivation* agents; the *pain* sensor returns a direction for the source of pain; and the *eye* sensor (see Figure 6.2-3) – under attentive control – can be actively sensed at any time by posting a "sense_eye" message to the blackboard.

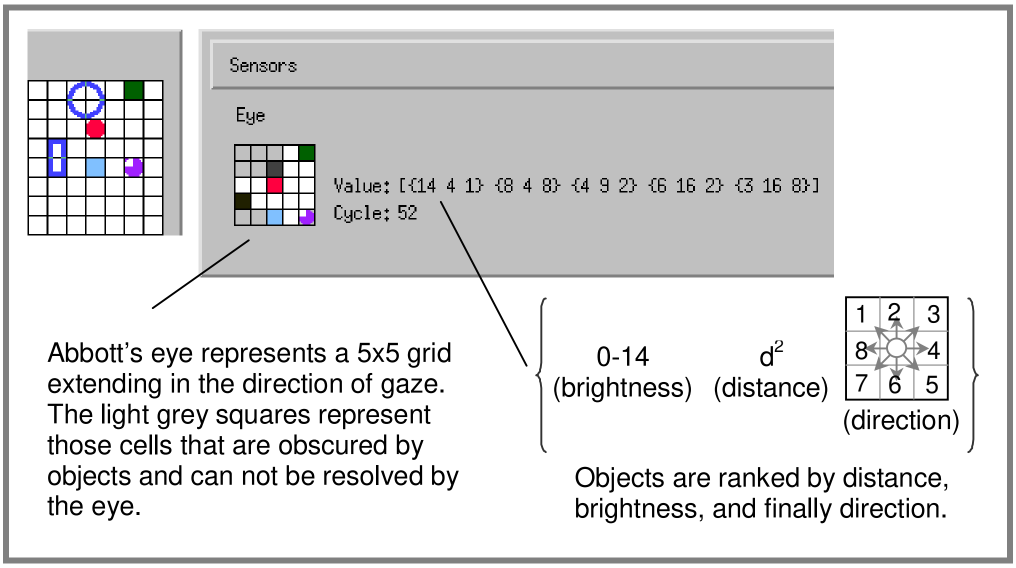


Figure 6.2-3 Abbott's Eye Sensor

Agent	Description of Agent
SensePain	Body state sensor for pain. Returns current intensity, intensity change, deviation from the set-point (*error signal*) and direction of pain.
SenseBloodSugar	Body state sensor of blood sugar level. Returns current level, change and deviation from the set-point (*error signal*).
SenseVascularVolume	Body state sensor of vascular volume. Returns current level, change and deviation from the set-point (*error signal*).
SenseTemperature	Body state sensor of temperature. Returns current level, change and deviation from the set-point (*error signal*).
SenseEnergy	Body state sensor of energy level. Returns current level, change and deviation from the set-point (*error signal*).
SenseOccupancy	Tactile sensor for occupancy of the surrounding grid cells. Returns a value to indicate occupied, partially occupied, or empty.
SenseHardness	Tactile sensor for hardness of objects in the surrounding grid cells. Returns a hardness value between 0 and 15
SenseOrganic	Tactile sensor for organic value of objects in the surrounding grid cells.
SenseBrightness	Visual sensor for brightness of objects immediately next to Abbott in the direction of gaze of the eye.
SenseEye	Visual/Proprioceptive sensor for Abbott's eye. Returns the eye direction and a list of line-of-sight object brightness', distances and directions.
SenseFoot	Proprioceptive sensor for Abbott's foot, giving current heading.

Table 6.2-1 Experiment 1 – Sensor Agents

Direction Neme, Map, and Recogniser Agents

Abbott's *direction neme* agents take the tactile and visual sensor information and perform simple object recognition/classification tasks on their respective cells. The *direction neme* agents are then accessed by the *map* agents to build a simple binary map for each known object in the cells adjacent to Abbott. These two classes of agent represent Abbott's pre-attentive perceptual abilities. Abbott also has a more sophisticated *recogniser* agent which is capable of filtering the *eye* sensor list to identify the particular object that Abbott is attending to – if Abbott is not actually looking for a particular object then the first (i.e. closest and brightest) object in the list is returned as a "percept".

Agent	Description of Agent
DirTopLeft	Extracts visual and tactile object information for the cell to the top left of Abbott.
DirTop	Extracts visual and tactile object information for the cell to the top of Abbott.
DirTopRight	Extracts visual and tactile object information for the cell to the top right of Abbott.
DirRight	Extracts visual and tactile object information for the cell to the right of Abbott.
DirBottomRight	Extracts visual and tactile object information for the cell to the bottom right of Abbott.
DirBottom	Extracts visual and tactile object information for the cell to the bottom of Abbott.
DirBottomLeft	Extracts visual and tactile object information for the cell to the bottom left of Abbott.
DirLeft	Extracts visual and tactile object information for the cell to the left of Abbott.

Table 6.2-2 Experiment 1 – Direction Neme Agents

Agent	Description of Agent
MapOccupancy	Takes direction neme information and builds a binary occupancy map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell is occupied.
MapWater	Takes direction neme information and builds a binary water map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains water.
MapFood	Takes direction neme information and builds a binary food map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains food.
MapLivingBeing	Takes direction neme information and builds a binary living being map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains a living being.
MapEnemy	Takes direction neme information and builds a binary enemy map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains an enemy.
MapBlock	Takes direction neme information and builds a binary block map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains a block.
MapAbbott	Takes direction neme information and builds a binary Abbott map for the cells surrounding Abbott. The map is only posted on the blackboard if at least one cell contains an Abbott.

Table 6.2-3 Experiment 1 – Map Agents

Agent	Description of Agent
RecogniserAttendTo	Returns either the closest object as a "percept", or the object Abbott is attending to at that moment in time.

Table 6.2-4 Experiment 1 – Recogniser Agents

Motivation Agents

Motivation agents have two separate levels of operation: (i) pre-attentive activation energy calculation; and (ii) attentive behaviour selection. At a pre-attentive level, *motivation* agents calculate their current activation level as a function of the error signal (*drive*) produced by *sensor* agents responsible for monitoring their controlled variables. The resultant drives can then be amplified by relevant external stimuli detected by *map* and *recogniser* agents – a drive to increase blood sugar level is amplified by the presence of a food map. Once selected (by the *attention filter* agent), *motivation* agents use the algorithm shown in Figure 6.2-4 to select the *manager* and *behaviour* agents they need to satisfy their drives.

1) *Behaviour* agents whose incentive stimulus is present (i.e. a food map for the *BehaviourEat* agent) post a "stimulus_observed" message on the black-board stating the drives they satisfy. The active *motivation* agent scans this list looking for a behaviour whose primary or secondary effect satisfies the current drive and sends an activate message to the chosen behaviour.
2) If no "stimulus_observed" messages are valid for the current drive, the *motivation* agent posts a "match_drive" message on the black-board. Any behaviour agent that can contribute to satisfying this drive responds by posting the incentive stimulus it needs to accomplish the task. The *motivation* agent then activates the *Finder* agent with the incentive stimulus as its "attend_to" object.
3) If at any point a *behaviour* agent returns a "failed" message the *motivation* agent chooses a new behaviour.

Figure 6.2-4 Abbott's Behaviour Selection Algorithm

Agent	Description of Agent
MotFatigue	Monitors the energy level of Abbott and generates a drive to increase energy if it drops too low.
MotCold	Monitors the temperature level of Abbott and generates a drive to increase temperature if it drops too low.
MotWarmth	Monitors the temperature level of Abbott and generates a drive to decrease temperature if it gets too high.
MotHunger	Monitors the blood sugar level of Abbott and generates a drive to increase blood sugar if it drops too low.
MotThirst	Monitors the vascular volume level of Abbott and generates a drive to increase vascular volume if it drops too low.
MotSelfProtection	Monitors the pain level of Abbott and generates a drive to decrease pain if it gets too high.

Table 6.2-5 Experiment 1 – Motivation Agents

Attention Filter Agent

Abbott's *attention filter* agent uses a simple winner-takes-all algorithm to select the *active motivation* agent.

Agent	Description of Agent
AttentionFilter	Chooses the *motivation* agent with the highest activation level as the *active motivation* agent – implicitly setting the filter threshold level. New motivations must then reach the filter threshold level before being considered for activation. Changes to the activation level of the *active motivation* agent are propagated using an "update" message posted on the blackboard – allowing changes in motivator strength to be propagated to *manager* and *behaviour* agents.

Table 6.2-6 Experiment 1 – Winner-takes-all Attention Filter Agent

Manager Agents

Manager agents represent the appetitive actions of Abbott – used to bring about the necessary incentive stimuli for *behaviour* agents. Abbott's *manager* agents form a strict hierarchy with *Finder* at the top. As soon as the behaviour's incentive stimuli has been detected by the *behaviour* agent, the *motivation* agent deactivates all the *manager* agents – allowing Abbott to take immediate advantage of serendipitous events.

Agent	Description of Agent
Finder	A general purpose agent used to find an incentive stimulus for a *behaviour* agent. *Finder* initially uses *LookFor* to check the immediate surroundings, then *LookForward* to look left and right for the stimulus as Abbott wanders through Gridland. Every few steps a new direction is chosen at random. Once the stimulus has been sighted, *GoTowards* is used to home in on the target.
LookFor	Causes the agent to stop and rotate its eye in a full circle in an attempt to locate the stimulus in the immediate vicinity.
LookForward	Causes the agent to look left and right for the stimulus as it moves through Gridland.
GoTowards	Causes the agent to walk towards the stimulus.

Table 6.2-7 Experiment 1 – Manager Agents

Behaviour Agents

Behaviour agents represent the consummatory actions of Abbott. As soon as a *behaviour* agent detects its incentive stimulus, it posts a "stimulus_observed" message on the blackboard to alert the active *motivation* agent. If the behaviour can contribute to satisfying the current drive it is activated and is then able to issue commands to the *effector* agents.

Agent	Description of Agent
BehaviourEat	Eat some food. Eating increases Abbott's blood sugar level, its energy and temperature.
BehaviourDrink	Drink some water. Drinking increases Abbott's vascular volume, its energy and decreases its temperature.
BehaviourRest	Rest on top of a block. Resting increases Abbott's energy, and allows its temperature to normalise.
BehaviourWalk	Walk around Gridland. Moving increases Abbott's temperature, and decreases energy, blood sugar and vascular volume levels.
BehaviourWithdraw	Withdraw from a source of pain, decreasing Abbott's pain, energy, blood sugar and vascular volume levels.

Table 6.2-8 Experiment 1 – Behaviour Agents

Effector Agents

Abbott has three effectors – a foot, a mouth, and an eye –, each controlled by an agent. Abbott's effectors interact with the outside world by posting messages to the parent "Abbott" agent.

Agent	Description of Agent
Foot	Abbott has eight degrees of freedom which gives him a competitive advantage over the enemies with their four degrees of freedom. To maintain synchronicity, Abbott can only move at the end of a "World" cycle (which is set at 5 internal clock cycles). Moving consumes energy and increases temperature. If Abbott attempts to move to an occupied cell it feels pain.
Mouth	The mouth's only action is to ingest. Abbott can take one bite per World cycle. Eating food increases energy, blood sugar and temperature. Drinking water increases energy, vascular volume and decreases temperature. Biting an enemy causes the enemy pain and reduces its health. Finally, biting empty space causes Abbott pain.
Eye	Abbott's eye can be rotated to any corner to give a 5x5 pixel view of the surrounding grid squares. The eye agent must wait a relaxation period of one cycle before it can move the eye again.

Table 6.2-9 Experiment 1 – Effector Agents

Body Agents

Abbott has a single *body* agent designed to help regulate temperature and simulate effects of resting and moving on the physiological variables that constitute its body state.

Agent	Description of Agent
BodyRegulation	Simulates the effects of the environment on Abbott's internal body state. Abbott has a very simple biological makeup which: helps to normalise temperature; reduces pain; accounts for food and water consumption; and increase energy.

Table 6.2-10 Experiment 1 – Body Agents

Analysis

Abbott performs well in its chosen Gridland environment and easily fulfils the goals of autonomous agency outlined in section 1.3 – (i) handling multiple sources of motivation with limited resources; (ii) having and pursuing an agenda; and (iii) being robust and adaptable in the face of a hostile and uncertain environment. Abbott is able to balance the multiple sources of motivation required to maintain a healthy body state; pursue an agenda to find food, water and blocks; respond opportunistically to food and water sources it finds on its travels; negotiate a path around objects; and change behaviours/motivations in response to attacks by enemies, and the spontaneous appearance of food or water (once exhausted a food/water source will regenerate itself at a random position in Gridland). Abbott is also capable of some surprisingly complex behaviours – in order to rest, Abbott must stop on top of a block, in a world in which Abbott has no concept of top to pass to *manager* agents. Abbott gets around this problem by passing the *Finder* agent the stimulus of block, but at the same time biasing the *GoTowards* agent to select upward directions when the path to the stimulus is blocked (possibly by the stimulus itself). This ensures that upon finding a block, Abbott will attempt to go around it until the *BehaviourRest* agent acknowledges that its incentive stimulus has indeed been found.

The Abbott architecture also performs well when measured against the behaviour selection criteria formulated by Tyrrell (see [Tyrrell 93a], and Figure 3.2-11). Abbott is able to deal with all types of sub-problems found in its environment. Activation of drives and motivations are proportional to offsets from optimum points. Abbott prefers consummatory behaviours over appetitive behaviours in the same motivational system, and is capable of taking advantage of opportunistic consummatory behaviour for other motivational systems. There is balanced competition at a motivational and behavioural level. The attention mechanism focuses activity towards contiguous action sequences in a motivational system, but is capable of being interrupted if a more urgent motivational system demands attention. The reactive nature of *motivation* agents ensures that there is no system-level winner-takes-all shutdown of motivational systems. Areas in which this implementation is deficient include:

(i) persistence; and (ii) the need to choose actions as compromise candidates between competing motivational systems.

Finally, running Abbott in the Gridland scenario identified some additional areas of weaknesses and possible improvement to the simple homeostatic drive-based motivational control architecture:

1) If Abbott is tired, and finds a block on the top edge of the Gridland world, it will persist in its attempts to rest on the top of the block even though it is prevented by the physical limits of the world. Unfortunately Abbott has no way of detecting this special case, and will repeatedly attempt to go around the block until its *Fatigue* motivation is usurped by *Hunger* or *Thirst*. Adding a *Frustration* or *Boredom* agent to monitor repetitive activity and instigate a change of motivation is one potential solution – an alternative approach would be to add self-monitoring capabilities to the *motivation* agent to instigate a change of behaviour.

2) Abbott's homeostatic drive-based motivations are unable to respond to specific objects – such as the presence of an enemy – and so Abbott is unable to generate a motivator to avoid an enemies' bite. *Emotion* agents provide one means of translating significant events/objects into chemical control signals that can then be monitored by *sensor* agents to produce the drives needed by *motivation* agents.

3) Motivations with similar activation levels often result in dithering action as Abbott switches between them. Giving priority to incumbent motivations will provide Abbott with a simple persistence mechanism. This can be achieved either through boosting the activation energy of the selected motivation, or making a distinction between actively managed and pre-management motivations – using the attention filter to block new motivators from surfacing.

4) Unfortunately Abbott suffers from a lack of stimulation in its Gridland world, and is often left choosing between very low-level motivations. In principle there is nothing wrong with this mode of operation – assuming some form of hysteresis is added as in 3) above –, nevertheless it would be more efficient for Abbott to continue in a behaviour until something more "urgent" needs to be attended to. In this case "urgency" and activation energy are not quite the same thing. The simple compromise of defining urgency as an activation energy above a certain threshold would result in more efficient behaviour.

6.2.3 Experiment 2: Affective Control of Motivation

This second experiment concentrates on identifying the value non-homeostatic *emotion* agents add to the drive-based Abbott architecture. Abbott is only able to support a limited range of emotion-like states – anger, fear, happiness, and boredom – which it uses to fulfil a number of different roles: (i) to generate motivators from external events; (ii) to modify perception – provide context evaluation; and (iii) to perform simple motivator management – switching tasks when current action is not working. The social/communication role of emotions are not considered.

Emotion Agents

Sloman (section 5.1; see also Damasio in section 5.2.2) identifies three classes of emotional state: (i) *primary* emotions triggered by reactive processes in early sensory input; (ii) *secondary* emotions triggered by attentive thought processes; and (iii) *tertiary* emotions which are differentiated from secondary emotions by a temporary loss of attentive control. Abbott does not have a meta-management layer to facilitate the "loss of control" of tertiary emotions, and so we will concentrate on the distinction between primary and secondary emotions. Abbott's *primary emotion* agents respond to simple output from *sensor* and *map* agents – such as a sustained high level of pain or the generation of an enemy map, whereas the *secondary emotion* agents are associated with attentive management processes such as the inability to achieve a goal.

Agent	Description of Agent
PriEmoAnger	Responds to a sustained high level of pain. Releases adrenaline.
PriEmoFear	Responds either to a sharp increase in error of any physiological variable, or the presence of an enemy map. Releases dopamine.
PriEmoHappy	Responds to a decrease in error of any physiological variable. Increases endorphine, and decreases adrenaline levels.
SecEmoBoredom	Triggered by low activation level motivations. Decreases endorphine level.
SecEmoFear	Presence of an enemy "percept". Increases dopamine
SecEmoHappy	Triggered by the accomplishment of a goal or playing. Increases endorphine, and decreases adrenaline levels.

Table 6.2-11 Experiment 2 – Emotion Agents

Emotion Selection Agent

Cañamero's [97] original design called on Abbott to always exhibit a clearly defined emotional state, but we would also like it to exhibit emotional states that vary in both duration and intensity. Abbott's emotional state is defined by a simple winner-takes-all selection algorithm. This allows Abbott to adopt a neutral-like state as the activation level of the selected emotion drops to zero. Emotional persistence is achieved by adding an emotion filter, and setting the filter threshold slightly higher than that of the active emotion. Abbott must also balance emotional states triggered by one off events with those caused by the continuing presence of a low intensity stimulus. Adding a relaxation constant to the emotion filter helps to achieve this balance.

Agent	Description of Agent
EmotionSelection	Selects the emotion agent with the highest activation level. Incorporates a simple emotion filter to add hysteresis and accommodate one-off and low intensity emotional stimuli.

Table 6.2-12 Experiment 2 – Emotion Selection Agent

Sensor Agents

The introduction of neurochemical control signals and physiological variables of blood pressure, heart rate, and respiration rate need supporting *sensor* agents.

Agent	Description of Agent
SenseBloodPressure	Body state sensor of blood pressure level. Returns current level, change, and deviation from the set-point (*error signal*).
SenseHeartRate	Body state sensor of heart rate. Returns current rate, change, and deviation from the set-point (*error signal*).
SenseRespirationRate	Body state sensor of respiration rate. Returns current rate, change, and deviation from the set-point (*error signal*).
SenseAdrenaline	Body state sensor of adrenaline level. Returns current level, change, and deviation from the set-point (*error signal*).
SenseDopamine	Body state sensor of dopamine level. Returns current level, change, and deviation from the set-point (*error signal*).
SenseEndorphine	Body state sensor of endorphine level. Returns current level, change, and deviation from the set-point (*error signal*).

Table 6.2-13 Experiment 2 – Sensor Agents

Motivation Agents

Maintaining the principle of homeostatic drive-based motivations, the addition of the adrenaline and endorphine neurochemicals facilitates the introduction of two new motivators.

Agent	Description of Agent
MotAggression	Monitors the adrenaline level of Abbott and generates a drive to decrease adrenaline if it gets too high.
MotCuriosity	Monitors the endorphine level of Abbott and generates a drive to increase endorphine if it drops too low.

Table 6.2-14 Experiment 2 – Motivation Agents

Behaviour Agents

The new motivations have associated behaviours that can satisfy their drives.

Agent	Description of Agent
BehaviourAttack	Behaviour to attack a living being. Attacking decreases Abbott's adrenaline level.
BehaviourPlay	Behaviour to play with blocks. Playing increases Abbott's endorphine level.

Table 6.2-15 Experiment 2 – Behaviour Agents

Body Agents

Finally the *BodyRegulation* agent needs to be modified to cope with the physiological effects of the neurochemicals.

Analysis

Abbott's non-homeostatic emotion system acts independently of its basic homeostatic drive mechanism, but is allowed to influence motivations through the release of neurochemicals. Taken at face value, it is possible for Abbott to be "happy" and "hungry", or "scared" and "curious" at the same time. However, before we can apply such terms we must first acknowledge the limitations of the underlying architecture – Abbott's notion of "happiness" and "hunger" is a far cry from human "happiness" and "hunger".

Abbott's motivational state is defined by the motivation that is being attended to at any given time – Abbott is "hungry" because its *MotHunger* agent has gained control precedence. The distinction is not as easy to make when it comes to Abbott's emotional state. Even though the *SecEmoHappy* agent may have been selected, its actions are limited to increasing the level of endorphine – there is no concept of having gained control precedence. *Emotion* agents operate covertly, the *PriEmoAnger* agent can trigger (via the release of adrenaline) the *MotAggression* motivation agent and thus indirectly gain control precedence – it is therefore debatable whether *emotion* agents represent "emotions" or simply affect.

Taking into account the above caveats, we can now start to identify those society members active in four out of the five classes of concern outlined in section 4.1: (i) agents associated with Abbott's physiological needs (W_{phys}) – *MotHunger*, *MotThirst*, *MotCold*, and *MotWarmth*; (ii) agents associated with safety (W_{safe}) – *PriEmoFear*, *SecEmoFear*, and *MotSelfProtection*; (iii) agents associated with achievement (W_{ach}) – *PriEmoAnger*, and *SecEmoBoredom*; and finally (iv) agents associated with learning (W_{learn}) – *SecEmoHappy*, and *MotCuriosity*. The fifth motivational class – affiliation (W_{aff}) – is not supported by the Gridland scenario. Although much still needs to be done, we are already laying the ground work for an agent that is capable of adapting its own motivational profile to better match that of its environment.

Aside from adding to Abbott's motivational repertoire, *emotion* agents also provide a mechanism for monitoring the effectiveness of different behaviour strategies. Abbott's drive-based motivational system only responds to changes in physiological variables that lead to the explicit generation of a motivator. There is no explicit feedback to allow Abbott to learn the associations between actions and drive satisfaction – the feedback loop to say when a drive has been satisfied is implicit in the absence of a motivator. Adding *emotion* agents allows us to make this feedback explicit. Abbott's *SecEmoHappy* agent releases endorphine which can be used to enhance the connections between *motivation* and *behaviour* agents, allowing our agent to adapt behaviours to motivations.

A related problem is that of adapting motivations to the environment – i.e. how do decide the relative importance of different sources of motivational stimuli. In an environment in which food is plentiful and water scarce it would be more advantageous to assign a higher priority to thirst-based motivations than food-based motivations. Monitoring Abbott's level of distress could allow the architecture to sensitise those motivations active when Abbott becomes distressed, and thus respond earlier to distress causing situations.

Finally, running the new Abbott in the Gridland scenario identified some additional areas of weaknesses and possible improvement to the control architecture:

1) Abbott's biological model is very primitive. Before any attempt can be made to realistically implement the actions of neurochemicals on learning and action readiness significant improvements need to be made in the area of cause and effect of neurochemical messengers. The biological model must also take into account automatic/reflexive biological regulation – an animal is likely to pant when hot, reducing temperature at a cost of increased water consumption.

2) There are many situations when motivators need to be generated for events that are not directly linked to biological drives – i.e. the presence of an enemy, or the achievement of a plan or goal. These motivations address Abbott's non-physiological concerns (safety, achievement, learning, and affiliation). At present the Abbott architecture can

only support non-physiological concerns in an indirect manner through intermediary physiological drives (adrenaline, endorphine, etc.).

3) Abbott's early sensory perception operates on simple binary representations of events – an enemy is either present or absent. By using the intermediary stage of neurochemicals, Abbott is able to translate these binary events into analogue drives. However, without a mechanism for introspection, Abbott must rely on interruption by *motivation* agents as the sole means of communication between the affect and cognition systems. Abbott has no analogue to arousal (see [Simon 67]; also section 5.2.2) in its affect/cognition interactions.

4) The current Abbott architecture roughly conforms to the classic BDI (belief, desire, intention) model discussed in section 3.1 [Bratman 87] – *map* agents and percepts form Abbott's *beliefs* about its environment; *motivation* agents can be thought of as Abbott's *desires*, and the active *motivation* agent is Abbott's current *intention* – its commitment to a particular course of action. However, Abbott's *emotion* agents fall outside this strict framework. Although *emotion* agents are related to motivations (*desires*), they neither form new *intentions*, nor are necessarily consistent with Abbott's current *intention* structure. Part of an *emotion* agent's role is to redirect current intentions towards new/important events – this idea needs further exploration through the development of Abbott's motivator management mechanism.

6.2.4 Conclusions

In this section we have demonstrated the benefits of motivational sharpening "emotional" mechanisms in drive-based ethologically inspired agent architectures. The addition of an "emotion" system alleviates many of the problems we identified with reactive behaviour-based architectures in section 3.2.

a) Persistence: *emotion* agents boost the activation level of current motivation. This has a motivational sharpening effect which leads to increased persistence of behaviours that satisfy that motivation.

b) Complexity: *emotion* agents provide a mechanism for non-homeostatic concern mediation, eliminating the need to inject (and balance) activation energy into a behaviour network from sensors.

c) Interruption: *emotion* agents focus on a few primary concerns, acting as a global alarm system – through motivational amplification – when a situation/event matches one of these concerns (this is in addition to opportunism achieved by allowing external stimuli to increase/decrease a motivation's activation level).

These "emotion"-based benefits are gained *in addition* to the more general advantages enjoyed by Abbott's concern-centric design stance. The recognition of motivational control states simplifies the accounting problem associated with spreading activation models – behaviours are now selected to satisfy a small sub-set of active motivations rather than attempting to simultaneously satisfy all the sources of motivation in one go. The architecture can avoid the problems of information-loss associated with *shutting down all-but-one system* (see section 3.2) by varying the number of motivations under attentive consideration at any one time (see chapter 4) – it is also possible to add extra levels of competence to consider more complex sources of motivation in situations that do not require immediate action. Finally, as we will see in the next chapter, Abbott can support many different concern-processing strategies – including motivational attitudes towards goal completion, or simply goal sufficiency (see section 5.2.2). In this sense, Abbott can be said to acknowledge the limits of bounded-rationality (or bounded-optimality), which reactive behaviour-based architectures are hard pressed to do.

There are a number of simplifying assumptions made in the Abbott architecture – not least the decision to give Abbott full grown primary and secondary "emotions." As we discussed in chapter 5, emotions are not generated by a discrete emotion module hidden somewhere in the depths of the limbic system, but emerge from the interaction of a number of different systems (that also take part in a number of other cognitive functions). Aside from this cognitive plausibility issue, there is the more pressing concern of the adaptability of such a discrete "emotion" system.

In the next chapter we will address these issues by looking at the raw mechanisms through which *primary*, *secondary*, and *tertiary* emotional states can emerge in a cognitively inspired agent architecture.

6.3 Summary

The emotion process can be viewed as a classic example of an information-processing system geared towards "serving" concerns at all levels of an agent architecture. In this chapter we have presented an information-level design-based analysis of concern-processing in "emotional" agents [Moffat and Frijda 95; Velásquez 96; Breazeal and Velásquez 98; McCauley and Franklin 98; and Cañamero 97]. We also presented an analysis of two complete implementations of an agent architecture that capture and extend Cañamero's original design within our motivated agent framework – as part of our investigation into the deliberative and reactive mechanisms of concern mediation. Finally, we identified a number of areas of weakness in these extended designs, that can in part be attributed to the inclusion of a discrete "emotion" system. These issues will be addressed in the next chapter when we look at concern-processing within an intelligent agent architecture with *emergent* "emotional" states.

7 Towards an Infant-Like "Emotional" Agent

A central theme that runs throughout this thesis is the need to design agent architectures on the basis of a solid *information-level* understanding of an agent's concern-processing requirements. In this chapter we will put theory into practice and describe the design for a cognitively inspired agent that addresses many of the problems associated with traditional deliberative, behaviour-based, and "emotional" architectures discussed in chapters 3 and 6.

7.1 Concern-Centric Design

Abbott3 (see Figure 7.1-1) represents the latest in our series of broad agent designs that address the requirements of virtual information-processing architectures for human-like minds [Beaudoin 94; Wright 97; Complin 97; Sloman 99].

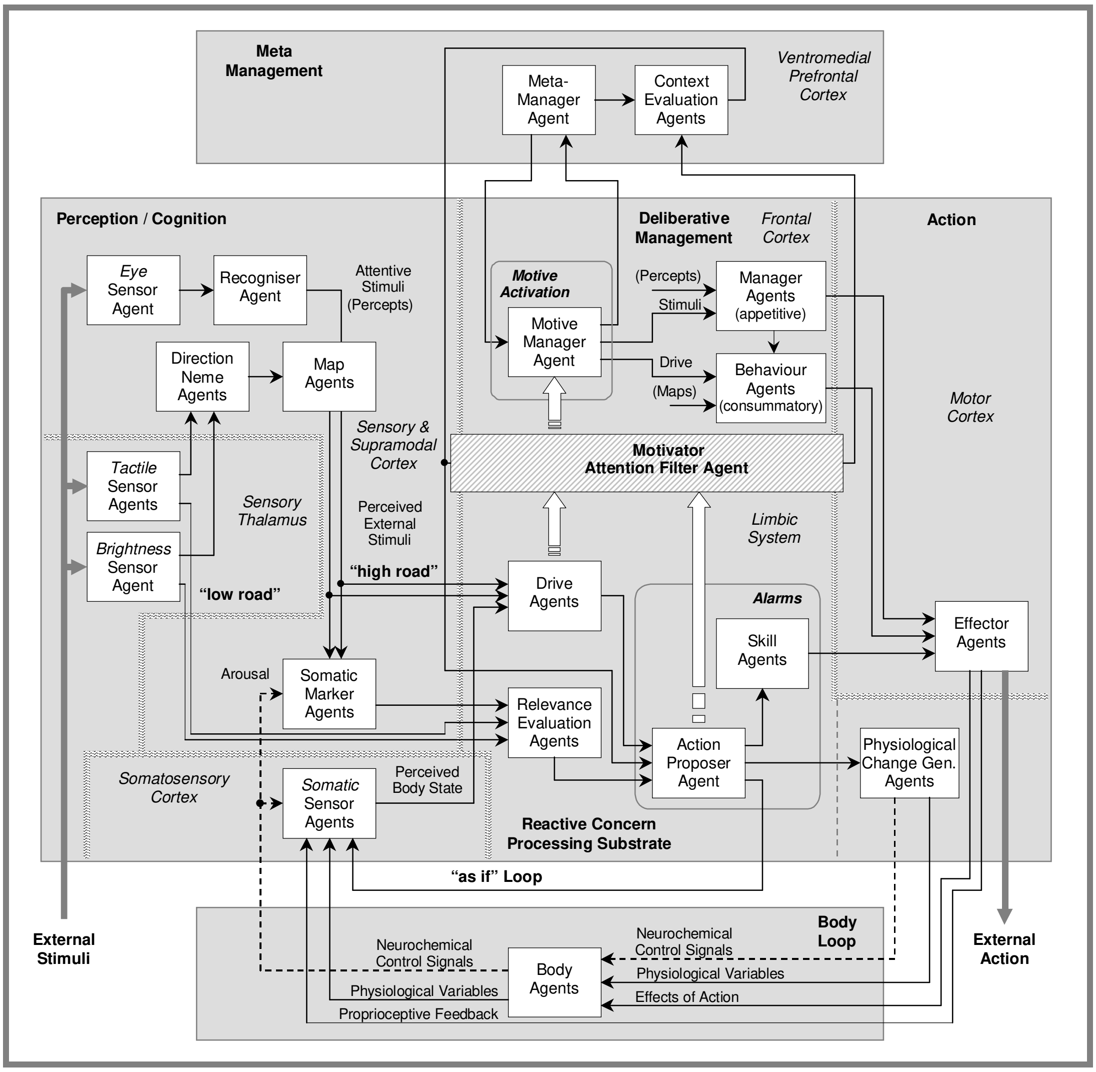


Figure 7.1-1 The Abbott3 Architecture

Our latest incarnation of Abbott marks quite a big departure from Cañamero's [97] original design. However, we feel that it is still appropriate to maintain the convention of using the label "Abbott" to refer to the collective *Society of Mind*. In cases where ambiguity is likely to exist, we will explicitly refer to the two designs as *Abbott3* and *Cañamero's original design*.

In this section we will describe how the competence layers in Abbott's concern-centric architecture co-evolve, and identify the mechanisms that lead to the emergence of "emotional" states. We will keep our descriptions of the design fairly concise, concentrating on the effects of the interactions between the different society members, and building on the concepts presented in earlier chapters – readers are encouraged to refer back to chapters 5 and 6 for more detailed background information.

7.1.1 Co-evolution within Abbott

One of the deficiencies we identified in the subsumption-style architecture [Brooks 86] was the problem of command fusion and the potential for a reversal of concern-processing priorities (see section 3.2.1). A subsumption-style architecture should be capable of being partitioned at any level, with the layers below forming a complete control system. This places an implicit requirement on the designer to capture the primary concerns of the agent in the behaviours of the base layer of the architecture. In Brooks' original proposal, these low-level concern-processing mechanisms are subsequently subsumed by the more specific behaviours represented in the higher-level competence layers. This means that unless the primary concerns are then replicated at each level, there exists the real danger of overriding these high-priority concerns with lower priority concerns as the architecture evolves.

We address the problem of command fusion by adopting a concern-centric design stance that recognises the need to allow the different levels of competence (coping strategies) to co-evolve. In the following discussion we will describe this evolution process as we grow our agent architecture from the base level Abbott3a (Figure 7.1-2) into the fully fledged Abbott3 (Figure 7.1-1). Here our use of the term "evolution" does not mean that we create our architecture through some process of genetic mutation, but that the addition of competence layers is comparable to the evolution of an organism in nature (and distinct from the development of skills or behaviours during the life-time of a single agent) – at each stage of this evolution process Abbott is a fully functional agent architecture. The term "co-evolve" is then used to indicate that as the architecture as a whole evolves, so too do the individual competence layers – facilitating the grounding of the higher-level competencies in the lower ones.

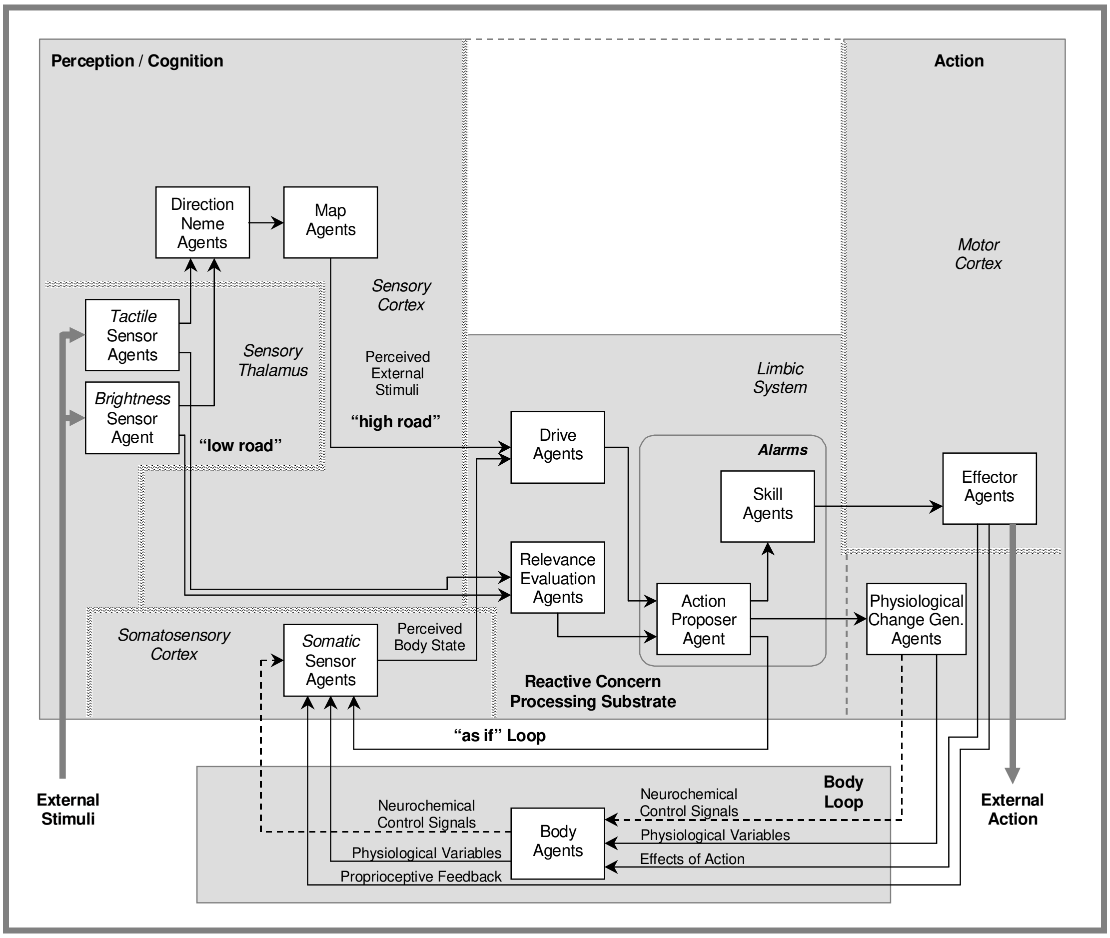


Figure 7.1-2 Abbott3a (base competence level)

In keeping with the vertical decomposition philosophy, Abbott3's base competence level (Figure 7.1-2) can actively sense the environment and respond to its basic needs. Our agent is also able to pursue an agenda (that of ensuring that a number of physiological variables are maintained within a desired range), and can even be said to possess a primitive personality in the form of a motivational profile that can be biased towards self-preservation, eating, or drinking (see section 4.1). We will call our base agent Abbott3a.

Abbott3a has two sources of motivational drive: (a) homeostatic drives – represented by *drive* agents; and (b) non-homeostatic drives – represented by the actions of *relevance evaluation* agents. The *drive* agents respond to error-signals in a controlled variable (i.e. blood sugar level or vascular volume), whereas the *relevance evaluation* agents detect significant features of the environment such as the colour/brightness of objects. These motivational drives are then able to trigger action through *skill* agents and/or generate physiological change through the *physiological change generator* agents. This latter *affective* pathway allows sustained activity to be initiated by one-off external events (partially addressing the persistence problem by providing a primitive motivational sharpening mechanism), and enables Abbott to respond to external incentives. Unfortunately, competence level 0 does not support learning, and Abbott3a is therefore unable to generate motivational control states from internally generated, non-sensor based, incentives (see section 6.2 for a description of the different types of motivational control states supported by Abbott).

Change in physiological arousal affects the perception of Abbott's internal somatic state, and thus allows one motivational drive to inhibit (or enhance) another. Inhibition occurs at the level of agent concerns, and not the individual behaviours as in the more common ethologically inspired behaviour-based architectures [Maes 89; Blumberg 94; Tyrrell 93a] described in section 3.2. We are thus able to implement a primitive attention mechanism to direct behaviour towards satisfying the most pressing concerns – here we are not advocating a system-level 'winner-takes-all' mechanism, but rather a selective mechanism that biases the type and number of concerns attended to at any one time. Concern inhibition acts as a first stage attention filter, with more complex *active* mechanisms evolving as additional competence levels are added to the architecture.

Our level 0 behaviour selection mechanism is self-contained in the *action proposer* agent, with complex behaviour mediation occurring within the deliberative layer (see Figure 7.1-4) of our agent architecture. Using a multi-layered approach eases many of the scalability issues associated with purely reactive designs. At the reactive level we implement skills as "learnt, or innate, action patterns which can be executed with simple perceptual feedback", receiving input from the early *sensory* agents situated in the sensory thalamus (not actually shown in Figure 7.1-2).

Finally, by building on our analysis of the emotion process in chapter 5, we are also able to start to map the various components of the level 0 concern-processing mechanisms on to the corresponding information-processing regions of the human brain – see appendix C.

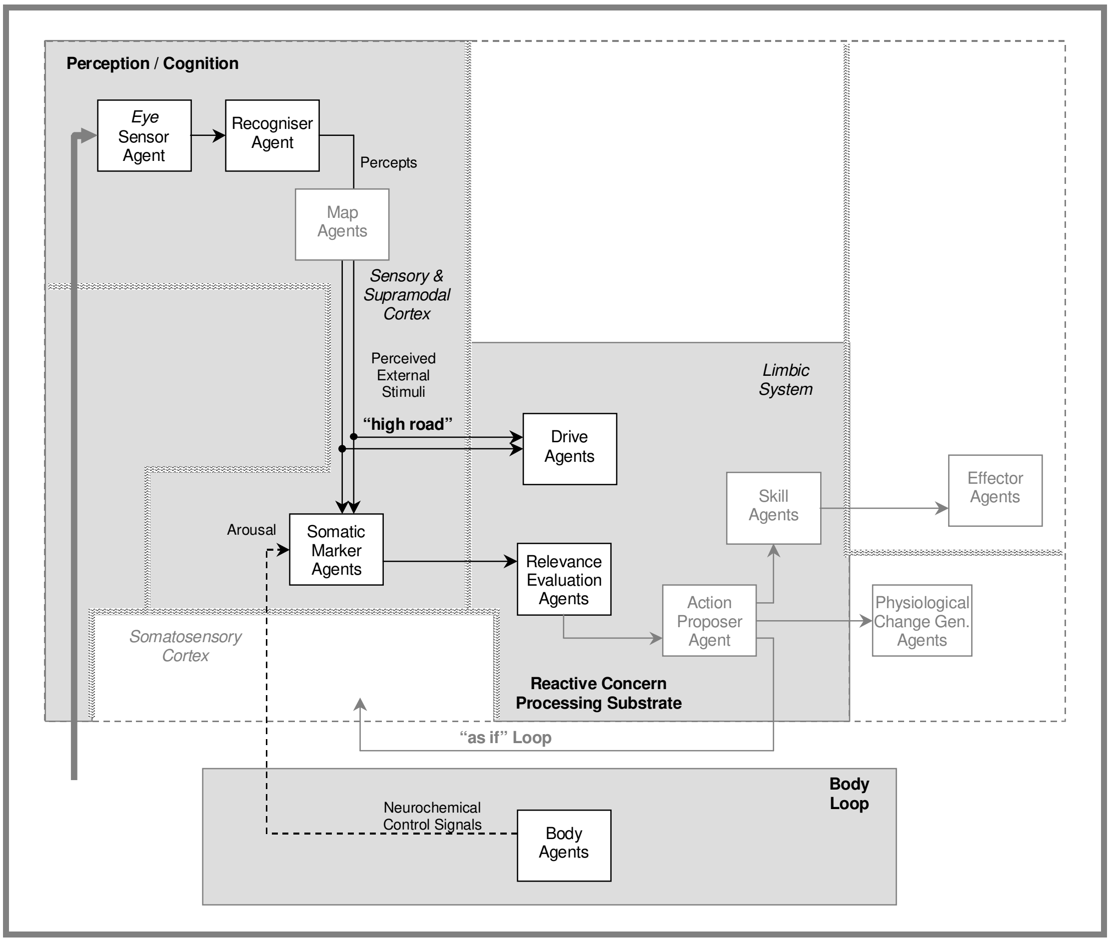


Figure 7.1-3 Abbott3b (showing competence level 1)

Figure 7.1-3 shows competence level 1 of the Abbott3 architecture. These level 1 competences dove-tail into the lower level 0 mechanisms without extensively redesigning the agent – i.e. we adopt the vertical decomposition ethos of the subsumption architecture, but take the practical stance of allowing our levels of competence to co-evolve. We will call our new agent Abbott3b.

The *direction neme* and *map* agents integrate Abbott's sight and touch modalities to produce *percepts* – internal representations of distinct objects in Abbott's immediate environment. These short-range *percepts* are augmented by long-range *percepts* from the *recogniser* agent, and grounded in Abbott's level 0 concern-processing mechanism through *somatic marker* agents (somatic markers are used to mark percepts that are coincident with aroused body states). Abbott is thus able to supplement its innate *primary appraisal* mechanism (relevance evaluation [Frijda 86, page 401]; see also section 5.2.3) with signals generated with respect to learnt "affective" experiences.

The addition of somatic markers, in combination with a primitive attention mechanism, allows our simple agent to exhibit rudimentary *primary* and *secondary* emotions. Objects or events in the external world reach the *relevance evaluation* agent (through the *sensory* agents in the case of *primary* emotions, and via *somatic marker* agents in the case of *secondary* emotions) and cause the action proposer to switch attention and signal a change in the agent's somatic state (through real or "as if" pathways to the *somatic sensor* agents). Although there is no *self* to feel the emotional episode, we have at least met the main criteria normally associated with emotional states – i.e. that of interruption of attention, valence, and motivational attitude. We however attach the label "rudimentary" as there is still no deliberative functionality within these first two levels of competence (Damasio and Sloman attach the label *secondary* emotions to events triggered by deliberative thought processes – see chapter 5). We will discuss this point further in section 7.1.2.

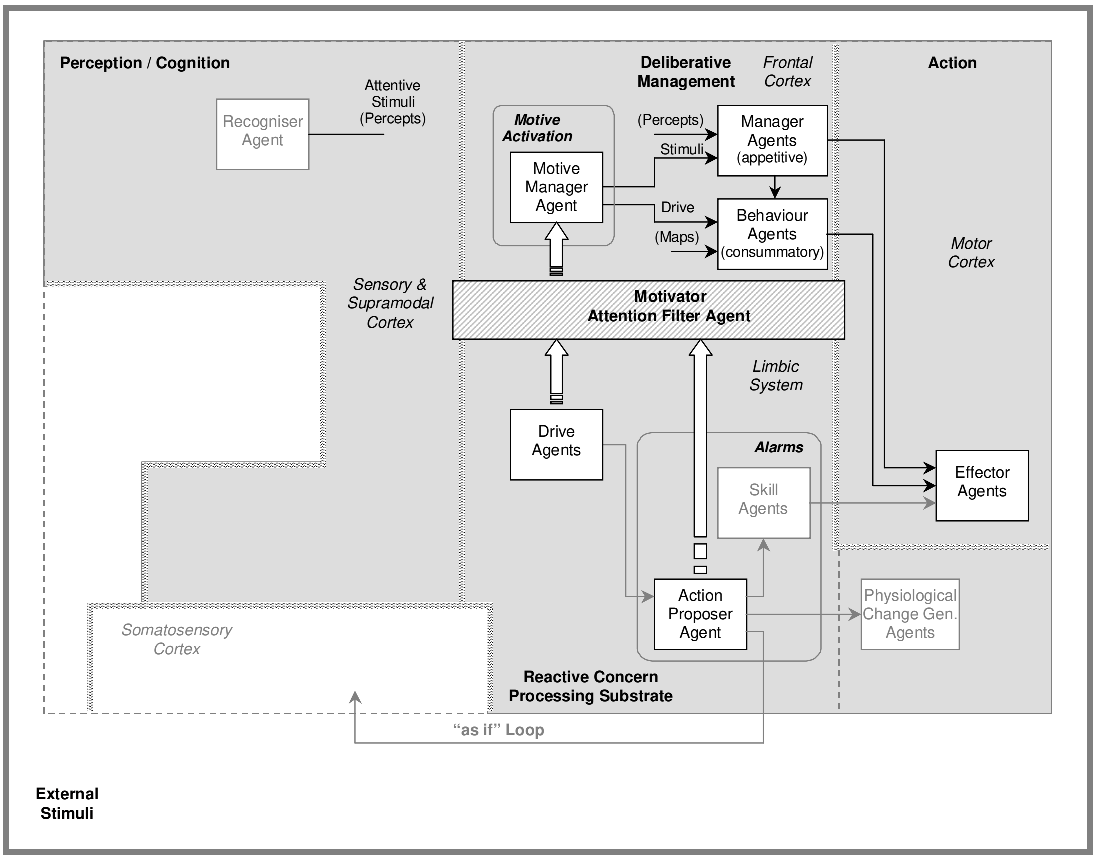


Figure 7.1-4 Abbott3c (showing competence level 2)

Figure 7.1-4 shows competence level 2 of the Abbott3 architecture. Each additional level adds new degrees of competence to the level below, without completely subsuming the original functionality (levels of competence increase the coping strategies available to the agent).

Our simple skill-based action selection mechanism is enhanced by the addition of a deliberative motive management layer capable of actively managing behaviours to meet Abbott's many competing needs. As deliberative action both takes time and requires access to limited deliberative reasoning resources, the original functionality encapsulated within Abbott's *skill* agents is still utilised by the *action proposer* agent in situations that require immediate action using simple perceptual (i.e. proprioceptive) feedback.

With the addition of deliberation, action selection can now be said to take place on two complementary levels:

a) *Deliberative* action selection basically follows the scheme outlined in Cañamero [97] and implemented in Abbott2 (see section 6.2.2). In Abbott3 we simplify this process by assigning a single *motive manager* agent the task of selecting behaviours to satisfy the active motivators (as chosen by the *attention filter* agent). This has the added benefit of allowing Abbott to attend to more than one motivator at a time – behaviours often have a number of different effects on the internal state of the architecture.

b) *Reactive* action selection focuses on generating immediate reactions to significant situations and events in the environment, and extreme levels of drives.

Deliberative and reactive action selection are intimately connected. As well as driving behaviours, the *recogniser* agent also activates the reactive *map* agents – which in turn feeds into the reactive concern-processing substrate. This allows our agent to produce affective reactions to deliberatively triggered images. Furthermore, Abbott's reactive concern-processing mechanism can also interrupt deliberative management and thus influence all levels of the architecture – receiving input from attentive perception, and acting through a global alarm system.

Abbott's primitive inhibition and fatigue motivator attention mechanism is supplemented with an active *attention filter* agent. This gives us the flexibility to produce a single drive-based motivator for urgent (highly insistent) sources of motivation, and still allow multiple non-urgent motivators to surface and be decided simultaneously. Finally, we have added the restriction that a *manager* agent's incentive stimuli can only come from the *recogniser* agent (i.e. the attended to object), implementing a form of perceptual attention within Abbott.

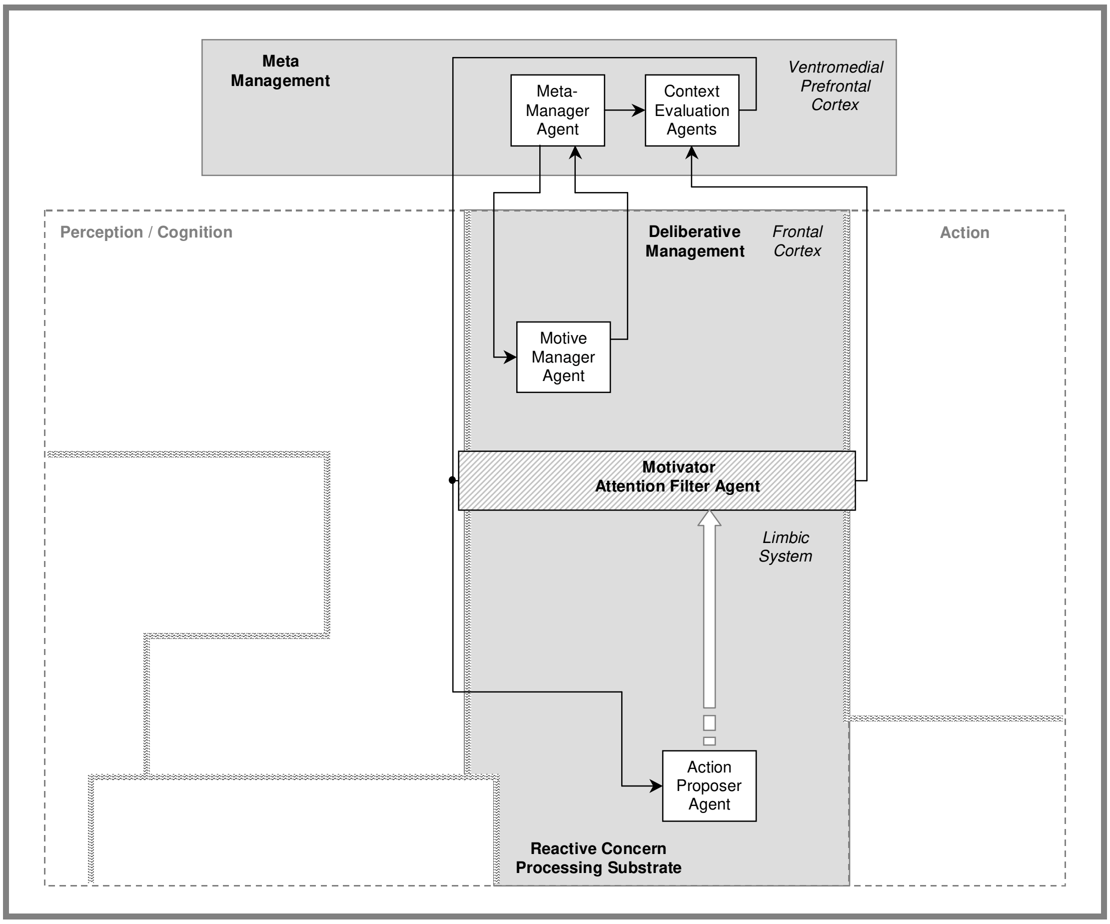


Figure 7.1-5 Abbott3d (showing competence level 3)

Figure 7.1-5 shows competence level 3 of the Abbott3 architecture. With the addition of a meta-management layer, we finally attain the basic three-layered model for intelligent autonomous agency introduced in section 2.2. Meta-management has two main responsibilities within the architecture: (i) it actively manages the processes that handle motivator evaluation, selection, and related processes like planning and plan execution – for example, ensuring that the agent does not expend too much energy attaining low importance goals; and (ii) it provides the reactive concern-processing substrate with more accurate input as to the coping ability (context evaluation) of the deliberative management layer.

Although the fine detail still needs to be worked out (i.e. how behaviours, skills and managers are created and adapted), we have demonstrated a plausible pathway for the computational evolution of a cognitively inspired mind – through added levels of competence in a subsumption style *Society of Mind* architecture. By grounding each competence level in the concern-processing mechanisms of level 0, we are also able to offer a plausible pathway for the development of mind during the lifetime of the individual – for example, Dux's *Weltbilder* theory [Dux 90] argues that the development of mind during the lifetime of an individual follows the same basic pattern of development that occurred during the evolution of the species.

7.1.2 Emergent Emotional States

The main difference between the design of Abbott3 and its predecessors (Abbott and Abbott2), is the absence of a clearly demarcated emotion system (Cañamero's [97] original design called for *emotion* agents to act as proto-specialists in a similar style to *motivation* agents – see Figure 6.1-6). As we argued in chapters 4 and 5, we believe that "emotions" are emergent mental states *caused* by the interaction of a variable number of intricately connected cognitive systems (i.e. systems that mediate arousal, attention, perception, concepts, memories, and physiological change) operating at different information-processing levels of the brain. Our approach towards elucidating emotions in Abbott3, is to replicate some of these systems at the information-level, and then explore the possible pathways through which emotional states can emerge.

In a sense, we are advocating a systems theory of emotion (as part of the more general requirements for intelligent autonomous agency). We must therefore remain vigilant to the accusation that the flexibility of our approach makes it possible to explain almost any data – if something does not fit the theory then the theory can easily be expanded to fit the data – and therefore makes it hard to formulate concrete predictions with which to test its validity. To counter such a claim we must re-emphasise the fact that our approach not only consists of an architectural framework (our three-layered model derived from an information-level analysis of the requirements for intelligent autonomous agency – see sections 1.3, 2.1, and 2.2), but also a design-based research methodology. As we add more depth to our agent architectures, we will be forced to make design decisions that will lead to concrete predictions (such as the number and type of reactive concern-processing mechanisms required for a particular niche, and the implications this has for the types of primary, secondary and tertiary emotional states supported).

We can make some tentative predictions (even if they are hard to verify) such as: (a) we would expect tertiary emotions to be more cognitive in nature and not easily distinguishable by physiological measurement alone – as they are unlikely to map on to unique/distinct reactive concern-processing systems; (b) we would expect secondary and tertiary emotional states to appear later in the development of a human mind, with secondary emotions subject to more cultural variability. Tertiary emotional states (such as those normally associated with grief, infatuation, and anxiety) should exhibit statistically less cultural variability than secondary emotions, as their perturbant nature arises out of a mismatch between cultural conditioning and the more universal mechanisms of primary emotions.

A single emotion type can cover a wide range of forms and intensities (with an even wider range of associated securities, insecurities, dreams, and feelings) – i.e. *being in love* covers: romantic interest; infatuation; longing; intense passion; mature love. To aid the verification of our predictions, one possible avenue for future research would be to develop a robust scheme for mapping sub-classes of common emotion types (possibly by linguistic labels) on to the underlying concern-processing mechanisms active in the emotion process. But first, we need to develop a series of emotional agents within which to elucidate the human emotion process, and clarify what exactly we mean at the information-level when we talk about love, hate, envy, and joy.

Our design is still far too shallow to claim that we can actually support human-like (or even infant-like) emotions, but as the following discussion will show, we can still usefully elucidate the emotion process within such a framework.

The role of Cognition and Self in the making of Emotion

Cognition (and by inference *self* – as represented by concern-processing mechanisms) is necessary for both the generation and interpretation of emotional states. However, this does not mean that the same cognitive processes we use to interpret our emotive states are active in their initial creation – generation and interpretation are two separate components of the emotion process.

Ortony et al.'s [88] *cognitive structure of emotions* is able to categorise the common emotion types according to the cognitive attributes of their eliciting condition – i.e. as valenced reactions to situations and events that impinge on agent *goals*, *standards*, and *attitudes*. Thus, the difference between being *proud* and being *pleased* is seen as one of whether the cognitive focus of the emotion eliciting condition is on: (a) the actions of the agent upholding a *standard* it wishes to maintain; or (b) the consequences of an event being desirable for the achievement of an agent *goal*. Mixed emotions, such as being proud and pleased at the same time, are explained by the fact that a single sequence of events can be appraised in a number of different ways – leading to both pride and happiness (or even pride and sadness, if the actions deemed necessary to maintain a standard, frustrate one of your goals).

However, cognitive theories of emotion must also explain the fact that: (a) the time taken to trigger the emotional response can be too short to include cognitive appraisal [Zajonc 80]; (b) *opponent process* [Solomon 80] and *reversal* [Apter 89] theories point out that emotions sometimes come in pairs – the termination of one emotion automatically brings about the onset of the opposite emotion and in circumstances whereby the second emotion cannot easily be explained by a change in appraisal of the same event [Mauro 88; Ellsworth 94, page 196]; and (c) it is hard to see how music can be appraised as being relevant to our "well-being," and yet music can trigger very complex emotional experiences.

We address these issues by: (i) separating the role cognition plays in the initial generation of an emotional episode, from its role in subsequent classification or modification; and (ii) acknowledging that evaluations relative to *self* occur on many different levels within the emotion process, and not just at a conscious level.

1) *The role of cognition*: Emotion is an emergent property of mind that appears in a variety of different forms and degrees of cognitive richness. There is clearly a difference between: (a) simply experiencing an emotional episode (i.e. a perturbant state which includes repeated interruption of attentive processing); (b) being aware of experiencing an emotional episode; and (c) being able to classify (or know why we are experiencing) an emotional episode. Each degree of emotional awareness adds an extra level of cognitive richness to the process, and gives the emotional state a different hedonistic tone – without changing our objective classification of the emotion type.

   Conscious cognitive appraisal *is* needed to describe and classify an emotional episode, but its role in the generation and experience of emotions varies greatly from person to person, emotion to emotion, and hedonistic tone to hedonistic tone. It is hard to imagine how the cognitively rich emotion of shame can exist without the *conscious* cognitive appraisal of the fact that you have done something wrong (in the sense that the cognitive appraisal component is included in our folk-psychology definition of shame) – i.e. can simply listening to music ever make us feel ashamed if we neither cognitively appraise our actions, nor the memories triggered by the music, as shameful?

   Although some emotion types clearly do require cognitive appraisal, the emotion process itself can still be triggered by reactive appraisal mechanisms (sufficient to interrupt attentive processing). As physiological responses are triggered by heuristic reactive appraisal mechanisms (see Figure 7.1-1), the time taken to trigger even complex emotions can be very short – it then takes a little longer for the emotional experience (with its specific hedonistic tone) to emerge and reach cognitive consciousness. The problem is again obscured by the emergent nature of emotion – i.e. is there really an exact point in the emotion process that an emergent emotional state can be said to exist? Do we define happiness when: (a) a *happiness* event triggers the interruption of attention? (b) we experience the feeling state normally associated with happiness? (c) somebody points out that we look happy? (d) when we understand why we are experiencing the feeling state normally associated with happiness? or (e) any/all of the above?

   Cognition plays an important, yet *relative*, role in the making of emotion. Reactive appraisal is certainly required in the generation of all emotional states – which in a sense goes without saying as any transformation of information by definition involves an evaluation (information exists in relation to a choice of process). If we restrict our definition of cognition to complex information-processing at the attentive/deliberative reasoning level, then we can make the distinction that cognition is not required for primary emotions, but is needed to trigger secondary and tertiary emotions – see section 5.1.

There are many neural pathways by which a deliberative evaluation of an event can lead to a reactive affective response (see our discussion on the somatic marker hypothesis in section 5.2.5), and/or subsequent cognitive context evaluation of an ongoing emotional episode gives the emotion its distinct flavour (i.e. identifies it as a particular emotion type). Appraisal theories show us how different appraisal types can be mapped to different emotion types, but not how those appraisal mechanisms are represented or evolve. The challenge we must now face is that of explaining how the brain (as part of the interaction of many different cognitive systems/agents) performs this appraisal.

2) *The role of self*: When asking questions about what types of cognition are active in the making of emotion, we also need to address the issue of what kind of *self* is required to support human-like emotional states. Even if we can give our agents human-like reactive, deliberative, and meta-management layers, it does not follow that they will exhibit recognisably human-like emotional states. We also need to give our agents the right set of basic concern-processing mechanisms – the *core-self* against which events are appraised (we can define the *extended-self* as the concern-processing mechanisms that evolve as the agent interacts with its environment – i.e. the mechanisms more closely associated with secondary and tertiary emotions).

   In section 6.2, we identified two different types of concern-processing mechanism that can be said to perform the necessary role of a *core-self*: (a) homeostatic drives which monitor an agent's internal state – for example, temperature regulation, hunger, sleep; and (b) non-homeostatic drives which respond to significant events (both internal and external) in the agent's environment – for example, sexuality, exploratory drive, and emergent states such as emotions.

   In the original Abbott design, Cañamero [97] used two different types of proto-specialist to represent these mechanisms: *motivation* agents and *emotion* agents.

   a) *Motivation* agents monitor the physiological variables (see Table 6.1-1) and produce an error signal proportional to the offset from a pre-defined range – these represent Abbott's homeostatic drives.

   b) However, Cañamero's use of *emotion* agents to represent the non-homeostatic drive mechanisms is a little misleading. All non-drive based concern-processing mechanisms were deemed emotional by virtue of the fact that Cañamero's *emotion* agents treat emotions as the product of discrete systems rather than emergent states – a useful abstraction in helping us identify the relative merits of affect modified motivation (see section 6.2.3), but of little use in identifying the basic non-homeostatic concern-processing mechanisms needed to support human-like emotional states.

In Abbott3 we replaced Cañamero's *emotion* agents with two different types of proto-specialist (*relevance evaluation* agents and *context evaluation* agents). We also added a global alarm mechanism (*action proposer* agent) to facilitate the interruption of attentive processing – allowing non-homeostatic drives to directly attain control precedence. With our new scheme we have created a fairly accurate information-level representation of the core components of the emotion process (see chapter 5). Our next task is to use this scheme to explore the design-space of concern-processing mechanisms (the *core*- and *extended-self*) capable of supporting human-like emotional states.

It is clear that Abbott's concept of *self* is not sufficient to support the notion of *standards* (beliefs about what ought to be the case as opposed to what one simply wants – or would like – to be the case), and so we cannot expect Abbott to be proud of its actions. However, we are still left with the question of what type of *self* is required to justify a claim that our agent can ever be happy? A question that can probably best be answered by building many iterations of "emotional" agents, with each generation standing on the shoulders of the generation before.

In the early stages of an infant's development, it has the cognitive abilities to support a very primitive concept of *self*, based on its biological needs (directly, or indirectly via attachment concerns). These basic concern-processing mechanisms lead to the emergence of a small set of primitive cross-cultural emotion types – roughly analogous to our concept of primary emotions. As a mind matures, it is able to support a richer concept of *self* (which includes control states such as *beliefs*, *standards* and *attitudes*), and thus facilitate more complex and cognitively rich emotional states – when events match or mismatch these new concerns. The initial set of infant emotion types are still present in the adult mind, but these too will have evolved, undergoing modification in both the forms of eliciting conditions, and their subsequent expression. Developing autonomous agents with concern-processing mechanisms that can adapt and evolve in a similar way to that of the human mind is one of the many challenges of intelligent autonomous agent design – for only then will people start to accept that our agents are truly "emotional."

Primary, Secondary, and Tertiary Emotions

One of the messages we have tried to drive home in this thesis is that there is no single system (or systems) that mediates emotion. The "emotion" phenomena emerges from the interaction of many different systems, performing many different roles, and operating at many different levels within a biological agent architecture. In the following discussion we will show how emotional control states emerge in the Abbott architecture (see Figure 7.1-1). We will not attempt to account for the specific folk-psychology emotion types (a task we describe in section 9.2), but rather explain how the different classes of emotional state emerge as part of the emotion process described in chapter 5.

There are a number of significant differences between the emotional states that can be supported within the Abbott architecture and true human emotions. In Abbott we have focused on capturing the minimal cognitive requirements for intelligent autonomous agency – hence our three-layered model refers to different levels of motivator management and not levels of human cognition. The differences are subtle, but nevertheless important:

1) Motivators are defined as motivational control states that move an agent towards a desired physical/mental state in light of agent beliefs and concerns. Motivator management is the process of managing (i.e. deciding, scheduling, modifying, or acting on) these motivational control states, and motivator meta-management is the process of managing motivator management mechanisms. Abbott's reactive and deliberative motivator processing layers map nicely on to the reactive and deliberative layers normally associated with human cognition. However, the motivator meta-management layer is not the same as a reflective or meta-cognition layer. Abbott has no *discrete* representation of self, and can only perceive and act on a *dispositional* self as represented by the active motivator management mechanisms – in practice this is not a problem as tertiary emotions are classified according to the emergence of a perturbant state and not some form of self-awareness.

2) As we mentioned above, Abbott suffers from a very impoverished concept of self. For example: (a) there are no explicit representations of agent concerns or beliefs in the deliberative layer (aside from the motivators generated within the reactive layer); (b) reactive concern-processing mechanisms can only respond to the actions of deliberative thought processes through the potential activation of *map* agents; and (c) there are no mechanisms for the generation of standards or "well-being of other" control states within the architecture. This does not detract from our ability to support emotional states, but it does severely limit the range of emotion types our agent can exhibit.

3) Finally, the hedonistic tone of human emotions is intimately connected with a sense of self-awareness and consciousness – both of which are missing in the Abbott3 design. The lack of these higher-level forms of cognition means that our emotional states will not have the same cognitive richness as human emotions.

In section 7.1.1 we briefly discussed the emergence of nascent primary and secondary emotional states within the context of co-evolution within Abbott's level 1 competence layer. In the following discussions, we will extend this analysis by investigating the theoretical emergence of *primary*, *secondary* and *tertiary* emotions within each successive competence level of the architecture. When describing the Abbott architecture, we will treat the *Society of Mind* members as reactive black-boxes – with relatively simple action selection mechanisms such as spreading activation, vector addition, condition-action rules, or ‘winner-takes-all’ networks. We will argue that emotions emerge as higher-level control states created by the interaction of individual society members, and not from a discrete emotion system added to the architecture.

Competence Level 0 Emotions

The *Society of Mind* agents, that make up Abbott’s level 0 competence layer are shown in Figure 7.1-3. At the heart of the Abbott3a architecture sits the *action proposer* agent – which is probably comparable in complexity to the behaviour-based agents discussed in section 3.2. However, the *action proposer’s* role in selecting actions to satisfy drives is only a small part of its functionality – it also takes part in a number of feedback loops to provide Abbott with primitive attention and global alarm mechanisms.

Abbott is able to support a primitive form of motivational sharpening (motivator attention) by allowing the *action proposer* agent to modify the perceived somatic state – and thus indirectly the magnitude of the error-signals generated by *drive* agents (Abbott’s homeostatic motivational state). Abbott is also capable of using the *action proposer* agent as part of a global alarm mechanism to switch motivational attention in response to significant objects/events in its external environment (as detected by the *relevance evaluation* agents). Together, the homeostatic *drive* agents and the non-homeostatic *relevance evaluation* agents provide Abbott with a static *core-self* (the primary concerns against which events are appraised to establish motivational attitude). *Body* agents, *physiological change generator* agents and *somatic sensor* agents provide Abbott with a sense of valence and a dynamic *core-self* – with Abbott’s affective state reflected in the relative levels of its neurochemical control signals. Although there is no deliberative management layer, Abbott still represents quite a sophisticated autonomous agent.

On the surface, these simple concern-processing mechanisms provide all the functionality needed to support emergent *primary* emotional states. External events are able to attain control precedence and redirect motivational attention towards satisfying specific concerns (as represented by the *relevance evaluation* agents). Emergency reflex-like action can be achieved by directly activating *skill* agents in response to external stimuli, and more sustained action through the release of neurochemical control signals and their affect on the perceived somatic state. Abbott differs from other reactive autonomous agents, such as Braitenberg’s [84] vehicles (which to an outside observer can also be said to exhibit simple emotion-like states), in that its internal architecture models the information-level emotion process as described in chapter 5 – with interruption of attention, valence, and motivational attitude. If we are not yet convinced that the level 0 Abbott3a architecture is capable of supporting primitive emotional states, then perhaps the addition of competence level 1 will take us over the threshold.

Competence Level 1 Emotions

The addition of Abbott's competence level 1 society members (Figure 7.1-3), allows Abbott3b to mark percepts in its external environment that are coincident with aroused affective states. *Somatic marker* agents thus give Abbott the ability to learn from previous encounters and anticipate future events – the ability to create an *extended-self*. Acquired associations between categories of objects and situations on the one hand, and primary emotions (albeit nascent ones) on the other, have been termed the *machinery of secondary emotions* [Damasio 94, page 136; also section 5.2.5). Unfortunately, Abbott is still devoid of a deliberative layer (the acquired associations only relate to immediate objects in Abbott's environment), and so we can only really claim a richer form of our level 0 emergent "emotional" state.

Competence Level 2 Emotions

The addition of competence level 2 (Figure 7.1-4) gives Abbott3c both a deliberative reasoning layer, and an attention filter to protect deliberative reasoning from excess interruption by reactive motivator generators. The deliberative layer is used to select the *behaviour* agent that will most likely satisfy the current active motivator(s) – as per Abbott2 (see section 6.2.2). Once a *behaviour* agent has been chosen, the *motive manager* agent uses the *manager* and *recogniser* agents to locate the behaviour's incentive stimulus (the object to eat, drink, rest on, or avoid). The *recogniser* agent also activates the *map* agents: (a) allowing attended to objects to alter the perceived somatic state – and thus increase a motivators insistence level; and (b) giving Abbott the ability to generate an affective state in response to an attended to object (under the control of deliberative thought processes) – a candidate secondary emotion.

If the chance sighting of an enemy whilst looking for food (a deliberative process involving *manager* agents) generates an interruption of deliberative attention and change in action readiness, would this count as an emergent secondary emotional state? If not, then perhaps allowing the *behaviour* agents to directly activate the *map* agents (i.e. create an imagination-like control state during behaviour selection), and thus provide an affective feedback path to deliberative thought processes, would be enough to qualify the claim that Abbott can support secondary emotions?

Competence Level 3 Emotions

We can complete our Abbott architecture by adding the society members that make up the level 3 competence layer (Figure 7.1-5). Abbott3d's *meta-manager* and *context evaluation* agents contribute the final cognitive elements to the emotion process picture described in chapter 5 – allowing us to elucidate the *primary*, *secondary* and *tertiary* emotion pathways:

1) *Primary Emotions* utilise two different eliciting pathways: (a) the "low road" from the early *sensory* agents (tactile and brightness); and (b) the "high road" through the

direction neme, *map*, and *somatic marker* agents. The low road represents the route for Abbott's *innate* emotional responses to external stimuli, and the high road for stimuli previously *associated* with earlier primary emotional episodes. The relevance of the external stimuli are assessed by a small number of *relevance evaluation* agents – resulting in the information-level equivalent of the relevance signals of pleasure, pain, curiosity and desire (see section 5.2.3).

The *action proposer* agent makes a heuristic estimate of the importance/urgency of the situation/event based on the relevance signals – resulting in any of: (a) selection of a *skill* agent in very urgent situations; (b) activation of a *physiological change generator* agent in situations that might require physiological arousal; (c) generation of a motivator in situations that require deliberative attention; (d) modification of the *somatic sensor* agents through the "as if" loop. Motivator generation either occurs directly via the *action proposer* agent, or indirectly through the changed somatic state and *drive* agents.

This control process replicates the information flow for a primary emotional state shown earlier in Figure 5.2-4. A primary emotional state emerges when the generated motivator attains control precedence and is adopted by the *motive manager* agent – i.e. when its insistence level is higher than the threshold defined by the *attention filter* agent. Valence is either attached to the situation/object through a change in somatic state (real or through the "as if" loop), or directly associated with the motivator itself.

2) *Secondary Emotions* are emergent emotional states that require deliberation at some point in the emotion process. For example, Damasio [96] concentrates on emotions generated in response to specific situations, events, or objects which have previously been paired with primary emotions, but are now triggered by deliberative thought processes (see section 5.2.5); whereas Sloman [99] also highlights emotions generated with respect to the planning process itself (see section 5.1) – i.e. when relevant risks are noticed, progress assessed, and success detected. We described Damasio's case in our discussion of competence level 2 emotions, and so here we will concentrate on emotions generated in response to inner perception and action within the deliberation process itself.

   Abbott's *meta-manager* agent continually monitors the *motive manager* agent, and is thus able to detect if, for example: a relatively low-level motivator is taking too long to satisfy; repeated behaviours are failing; the same low-level motivators are always being attended to; or a behaviour succeeds in satisfying a motivator. This type of information (along with the current threshold of the *attention filter* agent) allows the *context evaluation* agent to assess the effectiveness of the current coping strategy adopted by the deliberative layer. In situations where the current strategy is not working, the *action proposer* agent can use this context information to interrupt the *motive manager* agent with a new motivator, replicating the information flow for a secondary emotional state shown earlier in Figure 5.2-6.

3) *Tertiary Emotions* are normally associated with Damasio's class of secondary emotions – described in the context of competence level 2 emotions above. After the adoption of the new motivator (within the secondary emotion process), the *meta-manager* agent evaluates the new motivator as irrelevant and signals both the *context evaluation* agent and the *motive manager* agent. Control of attention is regained through context evaluation of the current situation, allowing the action proposer to evaluate a relevant event as non-urgent. However, repeated triggering of the secondary emotion via subsequent actions of *recogniser* (or *behaviour*) agents leads to a perturbant state. This temporary loss of control of attentive processing replicates the information flow for a tertiary emotional state shown earlier in Figure 5.2-7.

   In a sense, it is the difference in the adaptation rates of reactive and attentive meta-management processes to new situations that leads to the emergence of *tertiary* emotional states – the emergent state is terminated when the reactive motivator meta-management mechanisms (in the form of *somatic marker* agents) have had a chance to catch up with the new situation, which in the case of strong attachment concerns may never completely happen (i.e. with the tertiary emotional state of grief).

With the addition of a motivator meta-management layer, Abbott is able to exhibit simple emergent *primary*, *secondary*, and *tertiary* "emotional" states. We are definitely not claiming the title emotional for our agent architecture (and so will continue to make use of the scare quotes). However, we strongly believe that through the design, implementation, and subsequent analysis, of architectures such as Abbott, we will be able to achieve a greater understanding of the human emotion process.

7.1.3 Conclusion

Abbott3 represents our latest design for a *cognitively inspired* agent to meet the basic requirements for intelligent autonomous agency (see our discussion on the *Strengths and Weaknesses* of the design in section 8.1.2). Our design attracts the label "cognitively inspired" on two accounts: (a) the design has evolved from an information-level analysis of psychological and neurological models of human concern-processing mechanisms (chapter 5); and (b) it demonstrates a plausible mechanism for the development of mind on both an evolutionary time-scale, and throughout the lifetime of the agent (section 7.1.1).

In this section, we have presented a concern-centric autonomous agent design that recognises both the need to: (a) allow competence levels to co-evolve within a *Society of Mind* framework; and (b) ground higher-level competence levels in the concern-processing mechanisms of the lower-level competence levels. This approach allows the level 0 competence level to act as a fast, reactive, global alarm mechanism – alerting the agent to situations and events that impinge on the primary concerns of the agent. Grounding competence levels in the level below also allows an agent to develop an extended-self compatible with the core-self represented by the basic level 0 concern-processing mechanisms.

Finally, we discussed the mechanisms through which emotional states emerge within our agent architecture. By describing the characteristics of the different classes of emotional state each competence layer is capable of supporting, we are able to show the evolution of the emotion process within the context of the evolution of agent architecture itself. Although we do not claim to support emotional states of anything like a comparable complexity and richness to those experienced by humans, our discussion serves to illustrate how the extended motivated agent framework can be used to elucidate the emergence of human-like emotions in intelligent autonomous agent architectures.

7.2 Summary

In this chapter we have presented the design of a cognitively inspired agent architecture for elucidating infant-like emotional states – integrating the different research strands explored in chapters 1 through 6. We described how the different concern-processing competence levels of our three-layered architecture co-evolve, and identified the different processes active in the emergence of emotional states.

In the next section we will describe an implementation of Abbott3, and provide critique of our design – identifying its strengths and weaknesses, and discussing how it addresses some of the problems associated with the traditional deliberative and behaviour-based architectures described in chapter 3.

8 Implementation and Critique

Our Abbott3 design, described in the last chapter, provides the basic architecture to allow us to investigate the emergent “emotion-like” properties of our motivated agent framework.

8.1 Putting Theory into Practice

In this section we will present an implementation and critique of our design, discussing in more detail: (a) the requirements specification with respect to Abbott2; (b) an implementation of the Abbott3 design; (c) some experimental results exploring the Abbott design- and niche-space; and finally (d) the strengths and weaknesses of our design – i.e. how our approach contributes to the field of intelligent autonomous agency by addressing some of the problems identified with the agent architectures described in chapters 3, 4 and 6.

8.1.1 Requirements Specification

In keeping with the iterative nature of our design-based research methodology, we will now provide a brief overview of the new Abbott3 society members. Our discussion will focus on the differences between the ‘deeper’ Abbott3 design, and the existing Abbott2 design. The actual design details, and experimental results, will be discussed in sections 8.1.2 and 8.1.3.

Perception/Cognition

Abbott3’s perceptual system has changed very little from that employed by Abbott2, with the noticeable exceptions of the addition of *somatic marker* agents and a pathway from *recogniser* to *map* agents. The implications of theses changes are discussed below.

Somatic marker agents provide a useful mechanism to allow our agent to adapt to its environment. Abbott2 arrived in the world with pre-configured *map* and *primary emotion* agents capable of recognising and responding to all the objects it was likely to meet in the environment. With the addition of *somatic marker* agents, Abbott3 is able to mark objects in accordance with its current affective state, and thus learn how to respond to ‘significant’ objects in its environment (Cañamero’s original design called for *recogniser* agents to ‘learn’ new classifications of objects and not associations between objects and affective states). *Somatic marker* agents ground Abbott’s level 1 concern-processing mechanisms in the innate level 0 mechanisms – associations are actually created within the *relevance evaluation* agents, with the *somatic marker* agents acting as simple filters for significant external stimuli.

Abbott’s affect system is used to mark significant objects and events in its environment. As we discussed in the previous chapter, this allows the richness of the emotional states supported by the architecture to grow with the development of the agent. Affect grounded learning allows the different competence levels themselves to develop as our infant mind matures – i.e. it is possible to extend the architecture to allow behaviours to 'learn' the effects of their actions through affective feedback, and thus better assess their contribution to satisfying the active motivators of the agent (see *Pandemonium Theory* [Selfridge 59; Jackson 87]; and the *Theory of Neuronal Group Selection* [Edelman 87]).

Abbott's affect system is also used to attach valence to situations and events either directly via the *action proposer* agent or through the *somatic sensor* agents. By allowing the *recogniser* agent to activate the *map* agents during the competence level 1 perceptual process, it becomes possible to give valenced affective feedback to attended to objects and even future events – i.e. *behaviour* agents can attend to their incentive stimuli providing the *motive manager* agent affective feedback to any previous 'bad' experiences associated with the stimuli as part of the behaviour selection process (currently the *motive manager* picks the *behaviour* agent that 'shouts' the loudest). Abbott is above all an open architecture for exploring the space of possible intelligent autonomous agent designs.

Action and the Body Loop

Abbott3 continues the tradition of making the biological heritage of the affect system more explicit by modelling both *physiological change generator* agents and *body* agents. This is partly done to achieve more realism within the scenario and partly to simplify the control structure. Although *body* agents are an integral part of Abbott, from a control perspective they can also be said to belong to the external environment (as they model the physical structure of the biological agent) – separating the internal and external environment thus simplifies the control structure. Abbott is able to modify its internal state locally by manipulating the *somatic sensor* agents through the *as if* loop, or globally through the *physiological change generator* agents and the body loop (*body* agents can be simple glands or complex organs). We have further simplified the affective pathway by assuming that neurotransmitters are also generated by some form of *body* agent – this is a clearly an over simplification, but still a useful abstraction.

Reactive Concern-Processing Substrate

As we discussed in section 5.2, cognition is a label we use as a shorthand description for certain types of brain function, which can equally be applied to both deliberative reasoning and reactive heuristic evaluation. Relevance evaluation is a cognitive function that occurs within the confines of the reactive concern-processing substrate (i.e. a heuristic evaluation that could be made relative to: the inflection of voice; the suddenness of movement; a learnt somatic marker; or a simple mechanism that monitors deliberative reasoning processes). Abbott3's *relevance evaluation* agents mimic this process by matching situations and events against agent concerns. The optimum number and type of *relevance evaluation* agents will depend on a variety of factors (not least the particular niche the agent inhabits in the environment), and can best be established through empirical experimentation – one of the fertile areas for future research. We would however expect the number/type of *relevance evaluation* agents to remain quite small – roughly coincidental with the number of emotion circuits proposed by emotion theorists (i.e. of the order of 3 or 4).

On a superficial level, the *relevance evaluation* agents can be said to replace the *primary emotion* agents of Abbott2. There are however a number of significant conceptual differences between the two types of agent that makes any such comparison invalid. Abbott2's *primary emotion* agents attempt to capture both the relevance evaluation functionality and action tendencies associated with human emotion types – i.e. PriEmoHappy and PriEmoFear. This may have been a useful abstraction for exploring the added benefit of an emotion system, but fails to model the emotion process as outlined in chapter 5 – *emotion* agents do not represent "emotional" states. Abbott3's *relevance evaluation* agents simply evaluate situations and events for relevance to a small set of basic concerns. *Primary* emotions are supported through the "low road" from the early *sensor* agents and *secondary* emotions through the "high road" via the *somatic marker* agents. Here we are not attempting to capture specific emotion types, but provide the functionality needed to support the different emotion classes.

Within Abbott3, action tendencies are supported on four different levels: (a) direct action through the *skill* agents; (b) deliberative action through gaining control precedence; (c) indirect action by modifying the perceived somatic state through the "as if" loop; and (d) physiological action through the body loop. The actual response chosen will depend on the type and urgency of the relevance signal derived from the *relevance* and *context evaluation* agents. We leave the architecture open to the exact nature of the relevance signal – either Frijda's [86] signals of pleasure, pain, novelty and desire through a judicial partitioning of the *relevance evaluation* agents, or simply activation energy and possibly somatic state.

Deliberative Management

Abbott3's deliberative management mechanisms are essentially the same as those used in Abbott2 – with the addition of a *motive manager* agent to replace the dual functionality of Abbott2's *motivation* agents (motivational control states are now represented by messages rather than activation levels of physical *motivation* agents – giving the flexibility for the generation of multiple motivators by both *drive* and *action proposer* agents). The *motive manager* agent selects the *behaviour* agent that can best satisfy the current set of motivators under attentive deliberation (more than one motivator can pass through the attention filter at a time) – a process that relies on the *behaviour* agent's knowledge of its own competence (similar to *Pandemonium Theory* [Selfridge 59; Jackson 87]). Deliberation can be readily extended by allowing *behaviour* agents to assess their own levels of competence through affective feedback, and/or giving the *motive manager* agent access to affective memories via the *somatic marker* agents.

Meta-Management

Meta-management is the process of managing deliberative management processes. In Abbott3 a simple form of meta-management can be used to detect when a trivial motivator takes too long to satisfy (wasting time and resources), a motivator has been satisfied, or the *motive manager* agent is interrupted too often. The *meta-manager* agent can redress these inefficiencies by raising the filter threshold (via the *context evaluation* agent) or actively rejecting the current motivator.

A second role of meta-management is that of monitoring the current coping abilities of the agent. The *context evaluation* agent uses the filter threshold level and output of the *meta-manager* agent as a rough guide to how well the agent is coping at a particular moment in time. If the agent is "stressed" then the *context evaluation* agent can raise the filter threshold or signal a change to the global state on the society of agents via the *action proposer* agent. Active rejection of motivators by the *meta-manager* agent should be considered part of the normal management process, whereas a change in state (with interruption of attentive processing) via the *action proposer* agent corresponds to an emergent *secondary* "emotion".

Tertiary "emotional" states emerge when meta-management processes repeatedly lose control of motivator management due to a mismatch between reactive relevance evaluation and subsequent deliberative re-evaluation – i.e. an event is deemed relevant at a reactive level (generating an insistent motivator that attains control precedence), only to be rejected as non-urgent by deliberative management in the context of the current situation, but later regains control precedence again as the reactive concern-processing substrate has yet to adjust to the new situation. *Tertiary* emotional states can be supported by allowing the *meta-manager* agent to establish the true importance of a motivator based on previous history. For example, if a motivation to increase vascular volume is quickly satisfied then it is indicative that such motivations are not really urgent (although they may still be important). However, it should be stressed that the aim of meta-management is not to support tertiary emotions, but to allow Abbott to adjust to different niches during the lifetime of the agent – it just happens that such mechanisms (along with heuristic relevance evaluation) naturally lead to the emergence of tertiary emotional states.

8.1.2 Implementation Details

In this section we will cover the basic implementation details of a 'raw' Abbott3 design that meet the requirements specification of section 8.1.1 and support the emergence of "emotional" states as described in chapter 7. Here we use the qualifier 'raw' in anticipation of the fact the our exploration of design-space will add many levels of refinement to the simple algorithms used in this implementation. We will start this exploration/refinement process in the next section with the presentation of a series of experiments showing: i) how each competence level adds to the survival fitness of the Abbott design; ii) how individual society members contribute to the fitness of the design; and iii) how the interaction of the agents and layers naturally leads to the emergence of "emotional" states.

The full source code for the following Abbott design can be found on the poplog Abbott web site at: http://www.cs.bham.ac.uk/research/poplog/abbott

Competence Level 0

The base competence level allows Abbott to survive quite adequately in the Gridland scenario – the average survival time of a level 0 only Abbott is about 70% that of the complete Abbott, with 10% of runs lasting more than twice as long as the complete Abbott's average (the experiment consisted of a single Abbott with three enemies in a closed 30x30 world, with the average being taken over the shortest 70% of runs – see section 8.1.3). A quick feel for the complexity of the base competence level can be gained by noting that it contains 12 different agent types (see Figure 7.1-2), spread over 59 individual society members running a total of 166 condition-action rules – as a comparison, the complete Abbott architecture contains 76 individual agents and 342 rules (competence level 1-3 society members are themselves more complex, hence the higher ratio of rules to agents).

The *sensor*, *direction neme*, *map*, *drive*, and *effector* agents are identical to those used in the Abbott2 design, and have previously been described in section 6.2. The *skill* agents act as combined *manager* and *behaviour* agents, without the explicit representation of the incentive stimuli. *Skill* agents do not explicitly look for their incentive stimuli, but simply make Abbott wander around until the required conditions are encountered – i.e. food or water is found. The *physiological change generator* agents are responsible for releasing hormones in response to requests from the *action proposer* agent, and the body agent regulates Abbott's internal state to reflect its actions in the outside world (i.e. walking uses energy).

The *relevance evaluation* agent responds to different percepts within the different competence levels. At competence level 0, the *relevance evaluation* agent detects bright objects reported by the *map* agent, and generates a 'dangerRelEval' signal (bright objects correspond to enemies in our setting – nature adopts a similar strategy with bright red and yellow insect colouring). The relevance evaluation signals, and drives, feed into the *action proposer* agent, which sits at the heart of the reactive base competence layer and is responsible for selecting the appropriate internal/external actions.

The action proposer agent uses a fairly simple action selection algorithm to select the appropriate skill for the dominant drive – see Figure 8.1-1. If the activation level of the dominant motivator is low, and a deliberative management layer present, then the *action proposer* will defer overt action to the deliberative layer. If on the other hand, the activation level of the dominant motivator is high, the *action proposer* agent will usurp the deliberative layer by inflating the activation energy with which the *skill* agent is selected. This simple mechanism allows the reactive layer to gain control precedence in urgent situations, and yet still defer non-urgent situations to the consideration of the deliberative layer. The *action* *proposer* agent will also boost the activation energy of the *skill* if the *context evaluation* agent detects a "stressed" state (i.e. that Abbott's filter threshold has been high for an extended period of time). The numbers used by action selection algorithm are chosen to give the *action proposer* agent control precedence in urgent situations. As we learn more about the interaction of the different agents and layers, we will be better placed to refine our algorithm – for a start, the use of the single dimensional activation energy value to represent both urgency and importance is clearly an over-simplification.

1) Look for the most urgent new drive, record its activation level, and set activeSkill to the corresponding skill.
2) If the activation energy of the new drive is high, then multiply it by a factor greater than 1, else multiply it by a factor less than 1. If the activation energy is low, and the deliberative layer is present, then inhibit all skills by setting activeSkill to false.
3) If the 'stressed' message is present, then multiply the activation energy by a factor greater than 1.
4) If the 'relevanceEval danger' message is present, then set activeSkill to skillWithdraw, activation energy very high, and activate adrenalineChangeGen agent.
5) If the 'relevanceEval somaticMarkerPain' message is present, then set the 'as if' pain sensor to the strength of the marker.
6) If 'activeSkill' is defined, then activate the skill with the maximum activation energy, and record a change of skill if appropriate (allows meta-management to also monitor the reactive layer).

Figure 8.1-1 Action Proposer Agent Selection Algorithm

Competence Level 1

The addition of a *pain somatic marker* agent to the architecture allows Abbott to learn from affective experiences (i.e. aroused body states). A simple learning mechanism based on the output from Abbott's short-range map agents proved unworkable as the source of the pain had invariably moved before the aroused body state could be acted upon. The Gridland scenario has its own implicit 'laws of physics' which dictate, for example: that agents can only move at the end of a five-cycle time step; that sensors take one cycle to register a change in the environment; and that eyes can only move every two cycles. The internal message passing mechanism within the society-of-mind architecture also adds its own set of constraints which, amongst other things, ensures that deliberative actions require more time than single cycle reactive actions (we also use the currency of activation energy to ensure that reactive messages carry more weight at the effectors than deliberative messages).

Abbott's *pain somatic marker* agent relies on percepts gathered in the cycles immediately following an attack (Abbott instinctively turns towards the source of pain as it moves away). Multiple percepts are resolved over time (i.e. after a number of individual attacks) until single type of percept becomes dominant and can be marked as the source of pain – Figure 8.1-2 shows Abbott marking an ‘enemy’ as the source of pain when aroused during an attack. Once a percept becomes dominant it is hard to dislodge (mimicking the biological model), and so just occasionally blocks are marked when an enemy becomes obscured after an attack – Abbott’s eye sensor uses simple ray tracing to maintain a realistic perspective on the world (see Figure 6.2-3). The number of individual times a percept has been detected is used to create the strength of somatic marker – the more times Abbott is bitten, the stronger the reinforcement.

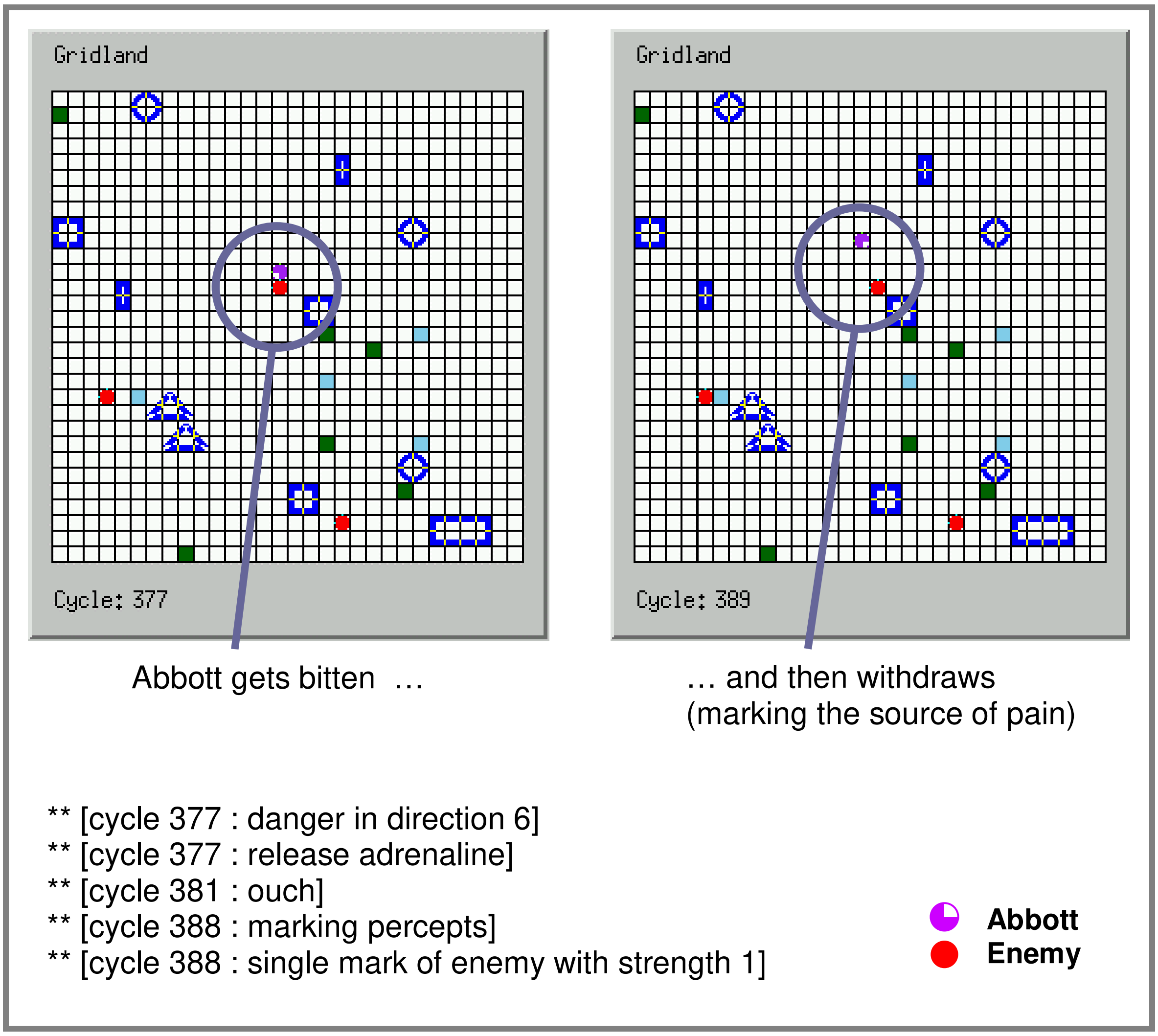


Figure 8.1-2 Abbott responding to being bitten

After an object has been marked, a ‘somaticMarkerPain’ motivator is generated by the *relevance evaluation* agent (the activation energy being a function of the strength of the marker and the distance of the new percept) whenever the object is next perceived. The *action proposer* agent then recreates the sensation of pain through the “as if” body loop, which can then lead to action via the ‘withdraw’ drive in the next cycle.

Competence Level 2

Competence level 2 differs from Abbott's lower competence levels, in that explicit representations of active motivators and incentive stimuli are held and compared over time – for which we use the term deliberation (see section 2.2). Abbott's basic deliberation mechanism (based on the *manager* and *behaviour* agents) remains the same as in our earlier Abbott2 design. However, the addition of an explicit *motivator manager* agent allows Abbott to select behaviours that contribute to the satisfaction of multiple motivators, and/or consider multiple motivators when selecting appropriate behaviours – the latter allows Abbott to select a consummatory behaviour for the non-dominant motivator over an appetitive behaviour for the dominant motivator (i.e. the motivator with the most activation energy). The basic behaviour selection algorithm is shown in Figure 8.1-3.

1) *Behaviour* agents whose incentive stimulus is present (i.e. a food map for the *BehaviourEat* agent) post a "stimulus_observed" message on the black-board stating the drives they satisfy. The *motivator manager* agent scans this list looking for a behaviour whose primary or secondary effect satisfies one of the active motivators and sends an activate message to the chosen behaviour.
2) If no "stimulus_observed" messages are valid for the active motivators, the *motivator manager* agent posts a "match_drive" message on the black-board (the match_drive message is actually posted as soon as the motivators are selected to speed up the operation). Any behaviour agent that can contribute to satisfying this drive responds by posting the incentive stimulus it needs to accomplish the task. The *motivator manager* agent then activates the *Finder* agent with the incentive stimulus as its "attend_to" object.
3) If at any point a *behaviour* agent returns a "failed" message the *motivator manager* agent chooses a new behaviour.

Figure 8.1-3 Behaviour Selection Algorithm

The addition of the *motivator manager* agent also allows us to replace Abbott's simple 'winner-takes-all' motivator selection algorithm with an algorithm for actively managing motivators – opening up the possibility for motivator meta-management. The gross filter threshold value is set by the *motivator manager* agent. However, the *filter* agent also uses a variable filter relaxation algorithm (slowly decreasing the filter threshold by a fixed percentage, and/or fixed amount, every world cycle), the parameters of which can be altered by the *motivator meta-manager* agent. Any motivator whose activation energy level is greater than the current filter threshold passes through to be evaluated by the *motivator manager* agent.

Once motivators have surfaced through the filter, they are evaluated by the *motivator manager* agent – although only a single motivator is finally adopted, the behaviour is selected during the motivator deciding process, and so can be selected to satisfy multiple motivators. Motivators are actively managed on two levels: i) motivators are evaluated, and the active motivators marked as "pending" or "adopted" according to the algorithm given in Figure 8.1-4; and ii) active motivators are then managed to remove those motivators no longer valid, according to the algorithm given in Figure 8.1-5. "Adopted" motivators can later be rejected by the *motivator meta-manager* agent (competence level 3), in which case they cannot be re-adopted for a fixed number of cycles.

1) Add the new surfaced motivators to the list of active motivators (updating the activation energy of the existing motivators and marking the motivators as "pending").
2) Remove existing motivators much less than the min activation level of the new motivators (unless marked as "rejected").
3) Deselect existing behaviours (consummatory or appetitive).
4) Activate the search for consummatory behaviours that match active motivators (irrespective of adopted motivator). This allows Abbott to take advantage of opportunistic situations if any exist.
5) Select the behaviour that matches the active motivators (the active motivator that is coincident with the selected behaviour is marked as "adopted").
6) If new motivator "adopted" then record fact that motivator has changed (used by motivator meta-manager).

Figure 8.1-4 Motivator Deciding Algorithm

1) Maintain a list of active motivators, removing all motivators that time-out (motivators have a fixed lifetime from the moment they penetrate the filter, which is updated whenever they resurface).
2) If no active motivators are present, reset the filter to zero.

Figure 8.1-5 Motivator Scheduling Algorithm

Competence Level 3

Abbott's final competence levels provide our agent with relatively simple motivator meta-management capabilities with which to monitor and modify the motivator management processes. This meta-management scheme is shown in Figure 8.1-6, and can readily be extended as more is learnt about the interaction of the agents, and layers, through exploration of design- and niche-space (see section 8.1.3). For example, it might prove advantageous to modify the filter relaxation rate in line with the "busyness" of the agent.

1) If the same motivator remains adopted for a long period of time then reject the motivator and reset the filter threshold. Rejected motivators cannot be re-adopted for a fixed number of cycles.
2) If the filter remains high for a period of time then mark the agent as 'stressed' (used by the action proposer to boost skill activation levels).

Figure 8.1-6 Simple Motivator Meta-Management Algorithm

8.1.3 Experimental Results

The final thread of the design-based approach (see section 1.2.2) calls for an analysis of similar designs in design-space – to give a deeper understanding of the tradeoffs inherent in the existing design. In this section we will report on a number of simple experiments that show how each layer and/or agent contributes to the overall fitness of our architecture.

Experimental Setting

Our experiments take place in the Gridland world (see Figure 6.2-2), with a standard configuration of a single Abbott, eight assorted blocks, five regenerating sources of food and water, and three enemies. Within this environment, Abbott must forage for food and water as it attempts to maintain a healthy internal state (represented by a number of varying physiological variables such as vascular volume and blood glucose level). Foraging is very tiring, and every so often Abbott must find a block to rest on. Finally, Abbott must also avoid the enemies which also inhabit the world – an enemies' bite will relieve Abbott of one of its five lives. When Abbott looses all its lives it dies, and the run ends. The run is also terminated after 20,000 cycles. More details of the Gridland toolkit, and experimental procedure, are given in appendices A and B respectively.

A 'survival time'/'fitness' profile for a particular architecture is obtained by plotting the frequency distribution of the survival times of Abbott over a number of runs – each run provides a single sample point. Using such an approach provides a convenient one dimensional graph with which to compare design trade-offs, without reducing the multidimensional fitness function to a single dimensional number. We can also group all runs greater than 10,000 cycles together to provide a highly visible measure of the architecture's longer-term performance. Finally, by specifying the seed used for the random number generator at the start of a run, we can ensure that each run starts in a random, but repeatable, initial state – i.e. the set of initial conditions remains invariant across different experiments.

The Abbott3 architecture has not been fine-tuned (aside from some initial gross tests on the filter characteristics), and as such represents a 'raw' implementation of the requirements specification. We have also tried to balance the functionality of the skills and behaviours so as not to inadvertently handicap the reactive competence layers. Unless otherwise stated, each profile is obtained from the frequency distribution of 500 experimental runs (about 13 hours of computation time on a 433 MHz Celeron™ PC).

Contribution of Competence Levels

Figure 8.1-7 shows the survival time profile for a single Abbott in the Gridland world with three enemies, an assortment of blocks, food and water – we have adopted the labels 'level 0', 'level 1', 'level 2', and 'level 3', for those designs with competence levels 0, 0-1, 0-2, and 0-3 respectively. These results clearly show an advantage for the longer-term survivability of architectures with competence level 2 – (i.e. levels 2 and 3 in Figure 8.1-8). However, as the capabilities of the different layers vary by more than the representation of future states and the management of motivators, we must refrain from simply chalking up a victory for active motivator management at this stage. For example, the advantages gained by level 2 could just as easily be attributed to the fact that the *finder* agent looks left and right when searching for the incentive stimuli, making it more likely to spot enemies.

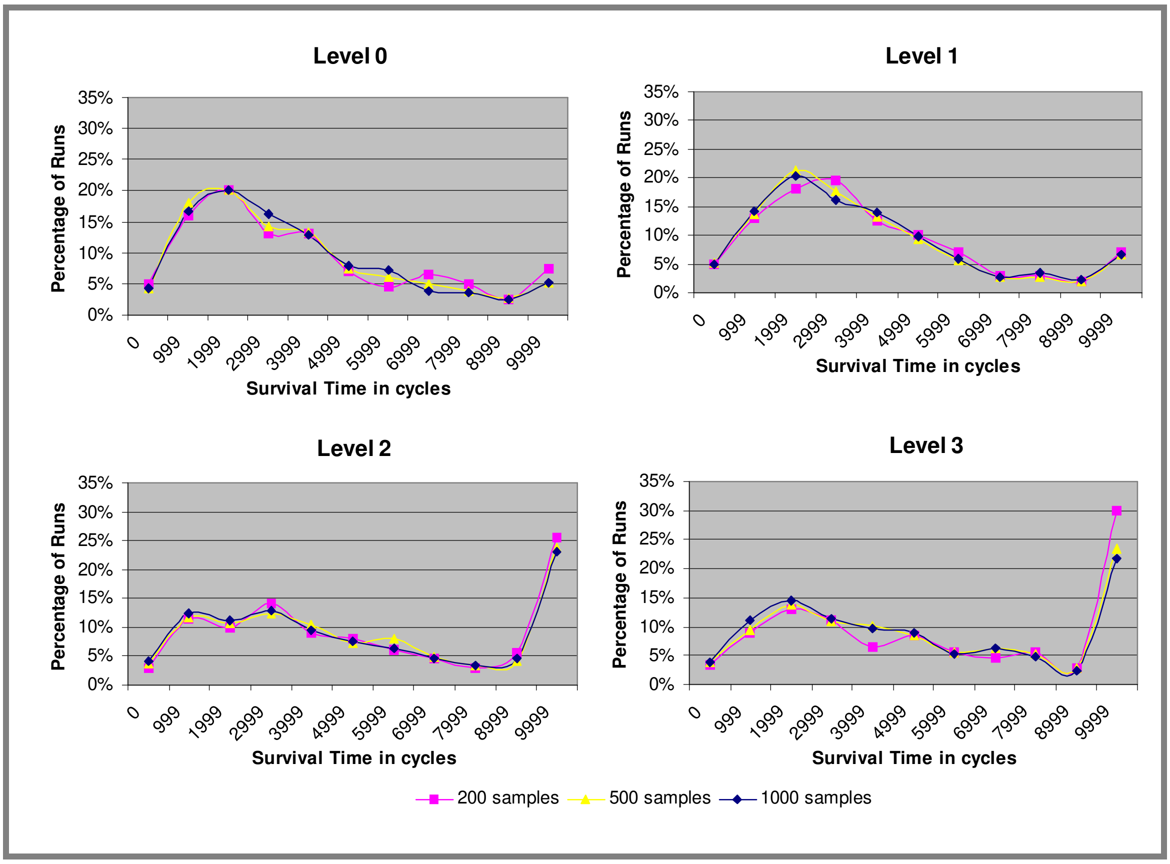


Figure 8.1-7 Survival Times for Competence Level 0-3 Architectures

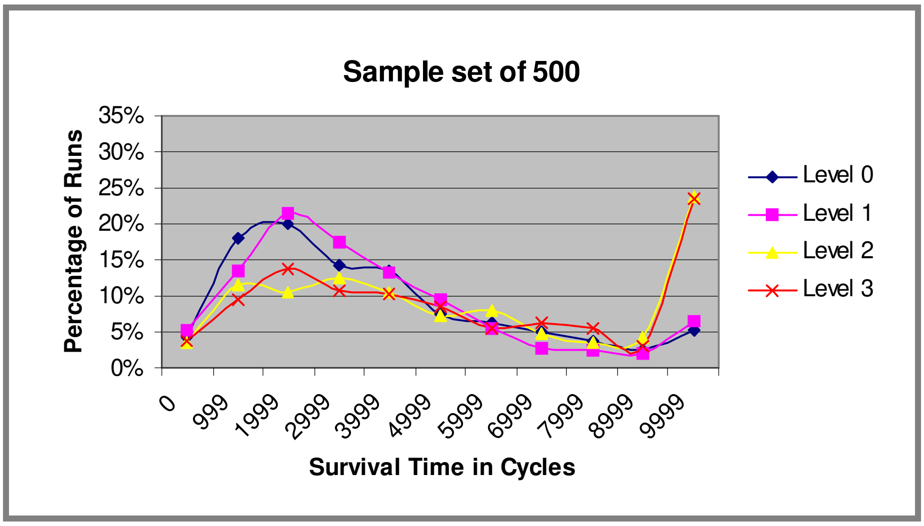


Figure 8.1-8 The Advantage of Level 2

Stressing The Architecture

This next set of experiments is designed to explore the niche-space by adding elements to the Gridland world that stress our design. Figure 8.1-9 shows the performance profile for Abbott as we add more enemies to the world, and then compensate with added lives. As we add more lives the relative advantage of the somatic marker learning mechanism in level 1 (over level 0) becomes apparent – as shown by the increased longer-term survival time for level 1 in Figure 8.1-9, and the delayed peak of level 1 in Figure 8.1-10.

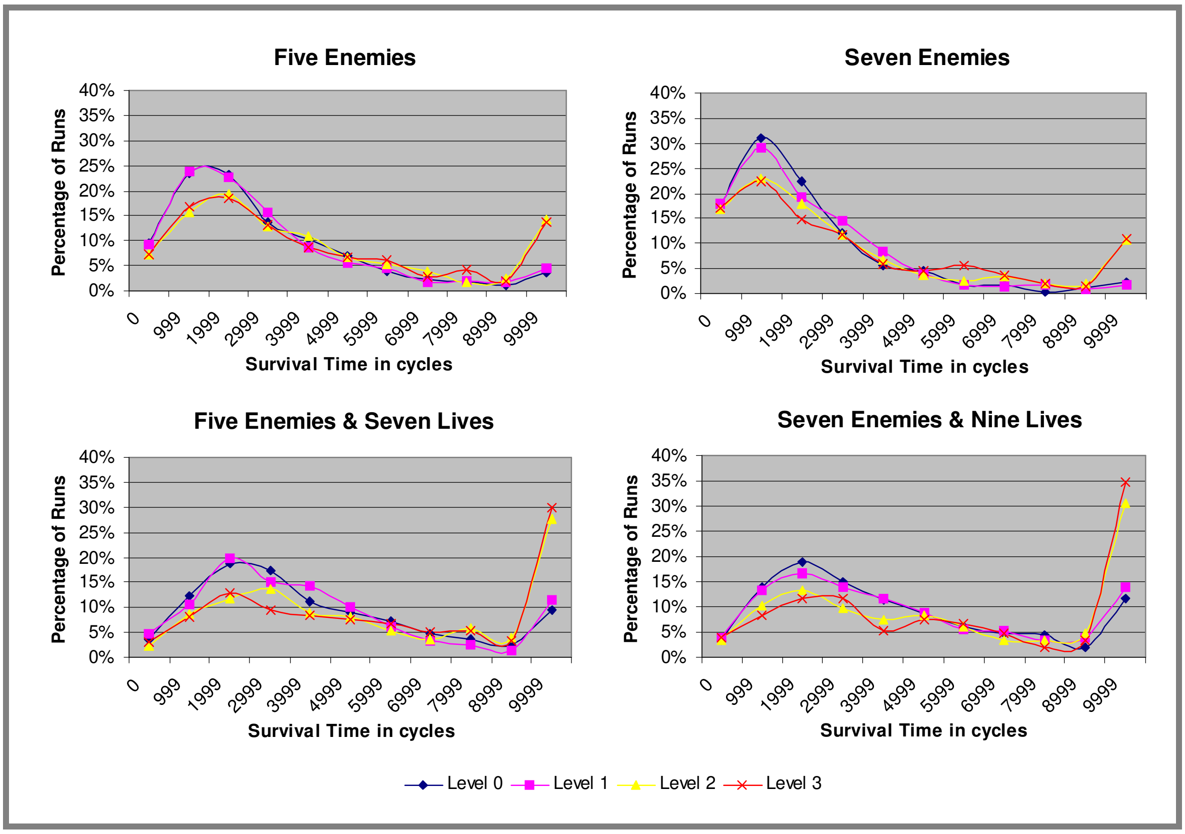


Figure 8.1-9 Stressing the Architecture

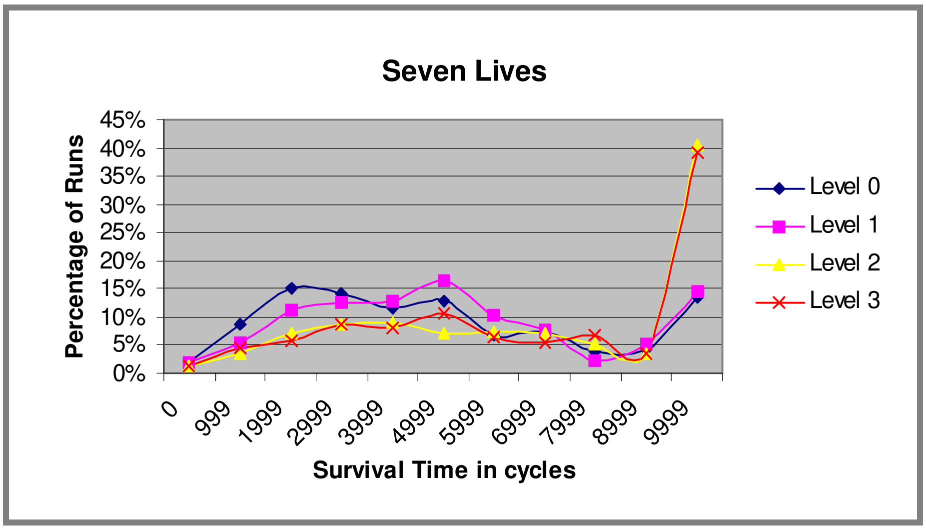


Figure 8.1-10 The Relative Advantage of Somatic Markers in Level 1

We are also able to change the dynamics of niche space by giving the enemies an excess of lives – effectively making them immortal (enemies can attack each other, or be attacked by Abbott, and so die within the normal course of an experimental run). Figure 8.1-11 gives the survival time profile for the standard experimental set-up with three immortal enemies. The profile of the first half of the graph is reassuringly similar to the standard case in Figure 8.1-8, with the effect of the immortality of the enemies only showing itself as the simulation progresses. Although the longer-term survivability of the Abbott architecture is much reduced, the comparative advantage shown by the level 2 architecture is still clearly visible.

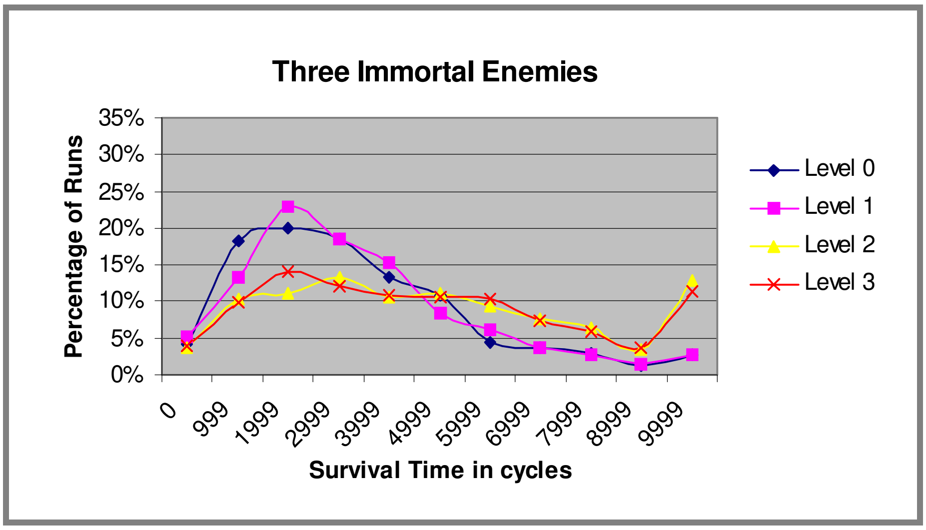


Figure 8.1-11 With Enemies that do not die

Contribution of Individual Agents

It is only natural that the higher competence levels are able to support more complex behaviour – hence the ability of *finder* agents to look both left and right. However, this has the inconvenient side-effect of making the relative utility of the motivator processing mechanisms inherent in the different levels hard to measure. By removing agents from the society, we can attack the problem from a different angle.

Figure 8.1-12 shows the survival time profile for a level 2, and level 3, Abbott with various agents removed from the architecture. Removing the *action proposer* agent effectively removes all the reactive action selection functionality from the Abbott architecture, leaving just the low level perception intact (the 'no reactive' trace on the graph). Even in this state, Abbott is still able to survive better than in the simpler competence level 1 case (see Figure 8.1-8), however its performance is considerably impaired against the complete level 2 or 3 architecture. The advantage of the reactive competencies are further emphasised when we look at the case in which the *skill* agents have been removed. Here, the infant mortality rate is more pronounced (the earlier peak in the graph), as a direct result of the removal of the reactive 'withdraw' skill from service. Removing the *relevance evaluation* agent does have a significant effect on the longer-term survivability of Abbott (see Figure 8.1-13), but not enough to validate the hypothesis that the advantage gained by the level 2 competence is due to Abbott's ability to spot enemies earlier (without the *relevance evaluation* agent, enemies are not evaluated as a source of pain and so spotting them early makes no difference).

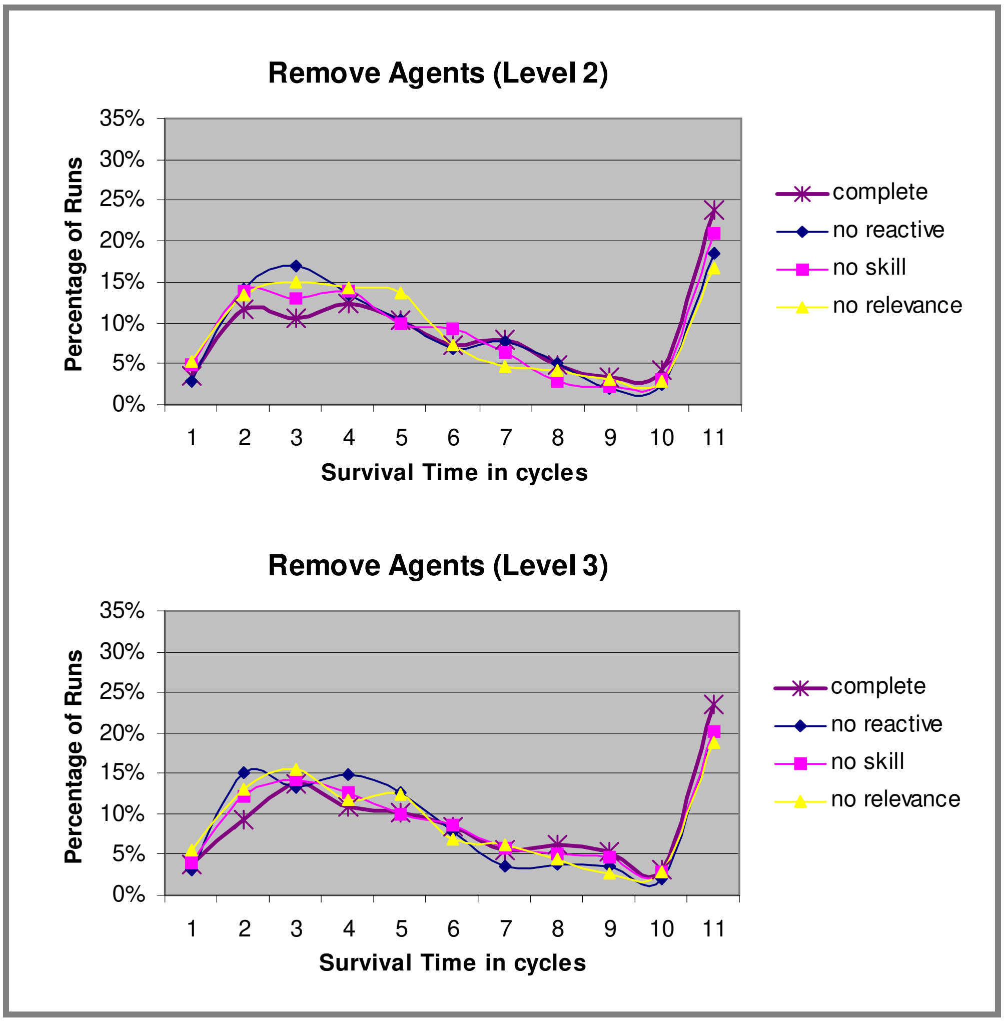


Figure 8.1-12 Removing Individual Agents

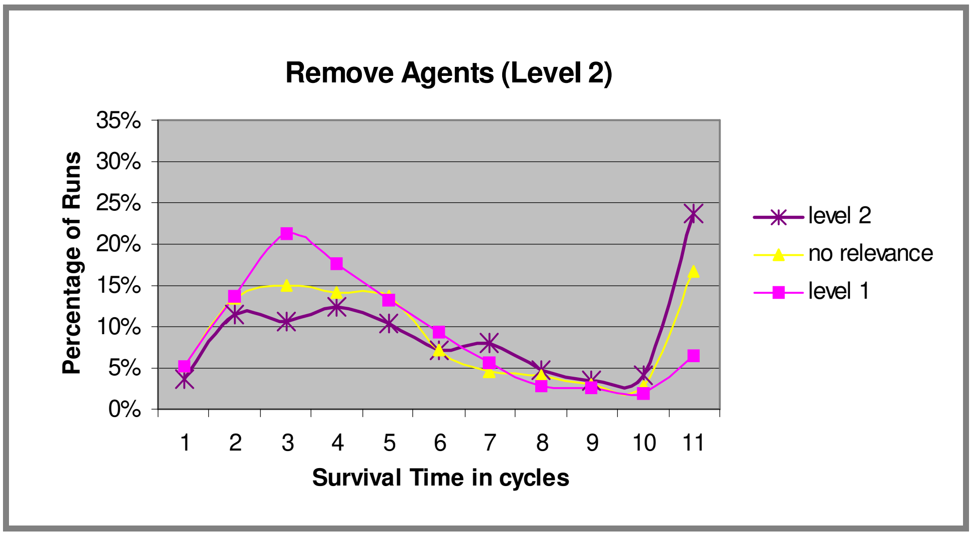


Figure 8.1-13 Contribution of Relevance Evaluation Agent

We can also repeat the experiment in a different part of niche-space – i.e. with more enemies or more lives. As we would expect, the relative loss of the *relevance evaluation* agent is more pronounced when Abbott has more lives with which to learn from experience (see Figure 8.1-14)

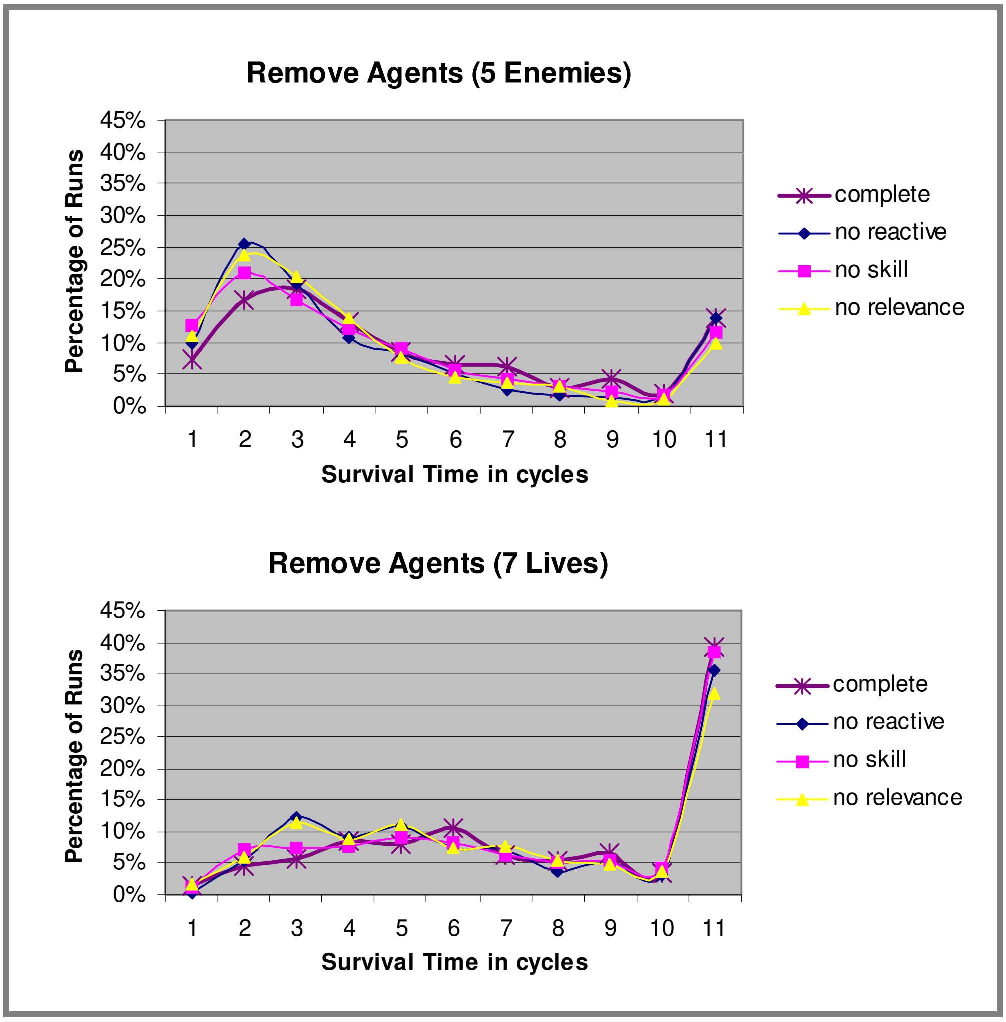


Figure 8.1-14 Removing Agents Under Different Conditions

Although the above results clearly demonstrate the usefulness of the reactive sub-system within the Abbott architecture, we have yet to show the utility of the *motivator manager* agent over the simple 'winner-takes-all' strategy adopted in Abbott2.

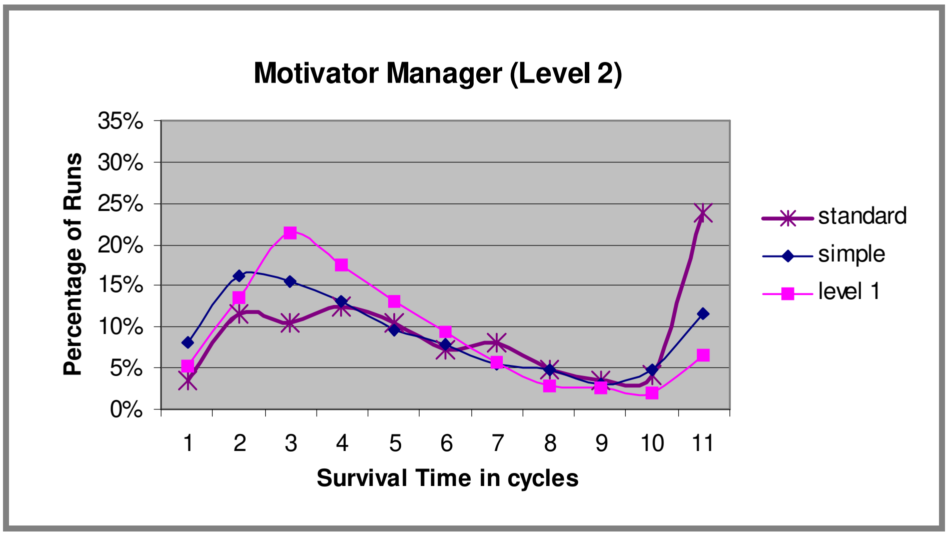


Figure 8.1-15 Simple Motivator Manager

Figure 8.1-15 shows the profile for the Abbott architecture with only a simple 'winner-takes-all' selection mechanism. To achieve the desired effect, the *filter* agent was modified to only allow a motivator to penetrate if it: a) had the highest activation energy; and b) was different to the current adopted motivator. Then, by using the *motivator meta manager* agent to reset the managed list of motivators, we guarantee that any motivator that surfaces through the filter is immediately adopted. These two simple changes allow us to create the 'winner-takes-all' selection mechanism with minimal disruption to the existing society members, giving us a fair degree of confidence that no other competencies were dramatically impaired by the operation. As we would expect, the active management of motivators does not account for all the gain exhibited by the level 2 architecture, but it does nevertheless make a significant contribution to the longer-term survivability of the agent. The increase in infant mortality shown by the 'winner-takes-all' Abbott, over the level 1 Abbott, is probably a result of the decision making time associated with the *manager* agents – without the persistence associated with the management of motivators, the motivator deciding process occupies a greater percentage of the overall time, making Abbott more vulnerable to attack as it sits thinking about what to do next.

Attention Filter Setting

The relative advantage of the meta-management layer is not immediately obvious from the experimental results shown thus far. This can, to a large extent, be attributed to the relatively minor role meta-management plays in the current implementation (see Figure 8.1-6). The main rationale for the meta-management layer is to adapt the motivator management process to changes in the internal and external environment – i.e. as an agent gains more experience. As Table 8.1-1 and Table 8.1-2 show, one particularly sensitive component of the motivator management layer is the filter relaxation parameters (the amount by which the threshold of the filter is decreased between cycles). The relaxation rate of *attention filter* agent is controlled by two components – a fixed amount, and a percentage of the current value. We can gain a rough measure of the performance of the filter by looking at the average survival time for the worst 70% of runs (eliminating all runs over 10,000 cycles) – we also include the percentage of runs greater than 10,000 and less than 1,000 in brackets. The default filter setting is 2% and 0.05 fixed (shown in italics).

Level 2		Fixed		
		0.00	0.02	0.05
Percentage	0%	3644 (18%, 6%)	3989 (24%, 5%)	3940 (21%, 4%)
	2%	3432 (15%, 4%)	3751 (24%, 4%)	*3928 (24%, 4%)*
	5%	3457 (16%, 5%)	3500 (20%, 4%)	3581 (21%, 5%)
	7%	3581 (17%, 5%)	3496 (18%, 6%)	3966 (21%, 3%)

Table 8.1-1 Filter Relaxation Parameters for Level 2

Level 3		Fixed		
		0.00	0.02	0.05
Percentage	0%	3362 (18%, 6%)	3791 (22%, 4%)	3761 (22%, 4%)
	2%	3594 (18%, 4%)	3583 (20%, 4%)	*3992 (23%, 4%)*
	5%	3253 (16%, 5%)	3605 (20%, 4%)	3767 (21%, 4%)
	7%	3250 (18%, 5%)	3582 (20%, 5%)	3770 (23%, 3%)

Table 8.1-2 Filter Relaxation Parameters for Level 3

Changing the parameters can have a dramatic effect on the survival rate of Abbott – decreasing the average survival time by as much as 20% (4000 down to 3250). The profile plots in Figure 8.1-16 also seem to indicate that different parameter settings might be more advantageous at different stages of the agents life. However, it is more likely that the parameters are better correlated to the internal state of the agent – i.e. if the meta-management layer detects a stressed state or not.

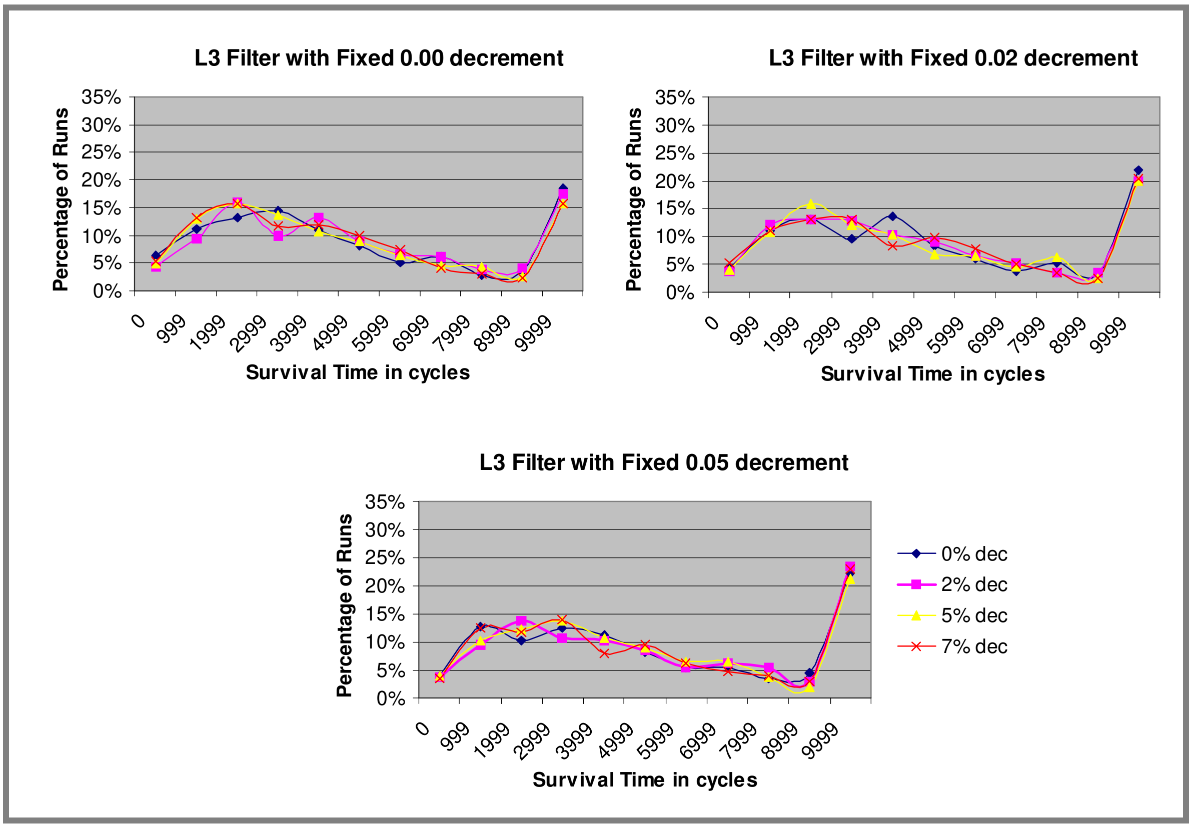


Figure 8.1-16 Filter Relaxation Parameters

Emergent States

Having established the 'utility' of Abbott's different competence levels, we can now describe in more detail the types of internal interactions that lead to what we term emergent "emotion-like" states. Figure 8.1-2 shows a typical example of the generation of a *primary* "emotion" state. The presence of a bright object (an enemy) causes Abbott's level 0 competence level to register 'danger', release adrenaline, and activate the *withdraw skill* agent with a high activation energy level – gaining control precedence over the higher competence levels. If Abbott is then subsequently bitten, a somatic marker is generated associating the enemy as the source of pain.

Once the somatic marker has been established, the mere presence of an enemy can then lead to a reactivation of the 'self-preservation' circuits through the "as if" loop. Figure 8.1-17 shows Abbott responding to the presence of a percept matching a known somatic marker (in this case enemy and pain – see the sample data trace). The debug trace reports the fact that the enemy has generated a bad memory of pain, with a strength proportional to the distance of the percept and strength of the marker. The *action proposer* agent triggers the *pain sensor* agent to register an "as if" pain somatic state which in turn causes the *self-protection drive* agent to be adopted.

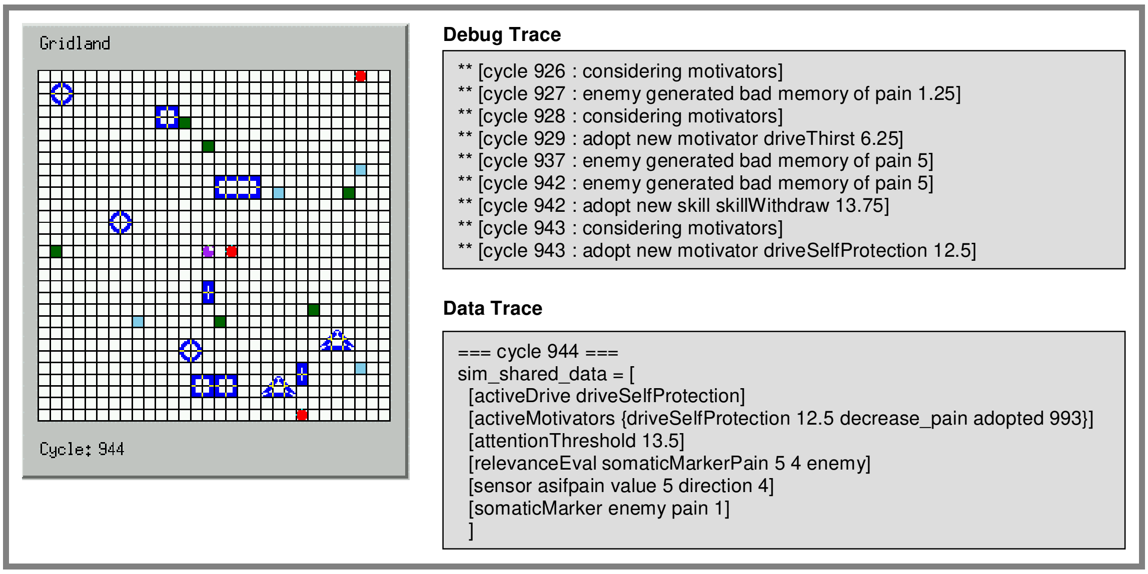


Figure 8.1-17 Generating a Self-Protection drive from a Marked Percept

If the *relevance evaluation* agent registers a more urgent threat – signified by a greater activation energy due to a closer percept or stronger somatic marker – then the *action proposer* agent will trigger the relevant skill directly, gaining control precedence over the level 2 competence level. Figure 8.1-18 depicts the generation of such a secondary "emotion-like" emergent state. The adoption of the withdraw skill in cycle 942 provides our agent with a critical window of opportunity over the slower adoption of the level 2 motivator in the next cycle (once adopted the level 2 motivator still needs to initiate action).

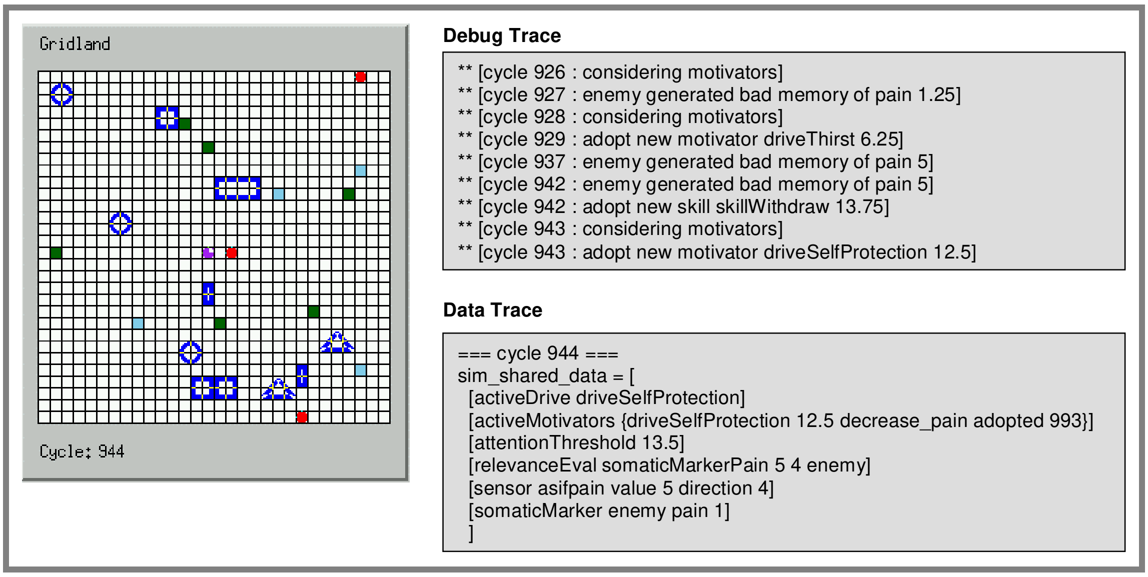


Figure 8.1-18 Generation of a Secondary "Emotion-like" State

Motivators are initially assigned an activation energy level based on relatively simple heuristics within the reactive competence level (such as deviation from a set point or distance of percept). This localised assignment can sometimes lead to a mismatch between the urgency attached to a motivator by the reactive level, and the more global consideration of the motivator management layer. Figure 8.1-19 shows the debug trace for such an emergent 'perturbant' state triggered after the motivator meta-manager agent has rejected a motivator which repeatedly surfaces through the filter. The fatigue motivator is adopted in cycle 859 and subsequently rejected by the motivator meta-manager 400 cycles later on cycle 1261 and the filter threshold is reset to 0.1 to allow other motivators to be considered by the motivator manager agent. Unfortunately, no other motivators are sufficiently insistent to penetrate the filter, and thus the agent is caught in a perturbant state rejecting the fatigue motivator every 50 cycles until another motivator can penetrate the filter and be adopted – the 50 cycle repetition is caused by the fact that the rejected motivator can only be re-adopted 50 cycles after being rejected (see Figure 8.1-6), but in this case, as there has been no change of motivator in the meantime it is immediately rejected again.

```
** [cycle 3 : considering motivators]
** [cycle 4 : adopt new motivator driveAggression 0]
** [cycle 142 : danger in direction 6]
** [cycle 142 : release adrenaline]
** [cycle 142 : adopt new skill skillWithdraw 20]
** [cycle 146 : ouch]
** [cycle 147 : danger in direction 5]
** [cycle 147 : release adrenaline]
** [cycle 148 : considering motivators]
** [cycle 148 : adopt new motivator driveSelfProtection 12.5]
** [cycle 153 : marking percepts]
** [cycle 153 : unable to resolve mark]
** [cycle 156 : stressed]
** [cycle 201 : considering motivators]
** [cycle 203 : considering motivators]
** [cycle 203 : adopt new motivator driveThirst 6.5]
** [cycle 206 : not stressed]
** [cycle 228 : adopt new motivator driveSelfProtection 6.5]
** [cycle 256 : considering motivators]
** [cycle 257 : adopt new motivator driveThirst 0.200027]
** [cycle 402 : danger in direction 3]
** [cycle 402 : release adrenaline]
** [cycle 476 : considering motivators]
** [cycle 478 : considering motivators]
** [cycle 479 : adopt new motivator driveHunger 3.72727]
** [cycle 856 : considering motivators]
** [cycle 858 : considering motivators]
** [cycle 859 : adopt new motivator driveFatigue 2.34768]
** [cycle 932 : adopt new skill skillFindRest 13.2337]
** [cycle 933 : considering motivators]
** [cycle 941 : stressed]
** [cycle 978 : considering motivators]
** [cycle 1033 : considering motivators]
** [cycle 1078 : considering motivators]
** [cycle 1133 : considering motivators]
** [cycle 1173 : considering motivators]
** [cycle 1218 : considering motivators]
** [cycle 1261 : rejecting driveFatigue as adopted too long]
** [cycle 1261 : considering motivators]
** [cycle 1263 : considering motivators]
** [cycle 1266 : not stressed]
** [cycle 1268 : considering motivators]
** [cycle 1273 : considering motivators]
** [cycle 1278 : considering motivators]
** [cycle 1283 : considering motivators]
** [cycle 1288 : considering motivators]
** [cycle 1293 : considering motivators]
** [cycle 1298 : considering motivators]
** [cycle 1303 : considering motivators]
** [cycle 1308 : considering motivators]
** [cycle 1313 : considering motivators]
** [cycle 1316 : rejecting driveFatigue as adopted too long]
** [cycle 1316 : considering motivators]
** [cycle 1318 : considering motivators]
** [cycle 1323 : considering motivators]
** [cycle 1328 : considering motivators]
** [cycle 1333 : considering motivators]
** [cycle 1338 : considering motivators]
** [cycle 1343 : considering motivators]
** [cycle 1348 : considering motivators]
** [cycle 1353 : considering motivators]
** [cycle 1358 : considering motivators]
** [cycle 1363 : considering motivators]
** [cycle 1368 : considering motivators]
** [cycle 1371 : rejecting driveFatigue as adopted too long]
** [cycle 1371 : considering motivators]
** [cycle 1373 : considering motivators]
** [cycle 1378 : considering motivators]
```

Figure 8.1-19 Trace of an Emergent Perturbant State

8.1.4 Strengths and Weaknesses

Autonomous Agency Credentials

Abbott3 represents a deepening, as well as a broadening, of Cañamero's [97] original architecture within the conceptual setting of the design-based research methodology and motivated agent framework described in **part I**. Abbott's autonomous agency credentials can readily be summarised as:

a) *Handling multiple sources of motivation*: Abbott3 is capable of handling multiple sources of homeostatic and non-homeostatic motivation on two different levels. At the reactive level, the *action proposer* agent is able to respond to signals from the *drive* agents, external events and somatic markers from the *relevance evaluation* agents, and the current state of deliberative processing from the *context evaluation* agents. At the deliberative level, the *motive manager* agent is able to respond to multiple sources of motivation by selecting *behaviour* agents that satisfy multiple concerns – the *attention filter* agent prevents the deliberative management resources from becoming swamped with trivial/non-urgent sources of motivation.

b) *Having and pursuing an agenda*: Abbott3's immediate agenda is determined by its active set of motivators pursed by the *motive manager* agent. However, Abbott can also be said to follow a more sophisticated agenda across all the competence levels of the architecture (see section 7.1.1). Abbott's level 0 purpose in life is to maintain a set of physiological variables within a desired range by responding to error signals generated by *drive* agents (with additional safety-based concerns represented through *sensor* agents and *relevance evaluation* agents). Abbott's level 1 concern-processing mechanisms allow the society to adapt to the environment through *somatic marker* agents. Competence level 2 determines Abbot's mode of deliberation. Finally, Abbott's level 3 concern-processing mechanisms pursue an agenda of continually fine-tuning the internal workings of the society.

c) *Robustness and adaptability in the face of a hostile and uncertain environment*: Abbott3 supplements Cañamero's original design with: i) affect-based learning, and ii) two new processing layers (reactive and meta-management) to enhance the robustness and adaptability of the architecture.

Design Heritage

In designing Abbott3, we have tried to pay particular attention to the lessons that could be learnt from existing agent designs – a key strand of the design-based research methodology. Cañamero's original design provided a valuable starting point, the motivated agent framework provided the structure, and the design analyses in chapters 3-5 provided the inspiration for Abbott3.

Abbott3 adopts the *vertical* decomposition strategy of the subsumption architecture [Brooks 86] – but along the lines of horizontal concern-processing competence layers rather than behaviours. The problems of inadequate command fusion normally associated with such an approach are avoided by both grounding the concern-processing mechanisms of each layer in the layer below (allowing the layers to co-evolve) and providing a global alarm mechanism to allow the primary level 0 competence layer to attain control precedence. By partitioning the Abbott design along the lines of concern-processing competence levels, we are also able to reduce the disruption to lower layers as a new layer is added – higher layers tend to represent new competencies that modify/enhance, rather than completely subsume, the competencies of the lower layers.

In a subsumption architecture, higher-level competencies again control precedence over lower-level competencies simply by inhibiting the output of lower-level behaviours (with the loss of all information inherent in the behaviour). In Abbott3, higher-level competencies can still gain control precedence over lower-level competencies, but only by modifying the input to the lower-level agents – i.e. re-focusing attention on a different part of the scene. Abbott3 is thus able to deliberate over actions in the normal course of events, and still respond rapidly to urgent sources of motivation at the reactive level.

The multi-layer approach also pays dividends in reducing the complexity of each of the individual competence layers, allowing Abbott3 to avoid the accountancy problems associated with spreading activation energy networks, without having to compromise on reactivity. Abbott's competence level 0 layer can concentrate on providing the global reactivity for the system based on the innate concerns/goals of the agent. Competence level 1 focuses on learning significant objects in the environment. Level 2 is free to provide deliberation and behaviour selection, with level 3 providing the self-monitoring self-adapting functionality. Competence level grounding also helps to ensure that all the layers work together for the good of the society/agent.

Earlier work by the Cognition and Affect Project [Beaudoin 94; Wright 97] concentrated on the requirements for goal/motivator processing within autonomous agents – identifying the need for a rich representation of motivational control states to support motivator deciding and scheduling. This research has adopted a broader objective, looking at the concern-processing requirements of autonomous agent designs. One side-effect of this stance is that the motivator management strategy used within Abbott is weaker than it could be. Motivators are simply adopted (or rejected by meta-management) with little further consideration of their scheduling requirements.

Although Abbott is capable of learning and modifying the internal structure of the society, there is currently no mechanism for adding new society members short of hand-coding them. Abbott cannot experiment with new behaviours or provide a mechanism that allows *manager* and *behaviour* agents to morph into *skill* agents. In principle it should be possible to add a new competence layer to manage the dynamics of the society or even allow individual agents to mutate and replicate. However, it is not immediately obvious how such a mechanism could be grounded in the level 0 competence layer – except by restricting the range and nature of such additions/deletions/modifications.

Exploration of Design and Niche Space

With the experiments so far performed, we have started to build a picture of the design space of the Abbott3 architecture – and evaluate Abbott's relative performance in niche space. We have shown that the flexibility of the society-of-mind architecture, and the support offered by the extended SIM_AGENT toolkit, does provide the required degrees of freedom with which to explore these two related spaces. We have also been able to demonstrate the utility of the Abbott design and our motivated agent framework, and shown that subtle changes in the architecture (i.e. filter relaxation functions) can have a profound effect on the overall survival time of our agent – and how these effects change when different stresses are introduced into the agent's environment. Even with our relatively simple agent, and 'toy' environment, complex control states soon start to emerge. As we learn more about the interaction of the agents and layers, we will be better placed to understand these emergent states, in the recursive style of the design-based research methodology.

8.1.5 Conclusions

We have presented the implementation details, some experimental results, and a brief critique of our Abbott3 design from the perspective of its contribution to the requirements of autonomous agency. However, our design should be seen as more than just an exercise in elucidating concern-processing in autonomous agents – Abbott is also a vehicle through which we wish to explore and describe human-like mental states and processes. Although we still have a long way to go before we can claim to actually support or explain infant-like emotional states, we feel that some progress towards a better understanding of the mental phenomena we label "emotion" has been made. In the next section we will look at some of lessons we can draw from Abbott, and their implications on the requirements for basic human emotions.

8.2 Requirements for Basic Human Emotions

In chapter 5 we argued that emotions are emergent mental states associated with a particular class of information-processing architectures broadly synonymous with our motivated agent framework. Although we have focused on the role emotional control states play in detecting and communicating urgent/relevant situations/events within an agent architecture, emotional expression also plays an important role in the communication of one agent's internal state to other agents. This dual communication role places very different requirements on the expressive nature of emotional states: (a) the internal affective channel needs to be broad and diffuse with many different pathways to allow very subtle affective states to be generated and communicated; whereas (b) the external affective channel needs to be narrow and focused to communicate a clear and easy to interpret message. Furthermore, the universality of the recognition and expression of some emotion types has led many researchers to speculate on the existence of "basic" human emotions – which at first sight appears contradictory to the emergent nature of emotions we have continued to paint.

Are There Basic Human Emotions?

Some emotion types (happiness, sadness, fear, anger, grief, and guilt) clearly appear across many different cultures, leading some emotion theorists to suggest that there are basic human emotions. Taken literally, this would mean that the carefree happiness of a child playing in the sea is the same emotion as the happiness experienced by an adult solving a hard cognitive puzzle. Both forms of emotional state may attract the same emotional label, but they will almost certainly utilise very different neural pathways and information-level representations in their generation and expression. However, both emotional states clearly do have something in common that gives them their distinctive *happiness* flavour, independent of their information-level pathway and emotion classification – i.e. the eliciting situations/ events are appraised as matching the same type of well-being concern.

Reactive concern matching is the role of the *central machinery of primary emotions* (see sections 5.2.3-5.2.5) – which is sometimes referred to as emotion-circuits in the literature (emotion-circuits can be a little misleading as different circuits share the same physical structures in the brain). The fact that there are only a few basic emotion types can be attributed to the fact that there are only a few emotion-circuits performing distinct types of reactive relevance evaluation. However, this does not mean that these emotions are basic in the sense that all other emotions are then mixes or blends of these universal emotion types.

What does Basic Really Mean?

All emotions are emergent mental states – in the sense that the emotional episode is the result of the interaction of a number of emotion specific, and non-specific, brain circuits. As the emotion process itself is a dynamic and ongoing process, it is therefore impossible to say where a distinct emotional episode starts and finishes – even basic emotions may take many different pathways (utilising the "as if" or the body loop) and subsequent conscious evaluation to give a myriad of internal hedonistic tones. At some point the fuzziness of the affective emotional state is de-fuzzified and an emotional type label is applied. If this defuzzification process produces a label that matches one of our universal emotion types, then we call the emotional experience a basic emotion.

The tertiary emotions of grief and guilt are often classified as basic human emotions. Grief appears universal because it emerges as a perturbant state created by the repeated interruption of attentive processing by the unconscious triggering of a *universal* attachment/ loss concern in the reactive concern-processing substrate [Wright et al. 96]. Even if we were to confine our search for basic emotions to primary emotional states, we must acknowledge that it becomes meaningless to attribute basic emotions to emotion-circuits when such circuits do not map on to distinct physical structures. It is therefore simply wrong to look for mechanisms that explain basic emotions and then expect to explain all other emotions as a mix/blend of these basic emotion mechanisms.

Emotional experiences are complicated by the fact that deliberation and past experiences play a significant role in our day-to-day activities, and so the vast majority of emotions we experience in everyday life will be *secondary emotions*. This partly explains why the concept of emotion blends is so appealing to theorists who simply look at the external expression of emotion without attempting to understand the internal information-level structures. Some emotions will certainly result from the activation of a single relevance evaluation mechanism and will be called basic emotions. Others will be the result of cultural conditioning or complex (re-)appraisals and could very likely result in the activation of a number of different emotion-circuits to give a mixed emotion. However, this still leaves those emotions that result from meta-management processes and/or are more cognitive in nature – i.e. where the basic emotion-circuits and body loop play a smaller role in the generation of the hedonistic tone of the emotional episode.

The beauty of the motivated agent framework and emotion process, as we described in chapter 5, is that it explains a plausible mechanism for the emergence of *all* emotional states – acknowledging the existence of universal emotion types, but without resorting to the need for blends and mixes to explain the rest. We have started to make progress, but there are still many more questions that need to be addressed before we can claim to understand the nature and requirements of basic human emotions.

Questions that still Need to be Addressed

The requirements for basic human emotions can be broken into two factors: nature and nurture (or architecture and society). In this thesis we have attempted to address the first of these factors within our motivated agent framework and Abbott architecture. To move forward, we must now start to address the second factor by embedding our agent in a socially rich environment – i.e. the Nursemaid scenario with which we opened this thesis. We will describe the requirements for such a scenario in the final chapter.

A number of important questions still need to be addressed as we continue to extend our Abbott architecture. Some of the more pressing questions are: (a) what are the level 0 concern-processing mechanisms needed for the generation of basic human emotion types?; (b) what forms of external expression are needed to reinforce these emotion types? (c) what internal reinforcers are required to facilitate the learning of secondary emotions?; (d) how do culture and emotional conditioning co-evolve?; (e) how much of the semantic content of the emotion process do we need to replicate – i.e. eliciting conditions, external/ physiological expression, and hedonistic tone – to qualify as a true emotional agent?; and finally (f) is it even meaningful to talk about a true emotional agent?

8.3 Summary

In this chapter we have presented the implementation details of our design of a cognitively inspired agent architecture for elucidating infant-like emotional states. We described a series of experiments showing how the different concern-processing competence levels contribute to the overall competence of our three-layered architecture, and identified the emergence of "emotion-like" states. We also presented a critique of our design, and started to address some of the requirements needed to support basic human emotional states.

We are still a long way from creating "emotional" agents with recognisably human mental states, but have taken some significant steps in the right direction.

Part IV

Conclusions

9 Conclusions

In this chapter, we summarise the main contributions our research makes to the ongoing task of elucidating the concern-processing requirements of intelligent autonomous agents. We also describe some of the possible directions in which we hope our work will be extended in the future.

9.1 Main Contributions of Thesis

This thesis makes a number of contributions to the field of concern-processing in both human and artificial autonomous agents. An extended overview of these contributions is given below in chronological order – with references to the supporting arguments to be found within the thesis itself.

Framework for Analysing and Designing Intelligent Autonomous Agents

There is a real and pressing need to develop a systematic framework within which we can compare, analyse, and design intelligent autonomous agents. Some significant progress has been made towards meeting this need by the ongoing research within the Cognition and Affect project – as we elucidate the architectural requirements needed to support human-like mental states and actions. In this thesis we present an extended *motivated agent framework*, consolidating and enhancing our earlier work in a number of important ways:

1) We consolidate and clarify the earlier work of the Cognition and Affect project by bringing together all the different strands of the motivated agent framework for the first time – chapters 1 and 2. Specifically, we: (a) present our design-based research methodology, and describe how it can be used to provide a powerful explanatory framework for elucidating complex systems such as intelligent autonomous agents – section 1.2; (b) describe how viewing the human mind as a complex control system allows the use of certain mentalistic terms and concepts to be justified by referring them to information-level descriptions of the underlying architecture – section 2.1; (c) introduce the concept of motivational control states and describe the *functional* attributes of some of the many control states that are likely to play an important role in intelligent autonomous agency architecture – section 2.1; and finally (d) describe a cognitively inspired three-layered architectural framework for elucidating the *structural*, *dimensional*, and *functional* attributes of these control states – section 2.2.

2) We argue for a concern-centric stance to autonomous agent design and provide a design-based analysis of *motivational* control states in both deliberative and behaviour-based agent architectures – chapter 3. We identify a number of problems with these designs that can in part be attributed to the traditional

approach of partitioning a design either along functional (section 3.1) or behavioural (section 3.2) lines, without due consideration of the concern-processing requirements of the agent. We address these problems with our design for an intelligent autonomous agent in chapter 7.

3) We provide an analysis of two broad intelligent agent designs by members of the Cognition and Affect project [Beaudoin 94; Wright 97] – chapter 4. We describe the importance of motivational control states in these architectures (section 4.1), and the emergence of affective states in relation to Sloman's [92] Attention Filter Penetration theory of emotion (section 4.3). We also identify the need for greater depth in both the perceptual and reactive concern-processing abilities of these agent architectures.

4) The human emotion process can be viewed as a classic example of an information-processing system geared towards "serving" concerns at all levels of an agent architecture. In chapter 5 we provide a broad requirements specification for such an emotion process and use recent theories from psychology and neurology [Frijda 86; Damasio 94; LeDoux 96] to explain the mechanisms inherent in the different classes of emotional states (*primary*, *secondary*, and *tertiary*) from an information-level design-based perspective. This chapter (a) adds depth to the motivated agent framework by making explicit the reactive concern-processing requirements; and (b) provides supportive evidence for our approach by mapping leading theories of emotion on to our framework.

5) Finally, we present our design for an intelligent autonomous agent (chapter 7) – building on the lessons learnt from chapters 3 through 6. This design extends the motivated agent framework by providing a concrete platform from which to test and analyse our theories about motivational and emergent emotional control states. We also start to bridge the gap between the fields of psychology, cognitive science, artificial intelligence, and neurology by mapping promising target brain regions on to this framework – see appendix C.

Analysis of Human and Artificial "Emotional" States

In the course of our research, we have collected and analysed a number of different theories of emotion. Our research can therefore also be seen as a contribution towards a better understanding of the emergent nature of human and artificial "emotional" states:

6) We provide an overview of previous work carried out by members of the Cognition and Affect project on emergent perturbant states within our motivated agent framework – section 4.3. We show how the information-level design-based approach can contribute to emotion research, and argue that the perturbant nature of some emotional states is afunctional – offering a warning against assigning

intrinsic function to the temporary loss of control sometimes associated with intense emotions.

7) We present a detailed information-level analysis of several important theories of emotion [Frijda 86; Damasio 94; LeDoux 96] and, by mapping them on to our motivated agent framework, show how these theories contribute to the emotion process puzzle – section 5.2. We argue that emotions are emergent mental states caused by the interaction of a variable number of intricately connected cognitive systems. We then provide an information-level description of these systems – sections 5.2.5 and 5.2.6.

8) We present an information-level design-based analysis of artificial "emotional" states in "emotional" autonomous agents [Moffat and Frijda 95; Velásquez 96; Breazeal and Velásquez 98; McCauley and Franklin 98; and Cañamero 97] – section 6.1. We also present an analysis of two broad-but-shallow implementations of an agent architecture that captures and extends Cañamero's original design – section 6.2. We identify a number of areas of weakness in these designs that can be attributed to the inclusion of a discrete "emotion" system.

9) We use our motivated agent framework to elucidate the architectural requirements for basic human emotions – section 8.2. We argue that some emotion types can be considered basic as a consequence of their universality, but that it does not then follow that all non-basic emotions are either mixes or blends of these universal emotions.

Toolkit for Building Intelligent Autonomous Agents

To support our experimental work, we designed and implemented a graphical front-end and development environment for the SIM_AGENT toolkit [Sloman and Poli 96]. This extended toolkit was then used in the development and analysis of the intelligent autonomous agent architectures described in this thesis. The extensions to the standard toolkit are described in appendix A.

Design of an Intelligent Autonomous Agent for Elucidating "Emotional" States

By designing and analysing complete agent architectures, we are able to provide a more tangible account of the phenomena of interest. Although this aspect of our research will remain ongoing, we have already made a number of important contributions through our design of an intelligent autonomous agents for elucidating "emotional" states:

10) We present the design of a cognitively inspired intelligent autonomous agent architecture partitioned along the lines of concern-processing competence. We show how such a design naturally supports *emergent* primary, secondary, and tertiary "emotional" states.

11) We use our intelligent autonomous agent architecture to elucidate the emergence of simple "emotional" states, and demonstrate how the richness of each supported state grows with the addition of new concern-processing competence layers – section 7.1.2.

12) Finally, we provide a concrete implementation of our design, and describe a series of experiments that: a) demonstrate the added value of each competence layer for the survival of the agent, and b) show that the interaction of these layers within the architecture leads to emergent "emotion-like" states.

9.2 Future Work

A good research programme should provide us with the sorts of new insights that allow us to ask better questions. In this section we will discuss some of the questions we now feel better placed to ask.

9.2.1 Autonomous Agency and Human Emotions

Although multi-layered agent architectures are not unusual in the field of autonomous agency, partitioning a design along the lines of concern-processing competence levels does add a new dimension to the game that has yet to be fully explored or exploited. By grounding each new competence layer in the concern-processing mechanisms of the base layer, we have created the necessary conditions to support emotional control states – providing both reactivity in the form of a global alarm system, and a useful mechanism for unsupervised learning. Having demonstrated the architectural requirements for artificial "emotions", we now need to explore in greater depth the utility of the emotion process, and understand its potential costs.

Replicators Replicate

We are all the survival machines of our selfish genes – replicators whose only purpose in life is to replicate (see appendix E). The key to the success of this replication process can in part be attributed to the fact that emotions influence our actions and colour our perception of the world – with the genetically pre-disposed machinery of primary emotions lying at the heart of the emotion process (see section 5.2.3). In principle, there is nothing to stop us using the machinery of primary emotions in our early stages of development, and then switching to the learnt associations between events and emotional markers to take us through the rest of life. To a certain extent this is exactly what happens – Damasio [94] calls this process the somatic marker hypothesis. However, our genes never completely hand over control, and secondary emotions still express themselves through the machinery of primary emotions.

An interesting direction in which this research can be taken in the future is to investigate the utility of grounding the upper concern-processing competence layers in the base (zero) layer as typified by the emotion process. Our analysis has pointed to useful gains to be made by such an approach (reactivity and a mechanism for unsupervised learning), but we have yet to demonstrate the practicality of capturing concerns at the base level, or explore the relative weighting of motivational control between the competence layers. Damasio [94] points to a useful role for "emotion" in automatically reducing our choice based on previous "affective" experiences (see section 5.2.5), but such conjectures still remain to be tested. Tertiary emotional states are created by situations in which a mismatch exists between the upper and lower concern-processing mechanisms. Grief may play a useful role in society (or rather having strong attachment concerns may be good for the gene pool), but it does not necessarily follow that it is equally advantageous to the individual.

Abbott offers a useful platform to testing the somatic marker hypothesis and determining the extent to which base level concern-processing mechanisms should have control over attentive processing. If we remove the requirement for reproduction, is it still necessary to always give our genes the upper hand?

Social Agents

One aspect of emotions that this research has not systematically addressed is the social/ communication nature of the emotion process. "Emotional" agent architectures [Moffat and Frijda 95; Velásquez 96; McCauley and Franklin 98; and Cañamero 97] typically attempt to capture human emotion types such as happy, angry, and sad. This is an entirely reasonable approach, as we will almost inevitably use the intentional stance [Dennett 87] in our day-to-day dealings with our agents, and so having a common baseline makes sense.

Our next task within our Abbott architecture will be to identify the number and type of *relevance evaluation* agents – the primary determinants of the *core-self*. Most emotion theorists are happy to accept that there are only a small number of such systems. For example, Gray [94, page 245] identifies three fundamental emotion systems: behavioural approach system; fight/flight system; and a behavioural inhibition system. Unfortunately, many theorists also attempt to associate emotion-circuits directly with emotions (or rather attempt to explain emotions only in terms of such circuits) – resorting to the terminology of basic emotions and emotion blends (see section 8.2). Once we have established the basic nature of the *relevance evaluation* agents, we can then turn our attention to the requirements for emergent human emotion types.

9.2.2 Nursemaid Scenario

The Gridland scenario (section 6.2) provided a useful tool for exploring the nature of the emotion process in a dynamic and hostile environment. However, it does not support the types of social conditions that are likely to lead to the emergence of recognisable human emotion types. We have therefore developed a simulated Nursemaid scenario based on the human scenario described in chapter 1 – with Abbotts taking the role of the infants in the nursemaids charge. Our next challenge is to grow the Abbott architecture into an architecture capable of supporting a competent nursemaid.

Technical Nursemaid Scenario

Our nursemaid's environment is the simulated world of the nursery. Within this environment, our nursemaid can perform a number of actions to affect what it senses in the future, i.e. pick up objects, wander from room to room, take Abbotts to food or water. The nursemaid can also be said to pursue an agenda which aims to satisfy a set of basic *concerns*: maintaining a healthy level of charge in its battery; keeping Abbotts away from the ditch; looking after the well-being of Abbotts; achieving goals; and exploring the nursery.

As with the Gridland scenario, the nursemaid, Abbott, and the different assorted objects, are all modelled as agents within the framework of an extended SIM_AGENT toolkit (see appendix A). Using a virtual environment allows us to abstract away many real world problems (the extended toolkit still provides a low-level sensory interface based on physical features of objects rather than convenient labels – we simply avoid problems such as shadows and reflections partially obscuring objects) and provides us with a controlled and repeatable environment for our experiments.

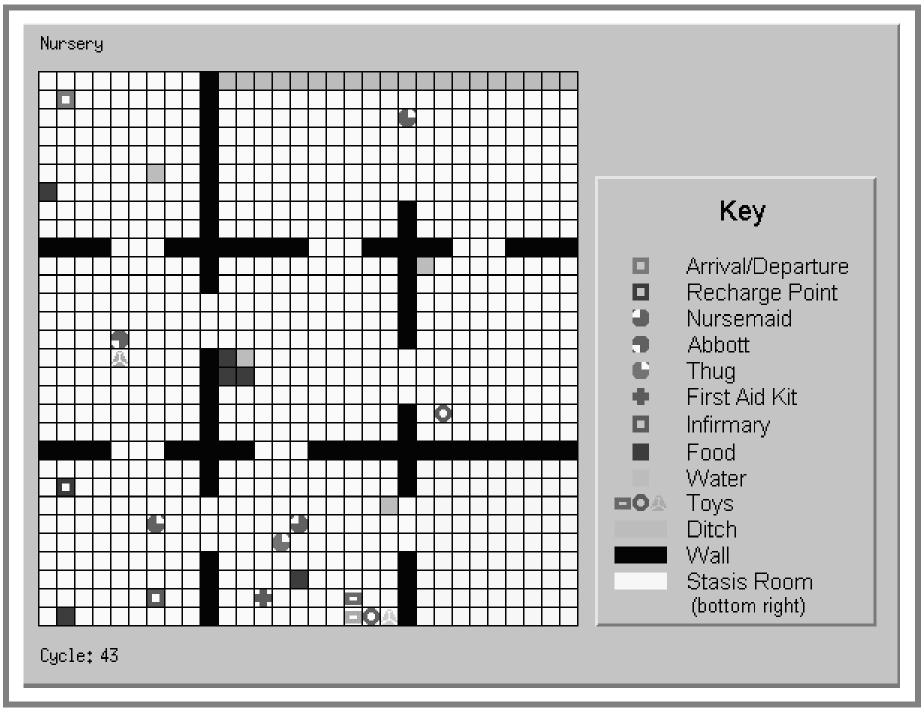


Figure 9.2-1 Technical Nursemaid Scenario

The scenario (Figure 9.2-1) includes a number of features specifically aimed at further extending the range of concern-processing requirements of our autonomous agents: (i) Abbotts, with their frequent needs, provide the nursemaid with multiple sources of asynchronous motivation; (ii) various objects (toys, infirmary, and first aid kit) provide both opportunities and different solutions for dealing with problems as and when they arise; (iii) the nursemaid must balance its own needs (it must remain recharged) against those of its charges; and (iv) the nursemaid can be stressed by the introduction of more Abbotts or the removal of resources (such as the first aid kit).

Additionally, the social aspects of the scenario provide a care-giving environment that is comparable to the situation in which real infants find themselves. Our conjecture is that this type of social environment is ideally suited for elucidating the emergence of basic human emotion types – similar to the approach adopted by Breazeal and Velásquez [98].

Details of the Nursery

The nursery consists of nine rooms, an arrival/departure point, a recharge point, a stasis room, an infirmary, and a ditch. Scattered around the nursery are a number of objects that can help the nursemaid in its duties.

Abbotts arrive at the arrival point and remain in the nursery until they reach a certain age or die. When Abbotts are old enough, or fall foul of the many dangers and die, they can be taken to the dismissal point. The infirmary (or first aid kit) can be used to heal injured Abbotts. The stasis room (bottom right-hand corner) can be used to place an Abbott in suspended animation allowing the nursemaid to regulate the number of adverse motivators in its environment.

The nursemaid has four effectors: a foot, a claw, a microphone, and a camera: (a) the foot moves the nursemaid around the nursery; (b) the claw can grab objects (including Abbotts) and transfer them around the nursery – taking a toy to an Abbott, or an Abbott to the infirmary; (c) the microphone can detect sounds from anywhere in the nursery; and (d) the camera (a 5x5 grid extending from the eye) can see into the room currently occupied by the nursemaid. Additionally, various short-range sensors (hardness, brightness, occupancy, etc.) can be used to probe the squares surrounding the nursemaid – giving a more detailed picture of the environment than is possible with the camera. Finally, the nursemaid moves at a finite speed, and must negotiate a path around walls, Abbotts, or other objects.

Requirements for a Competent Nursemaid

The scenario contains multiple sources of adverse motivation. It is assumed that the well-being of the Abbotts is a primary concern of a competent nursemaid, and so the nursemaid must remain vigilant to the prospects of: (a) Abbotts falling into the ditch; (b) Abbotts turning into thugs (when angry for too long); (c) Abbotts being injured by thugs; (d) Abbotts running low on food; and (e) the nursemaid running low on charge. Some sources of motivation will require fast reflexes (when an Abbott wanders too close to the ditch), and others more longer-term planning (i.e. moving from room to room to proactively search for potential trouble spots). There will be occasions when simply comforting an Abbott will be enough to calm the situation, and others when a toy or first-aid kit must be retrieved first – the nursery represents a very open environment.

Aside from the issue of handling multiple sources of motivation, a competent nursemaid must also attempt to understand the needs of its charges – is an Abbott crying because it is hungry, hurt, scared, or simply wants attention? It is in this context that we hope to make further inroads into the requirements for the emergence of the basic human emotion types – i.e. the actual emotion types supported by an architecture will be dependent on the concerns *and* communication needs of the agent.

9.3 Summary

In this chapter we have described: (a) the contributions our research makes to the fields of autonomous agency and emotion research; and (b) the directions in which we hope our research will be extended in the future. Although we cannot yet provide the answers to the emotion process puzzle, we are learning how to ask better questions – bringing us closer to our goal of elucidating concern-processing in intelligent autonomous agents.

10 List of References

Abbott, E. A. (1884). *Flatland: A Roman of Many Dimensions*. London: Seeley & Co., Ltd.

Agre, P. and Chapman, D. (1991). What are Plans for? In P. Maes (Ed.), *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*, pages 17-34. MIT Press, March 1991.

Alami, R., Chatila, R., Fleury, S., Ghallab, M., and Ingrand F. (1998). An Architecture for Autonomy. In *International Journal of Robotics Research*, Vol. 17, No. 4, pages 315-337, April 1998. Also as *LAAS Report N° 97352*.
(ftp://ftp.laas.fr/pub/ria/felix/publis/ijrr97.ps.gz)

Apter, M. J. (1989). *Reversal Theory: Motivation, Emotion and Personality*. London: Routledge.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Balkenius, C. (1993). The Roots of Motivation. Summary published in J-A. Mayer, H. L. Roitblat, S. W. Wilson, (Eds.), *From Animals to Animats 2*, Cambridge, MA: MIT Press, 1993.
(http://lucs.fil.lu.se/ftp/pub/Papers/Christian/RootsOfMotivation.ps)

Bateson, G. (1972). *Steps to an Ecology of Mind*. San Francisco: Chandler.

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)

Bloom, F. E. (1995). Cellular Mechanisms Active in Emotion. In Gazzaniga, M. S. (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: The MIT Press, pages 1063-1070.

Blumberg, B. (1996). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD Thesis. Media Lab, Massachusetts Institute of Technology.
(http://characters.www.media.mit.edu/groups/characters/thesis/blumberg_phd.pdf)

Blumberg, B. (1994). Action Selection in Hamsterdam: Lessons from Ethology. In Cliff, D. and Husbands, P. (Eds.), *Animals to Animats: Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior*.
(http://characters.www.media.mit.edu/groups/characters/papers/sab94.pdf)

Blumberg, B., Todd, P., and Maes, P. (1996). No Bad Dogs: Ethological Lessons for Learning in Hamsterdam. In *Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, Cape Cod, MA, September 1996. MIT Press/ Bradford Books. (http://characters.www.media.mit.edu/groups/characters/papers/sab96.pdf)

Bogner, M. B. (1998). *Creating a "conscious" agent*. Masters Thesis, The University of Memphis. (http://www.rhodesalumni.org/~myles/memphis/CreatingAConsciousAgent.pdf)

Bogner, M. B. (1999). *Realizing "consciousness" in software agents*. PhD Thesis, The University of Memphis. (http://www.rhodesalumni.org/~myles/memphis/RealizingConsciousnessInSoftwareAgents.pdf)

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: The MIT Press.

Bratman, M. E. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E., Israel, D. J., and Pollack, M. E. (1988). Plans and Resource-Bounded Practical Reasoning. *Computational Intelligence Journal*, Vol. 4, No. 4, pages 349-355, 1988. (http://www.ai.sri.com/pubs/papers/Brat88-349:Plans/document.ps.gz)

Breazeal C., and Velásquez, J. (1998). Towards Teaching a Robot "Infant" using Emotive Communication Acts. In the *Proceedings of the Socially Situated Intelligence Workshop, SAB '98*, Zurich, Switzerland. (http://www.ai.mit.edu/people/jvelas/papers/Breazeal-Velasquez-SAB98.ps)

Brodal, A. (1982). *Neurological Anatomy*. New York: Oxford University Press.

Brooks, R. A. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, Vol. RA-2, No. 1, pages 12-23. (ftp://publications.ai.mit.edu/ai-publications/500-999/AIM-864.ps)

Buck, R. (1985). Prime Theory: An Integrated View of Motivation and Emotion. *Psychological Review*, Vol. 92, No. 3, pages 389-413.

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)

Chalmers, D. J. (1996, 97). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press. (first published 1996, Oxford: Oxford University Press)

Cohen, P. R., and Levesque, H. J. (1990). Intention is choice with commitment. In *Artificial Intelligence*, Vol. 42, No. 3.

Complin, C. (1997). The Evolutionary Engine and the Mind Machine: A Design-based Study of Adaptive Change. Ph.D. Thesis. The University of Birmingham. (http://www.cs.bham.ac.uk/research/cogaff/Complin.thesis.ps.gz)

Damasio, A. R. (1994, 96). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac. (first published 1994, New York: G. P. Putman's Sons.)

Damasio, A. R. (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace & Company.

Dawkins, R. (1982). *The Extended Phenotype*. Oxford: Oxford University Press.

Dawkins, R. (1986). *The Blind Watchmaker*. London: Penguin.

Dawkins, R. (1989). *The Selfish Gene*. Oxford: Oxford University Press.

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

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and The Meaning of Life*. London: Allen Lane – The Penguin Press.

Dennett, D. C. (1996). *Kinds of Minds: Towards an Understanding of Consciousness*. London: Weidenfeld and Nicolson.

Denton, D. (1993). *The Pinnacle of Life: Consciousness and Self-Awareness in Humans and Animals*. St. Leonards, Australia: Allen and Unwin Pty Ltd.

Dux, G. (1990). *Die Logik der Weltbilder* (3rd Edition). Frankfurt: Suhrkamp Verlag.

Edelman, G. M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Edelman, G. M. (1992). *Bright Air, Brilliant Fire: On the Matter of the Mind*. London: Penguin.

Ekman, P., Friesen, W. V., and Ellsworth, P. (1982). Does the face provide accurate information? In P. Ekman (Ed.), *Emotion in the Human Face* (2nd edition). Cambridge: Cambridge University Press, pages 56-97.

Ekman, P. (1992). Are There Basic Emotions? *Psychological Review*, Vol. 99, No. 3, pages 550-553.

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)

Ellsworth, P. C. (1994). Levels of Thought and Levels of Emotion. In P. Ekman, and R. J. Davidson (Eds.), *The Nature of Emotion*. Oxford: Oxford University Press, pages 192-196.

El-Nasr, M. S. (1998). *Modeling Emotion Dynamics in Intelligent Agents*. Masters Thesis, Texas A & M University.

Eysenck, H. J. (1991). Dimensions of Personality: 16, 5, or 3? – Criteria for a Taxonomic Paradigm. In *Personality and Individual Differences*, Vol. 12, pages 773-790.

Firby, J. R. (1989). Adaptive Execution in Complex Dynamic Worlds. Ph.D. Thesis *Yale University Technical Report, YALEU/CSD/RR #672,* January 1989
(http://people.cs.uchicago.edu/users/firby/thesis/thesis.ps.Z)

Ferguson, I. A. (1992). Touring Machines: An Architecture for Adaptive, Rational, Mobile Agents. *PhD Thesis* – Technical Report 273, Computer Laboratory, University of Cambridge, UK, April, 1992.
(http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/TR273-iaf-thesis.ps.gz)

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

Franklin, S., and Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the Third International Workshop on Agents, Theories, Architectures, and Languages*. Springer-Verlag, pages 21-35.
(http://www.msci.memphis.edu/~franklin/AgentProg.html)

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.

Frijda, N. H. (1994). Emotions are Functional, Most of the Time. In P. Ekman, and R. J. Davidson (Eds.), *The Nature of Emotion*, Oxford: Oxford University Press, pages 112-122.

Frijda, N. H., and Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. In *Cognition and Emotion*, Vol. 1, No. 3, pages 235-257.

Georgeff, M. P., and Ingrand, F. F. (1989). Decision-making in an embedded reasoning system. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pages 972-982). Detroit, MI: IJCAL. Also as *Tech. Rep. 04*, Australian Artificial Intelligence Institute, Melbourne, Australia, Nov 1989. (ftp://www.aaii.oz.au/pub/aaii-technotes/technote04.ps.gz)

Georgeff, M. P., and Lansky, A. L. (1986). Procedural Knowledge. In *Proceedings of the IEEE: Special issue on Knowledge Representation*, Vol. 74, No. 10, pages 1383-1398.

Georgeff, M. P., and Rao, A. S. (1995). BDI Agents: From Theory to Practice. *Tech. Rep. 56*, Australian Artificial Intelligence Institute, Melbourne, Australia, April 1995. (ftp://www.aaii.oz.au/pub/aaii-technotes/technote56.ps.gz)

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin Company.

Goodall, J. (1991). *Through a Window: Thirty Years with the Chimpanzees of Gombe*. London: Penguin.

Gray, J. A. (1994). Three fundamental emotion systems. In P. Ekman, and R. J. Davidson (Eds.), *The Nature of Emotion*. Oxford: Oxford University Press, pages 243-247.

Hayes-Roth, B. (1985). A blackboard architecture for control. In *Artificial Intelligence*, Vol. 26, No. 3, pages 251-321.

Hayes-Roth, B. (1995). An Architecture for Adaptive Intelligent Systems. In *Artificial Intelligence: Special Issue on Agents and Interactivity*, Vol. 72, pages 329-365. (ftp://ftp-ksl.stanford.edu/pub/KSL_Reports/KSL-93-19.ps)

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)

Huber, M. J., Lee, J., Kenny, P., Durfee, E. H. (1995). UM-PRS V2.9 Programmer and User Guide. *Tech. Report*, Artificial Intelligence Laboratory, The University of Michigan.

Ingrand, F. F., Chatila, R., Alami, R., and Robert, F. (1996). PRS: A High Level Supervision and Control Language for Autonomous Mobile Robots. In *IEEE ICRA 96, Minneapolis, USA*. Also as *LAAS Report N°96027.* (ftp://ftp.laas.fr/pub/Publications/1996/96027.ps)

Izard, C. E. (1992). Basic Emotions, Relations Among Emotions, and Emotion-Cognition Relations. In *Psychological Review*, Vol. 99, No. 3, pages 561-565.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. In *Psychological Review*, Vol. 100, No. 1, pages 68-90.

Jackson, J. V. (1987). Idea for a Mind. *SIGGART Newsletter*, No. 181, pages 23-26.

Johnson-Laird, P. N., Oatley, K. (1992). Basic Emotions, Rationality, and Folk Theory. In *Cognition and Emotion*, Vol. 6, No. 3/4, pages 201-223.

Jung, C. (1999). *Theory and Practice of Hybrid Agents*. Ph.D. Thesis, The University of Saarland. (http://www.dfki.de/~jung/publications/doctoral99.ps.gz)

Kaelbling, L. P., and Rosenschein, S. J. (1991) Action and Planning in Embedded Agents, in P. Maes (Ed.) *Designing Autonomous Agents*, The MIT Press. Also in *Robotics and Autonomous Systems*, Vol. 6, No. 1.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (1995). *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton and Lange.

Kitano, H. (1995). *A Model for Hormonal Modulation of Learning*. Sony Technical Report SCSL-TR-95-044. (ftp://ftp.csl.sony.co.jp/CSL/CSL-Papers/95/SCSL-TR-95-044.ps.gz)

Konolige, K., Myers, K. L., Ruspini, E. H., and Saffiotti, A. (1997). The Saphira architecture: A design for autonomy. *Journal of Experimental and Theoretical Artificial Intelligence*. Vol. 9, No. 1, 1997, pages 215-235. (ftp://aass.oru.se/pub/saffiotti/robot/jetai96.ps.gz)

Kumar, D. (1993). *From Beliefs and Goals to Intentions and Actions: An Amalgamated Model of Inference and Acting*. Ph.D. Dissertation, Department of Computer Science, State University of New York at Buffalo, Buffalo, NY 14260, 1993. (ftp://ftp.cs.buffalo.edu/pub/tech-reports/94-04.ps.Z)

Kumar, D. and Shapiro, S. C. (1994a). The OK BDI Architecture. *International Journal of Artificial Intelligence Tools*, Vol. 3, March 1994, pages 349-366. (http://serendip.brynmawr.edu/~dkumar/ijait.ps)

Kumar, D. and Shapiro, S. C. (1994b). Acting in Service of Inference (and *vice versa*). In Dankel and Stewman (Eds.), *Proceedings of FLAIRS-94* (Seventh Florida AI Research Symposium), Florida AI Research Society, May 1994. (http://serendip.brynmawr.edu/~dkumar/flairs.ps)

Larsen, R. L., and Ketelaar, T. (1989). Extraversion, neuroticism and susceptibility to positive and negative mood induction procedures. *Personality and Individual Differences*, Vol. 10, pages 1221-1228.

LeDoux, J. E. (1987). Emotion. In *Handbook of Physiology I: The Nervous System – Vol. V: Higher Functions of the Brain*. Bethesda, MD: American Physiological Society, pages. 419-460.

LeDoux, J. E. (1992). Brain mechanisms of emotion and emotional learning. *Current Opinion in Neurobiology*, No. 2, pages 191-198.

LeDoux, J. E. (1994). Cognitive-Emotional Interactions in the Brain. In P. Ekman, and R. J. Davidson (Eds.), *The Nature of Emotion*, Oxford: Oxford University Press, pages 216-223.

LeDoux, J. E. (1995). In Search of An Emotional System in the Brain: Leaping from Fear to Emotion and Consciousness. In Gazzaniga, M. S. (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: The MIT Press, pages 1049-1062.

LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Ludlow, A. (1976). The behavior of a model animal. In *Behavior*, Vol. 58.

Ludlow, A. (1980). The evolution and simulation of a decision maker. In F. Toates, and T. Halliday (Eds.), *Analysis of Motivational Processes*. London: Academic Press.

MacLean, P. D. (1952). Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). In *Electroencephalogr. Clinical Neurophysiology*, Vol. 4, pages 407-418.

Maes, P. (1989). How to Do the Right Thing. In *Connection Science Journal*, Vol. 1, No. 3, pages 291-323, 1989. Also MIT AI-Memo 1180. December 1989. (http://agents.www.media.mit.edu/groups/agents/Publications/Pattie/consci/consci.ps)

Maes, P. (1994). Modeling adaptive autonomous agents. In C. Langton (Ed.), *Artificial Life Journal*, Vol. 1, No. 1&2, pages 135-162 (http://pattie.www.media.mit.edu/people/pattie/alife-journal.ps)

Maslow, A. H. (1954). *Motivation and Personality*. Harper. (3rd Edition, Addison-Wesley, 1987).

Mauro, R. (1988). Opponent process in human emotions? An experimental investigation of hedonic contrast and affective interaction. In *Motivation and Emotion*, 12, pages 333-351.

McCauley, T. L. (1999). *Implementing Emotions in Autonomous Agents*. Masters thesis, The University of Memphis. (http://www.msci.memphis.edu/~mccaulet/thesis.ps)

McCauley, T. L. and Franklin, S. (1998). An Architecture for Emotion. In *Working Notes of AAAI Fall 1998 Symposium*. (http://www.msci.memphis.edu/~mccaulet/Emotion.PDF)

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. In *Journal of Personality*, Vol. 60, pages 175-215.

Meyer, G. J., and Shack, J. R. (1989). Structural convergence of mood and personality: evidence for old and new directions. In *Journal of Personality and Social Psychology*, Vol. 57, pages 691-706.

Minsky, M. (1985, 87). *The Society of Mind*. London: William Heinemann Ltd. (first published 1985, New York: Simon & Schuster.)

Myers, K. L. (1996). A procedural knowledge approach to task-level control. In *Proceedings of the Third International Conference on AI Planning Systems*, AAAI Press, 1996. (http://www.ai.sri.com/~prs)

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.

Moffat, D., and Frijda, N. H. (1995). Where there's a will there's an agent. In M. J. Wooldridge, and N. Jennings (Eds.), *Intelligent Agents, ECAI-94. Lecture Notes in Artificial Intelligence 890*. New York: Springer-Verlag, pages 245-260. (http://www.cf.ac.uk/uwc/psych/moffat)

Morignot, P. and Hayes-Roth, B. (1994). Why does an agent act? In *Knowledge Systems Laboratory Report KSL-94-76*, December 1994. (ftp://ftp-ksl.stanford.edu/pub/KSL_Reports/KSL-94-76.ps)

Morignot, P. and Hayes-Roth, B. (1996). Motivated Agents. In *Knowledge Systems Laboratory Report KSL-96-22*, July 1996. (ftp://ftp-ksl.stanford.edu/pub/KSL_Reports/KSL-96-22.ps)

Müller, J. P. (1996). *The Design of Intelligent Agents: A Layered Approach*. New York: Springer-Verlag.

Newell, A. (1982). The knowledge level. In *Artificial Intelligence*, Vol. 18, pages 87-127.

Nilsson, N. J. (1994). Teleo-Reactive Programs for Agent Control. In *Journal of Artificial Intelligence Research*, Vol. 1, pages 139-158. (http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume1/nilsson94a.pdf)

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.

Oatley, K., and Jenkins, J. M. (1996). *Understanding Emotions*. Cambridge, MA: Blackwell Publishers Ltd.

Ono, T., and Nishijo, H. (1992). Neurophysiological basis of the Klüver-Bucy syndrome: Responses of monkey amygdaloid neurons to biologically significant objects. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction*. New York: Wiley-Liss, pages 167-190.

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

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: The MIT Press.

Prem, E. (1996). Motivation, Emotion and the Role of Functional Circuits in Autonomous Agent Design Methodology. *TR-96-04*, The Austrian Research Institute for Artificial Intelligence, Vienna, 1996.
(ftp://ftp.ai.univie.ac.at/papers/oefai-tr-96-04.ps.gz)

Rao, A. S., and Georgeff, M. P. (1995). Formal models and decision procedures for multi-agent systems. *Tech. Rep. 61*, Australian Artificial Intelligence Institute, Melbourne, Australia, June 1995.
(ftp://www.aaii.oz.au/pub/aaii-technotes/technote61.ps.gz)

Rao, A. S., and Georgeff, M. P. (1991a). Deliberation and intentions. *Tech. Rep. 10*, Australian Artificial Intelligence Institute, Melbourne, Australia, May 1991.
(ftp://www.aaii.oz.au/pub/aaii-technotes/technote10.ps.gz)

Rao, A. S., and Georgeff, M. P. (1991b). Modeling rational agents within a BDI-architecture. *Tech. Rep. 14*, Australian Artificial Intelligence Institute, Melbourne, Australia, Feb 1991.
(ftp://www.aaii.oz.au/pub/aaii-technotes/technote14.ps.gz)

Revelle, W. (1995). Personality Processes. In *Annual Review of Psychology*, Vol. 46, pages 295-328.
(http://pmc.psych.nwu.edu/revelle/publications/AR.html)

Rhodes, B. (1996). *PHISH-Nets: Planning Heuristically in Situated Hybrid Networks*. MS thesis. School of Architecture and Planning, Massachusetts Institute of Technology, September 1996.
(http://rhodes.www.media.mit.edu/people/rhodes/PHISH-Nets)

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, pages 109-114.

Robbins, T. W., and Everitt, B. J. (1995). Arousal System and Attention. In Gazzaniga, M. S. (Ed.) *The Cognitive Neurosciences*. Cambridge, MA: The MIT Press, pages 703-720

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. In *Philosophical Transitions of the Royal Society, London, Series B, Biological Sciences 335*, pages 11-21.

Rolls, E. T. (1995). A Theory of Emotion and Consciousness, and its Application to Understanding the Neural Basis of Emotion. In Gazzaniga, M. S. (Ed.) *The Cognitive Neurosciences*. Cambridge, MA: The MIT Press, pages 1091-1106.

Rolls, E. T. (1999). *The Brain and Emotion*. Oxford: Oxford University Press.

Rosenblatt, J. (1998). Behavior-Based Planning for Intelligent Autonomous Vehicles. To appear in *Proceedings of IAV'98 Symposium on Intelligent Autonomous Vehicles*, Madrid, Spain, March 27-29, 1998. (http://www.umiacs.umd.edu/~julio/papers/IAV98-formatted.ps.gz)

Rosenblatt, J. (1997). *DAMN: A Distributed Architecture for Mobile Navigation*. Ph.D. Thesis, Carnegie Mellon University Robotics Institute, Pittsburgh, PA. Also as *Technical Report CMU-RI-TR-97-01*. (http://www.cs.cmu.edu/afs/cs.cmu.edu/user/jkr/www/papers/Dissertation/DAMN_Diss.ps.gz)

Rosenblatt, J., and Thorpe, C. (1995). Combining Multiple Goals in a Behavior-Based Architecture. In *Proceedings of 1995 International Conference on Intelligent Robots and Systems (IROS)*, Pittsburgh, PA, August 7-9, 1995. (http://www.umiacs.umd.edu/~julio/papers/IROS_95.ps.gz)

Rosenblatt, J., and Payton, D. W. (1989). A Fine-Grained Alternative to the Subsumption Architecture for Mobile Robot Control. *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, Washington DC, June 1989, Vol. 2, pages 317-324. (http://www.umiacs.umd.edu/~julio/papers/Fine_Grained_Alternative.ps.gz)

Rosenschein, S. J., and Kaelbling, L. P. (1995). A Situated View of Representation and Control. In *Artificial Intelligence*, Vol. 73, 1995. (http://www.cs.brown.edu/people/lpk/sitaut.ps)

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)

Sacks, O. (1990). *Seeing Voices*. London: Picador.

Selfridge, O. G. (1959). Pandemonium: A Paradigm for Learning. In *Proceedings of the Symposium on Mechanisation of Thought Process*. National Physics Laboratory.

Schachter, S. (1964). The interaction of cognitive and physiological determinants of emotional states. In Berkowitz, L. (Ed.) *Advances in Experimental Social Psychology Vol. 1*. New York: Academic Press, pages 49-80.

Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum.

Simon, H. (1967). Motivational and emotional controls of cognition. Reprinted in *Models of Thoughts*, Yale University Press, (1979), pages 29-38.

Simon, H. (1982). Affect and cognition: Comments. In M. S. Clark, and S. T. Fiske (Eds.), The Seventeenth Annual Carnegie Symposium on Cognition: Affect and Cognition. London: Lawrence Erlbaum Associates, pages 333-342.

Sloman, A. (1992). Prolegomena to a Theory of Communication and Affect. In: Ortony, A., Slack, J., and Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*. Heidelberg, Germany: Springer-Verlag, pages 229-260. (ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_Prolegomena.ps.Z)

Sloman, A. (1993a). 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)

Sloman, A. (1993b). Prospects for AI as the General Science of Intelligence. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsey (Eds.), *Prospects for Artificial Intelligence*. Amsterdam: ISO Press, pages 1-10 (ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_prospects.ps.Z)

Sloman, A. (1997). *Designing Human-Like Minds*. University of Birmingham. (ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Sloman.ecal97.ps.gz)

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. (http://www.cs.bham.ac.uk/research/cogaff/Sloman.kd.pdf)

Sloman, A. and Croucher, M. (1981). Why robots will have emotions. In *Proceedings IJCAI 1981*, Vancouver. Also available as *Cognitive Science Research paper No 176*, Sussex University. (ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_why_robot_emotions.ps.gz)

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)

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. (ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/Aaron.Sloman_Riccardo.Poli_sim_agent_toolkit.ps.Z)

Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. In *American Psychologist*, 35, 691-712.

Śmieja, F. (1996). The Pandemonium System of Reflective Agents. In *IEEE Transactions on Neural Networks*, Vol. 7, No. 1, pages 97-106, 1996. Also available as *Technical Report 794*, German National Research Centre for Computer Science (GMD), St. Augustin, Germany, 1994.
(ftp://borneo.gmd.de/pub/as/janus/ref94_2.ps)

Swanson, L. W. (1983). The hippocampus and the concept of the limbic system. In W. Seifert (Ed.), *Neurobiology of the Hippocampus*. London: Academic Press, pages 3-19.

Thorpe, S. J., Rolls, E. T., and Maddison, S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. In *Experimental Brain Research* 49, pages 93-115.

Tomkins, S. S. (1981, 95) The Quest for Primary Motives: Biography and autobiography of an idea. In E. V. Demos (Ed.), *Exploring Affect: The Selected Writings of Silvan S. Tomkins*. Cambridge: Cambridge University Press, 1995. Originally published in the *Journal of Personality and Social Psychology*, Vol. 41, No. 2, pages 306-329, 1981.

Tomkins, S. S. (1984). Affect Theory. In K. R. Scherer and P. Ekman (Eds.), *Approaches to Emotion*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Turner, T. J., and Ortony, A. (1992). Basic Emotions: Can Conflicting Criteria Converge? In *Psychological Review*, Vol. 99, No. 3, pages 566-571.

Tyrrell, T. (1992). Defining the Action Selection Problem. In the *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates.
(http://www.soc.soton.ac.uk/SOES/STAFF/tt/AS/tt1.ps.Z)

Tyrrell, T. (1993a). *Computational Mechanisms for Action Selection*. PhD Thesis, University of Edinburgh.
(ftp://ftp.ed.ac.uk/pub/lrtt/)

Tyrrell, T. (1993b). The Use of Hierarchies for Action Selection. In the *Journal of Adaptive Behavior*, Vol. 4, 1993. Also to appear in J-A. Meyer, S. W. Wilson, and H. Roitblat (Eds.), *the Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. MIT Press/Bradford Books, 1993.
(http://www.soc.soton.ac.uk/SOES/STAFF/tt/AS/tt2.ps.Z)

Tyrrell, T. (1994). An Evaluation of Maes' "Bottom-Up Mechanism for Behavior Selection". In *Adaptive Behavior*, Vol. 2, No. 4, pages 307-348, 1994. Also as *Technical Report*, Plymouth Marine Laboratory.
(http://www.soc.soton.ac.uk/SOES/STAFF/tt/AS/tt3.ps.Z)

Velásquez, J. (1996). *Cathexis: A Computational model for the Generation of Emotions and their Influence in the Behavior of Autonomous Agents.* Masters Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Velásquez, J. (1997). Modeling Emotions and Other Motivations in Synthetic Agents. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.
(http://www.ai.mit.edu/people/jvelas/papers/Velasquez-AAAI-97.ps)

Velásquez, J. (1998). Modeling Emotion-Based Decision-Making. In *Proceedings of the 1998 AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition* (Technical Report FS-98-03). Orlando, FL: AAAI Press.
(http://www.ai.mit.edu/people/jvelas/papers/Velasquez-FS98.ps)

Velásquez, J. (1999). When Robots Weep: Emotional Memories and Decision-Making. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, pages 70-75.
(http://www.ai.mit.edu/people/jvelas/papers/Velasquez-AAAI-98.ps)

Watson, D., Clark, L. A., and Harkness, A R. (1994). Structures of personality and their relevance to psychopathology. In *Journal of Abnormal Psychology*, Vol. 103, pages 1-14.

Wilkins, D. E. and Myers, K. L. (1995). A common knowledge representation for plan generation and reactive execution. In *Journal of Logic and Computation*, Vol. 5, No. 6, pages 731-761, December 1995.
(http://www.ai.sri.com/~wilkins/papers/jlc-www.ps)

Wilkins, D. E., Myers, K. L., Lowrance, J. D., and Wesley, L. P. (1995). Planning and reacting in uncertain and dynamic environments. In *Journal of Experimental and Theoretical AI*, Vol. 7, No. 1, pages 197-227, 1995.
(http://www.ai.sri.com/pubs/papers/Wilk95-197:Planning/document.ps)

Wright, I. P. (1997). *Emotional Agents*. PhD Thesis, School of Computer Science, University of Birmingham.
(http://www.cs.bham.ac.uk/research/cogaff/Wright.thesis.ps.gz)

Wright, I. P., Sloman, A., and Beaudoin, L. P. (1996). Towards a Design-Based Analysis of Emotional Episodes. In *Philosophy Psychiatry and Psychology* 3(2):101-126. This is a revised version of the paper presented to the *Geneva Emotions Workshop, April 1995* entitled “The Architectural Basis for Grief.”
(http://www.cs.bham.ac.uk/research/cogaff/Wright_Sloman_Beaudoin_grief.ps.gz)

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. In *American Psychologist*, Vol. 35, pages 151-175.

Appendices

A Extended SIM_AGENT Toolkit

The Gridland toolkit provides the SIM_AGENT toolkit [Sloman and Poli 96] with a graphical interface and simulated environment in which to explore the design-space of autonomous agent architectures. The toolkit has been heavily influenced by Cañamero's work on the Gridland World [Cañamero 97], growing out of the design requirements for the initial implementation of the Abbott architecture. The key benefits offered by the toolkit are:

- A mouse driven interface
- Run, pause, and single step a simulation.
- Load, save, and reset a simulation.
- Set trace and debug options for any agent.
- Multiple windows to display the agent's internal status.
- Controllable scheduler loop for real-time interactions.
- Interactive control and display of agent status.
- Uses the standard SIM_AGENT toolkit.
- Easily expandable.
- Command batch operation.

A.1 Virtual Machines

The various files that form the Gridland toolkit are best described within the context of a stack of "virtual machines" – machines with no definable physical form – with the operating system at the bottom and the target autonomous agent architecture at the top. This whole structure is shown graphically in Figure A.1-1 below.

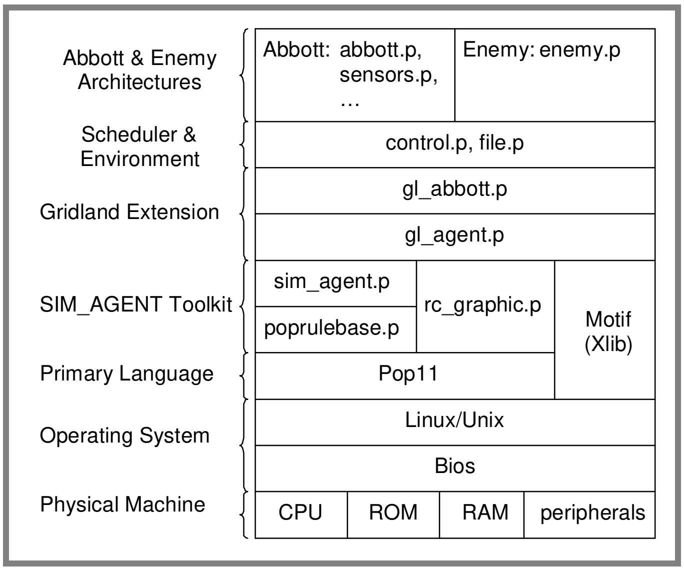


Figure A.1-1 The Gridland "Virtual Machine"

The first layer of the Gridland virtual machine is the Linux/Unix operating system, which in turn supports the Pop11 programming language – a flexible stack-based language which has some similarity to LISP. On top of the Pop11 virtual machine sits the SIM_AGENT toolkit (which provides the rule-based virtual environment used to support the architecture of the agent). The Gridland extension comes next, adding the object classes needed to support the Gridland environment and providing routines to access the Gridland data-structures and Motif graphical interface. The Scheduler and the Environment provide the simulator virtual machine – controlling the operation of the SIM_AGENT virtual machine below it. Finally, at the top of the tower sit the Abbott and Enemy virtual machines – implemented as a society of agents running condition-action rules.

A.2 The SIM_AGENT Toolkit

The SIM_AGENT toolkit is a general purpose toolkit for investigating different types of agent architectures. The toolkit supports both rule-based and sub-symbolic (i.e. neural) mechanisms, and comes with extensive on-line help and teach files. A description of the toolkit can be found in Sloman and Poli [96] – the main features are summarised below:

- *Minimal ontological commitment* supporting many different kinds of objects with very different architectures.
- *External behaviour* which can be detected by or affect other objects or agents.
- *Internal behaviour* involving mechanisms for changing internal control states (percepts, beliefs, maps, goals, …) that are not directly detectable by others.
- *Rich internal architecture within agents* allowing several rule-sets and rule-families to run in simulated parallelism. An architecture can therefore support several levels of sensory perception, reactive and deliberative processes, neural nets and other trainable sub-mechanisms.
- *Use of classes and inheritance* to allow generic behaviour.
- *Control of relative speed* allowing both agents and sub-mechanisms within agents to run at different relative speeds.
- *Rapid prototyping* through the incremental compilation of the Pop11 environment.

A.3 The Gridland Extension

The **gl_agent.p** and **gl_abbott.p** libraries form the Gridland extension to the SIM_AGENT toolkit. These extensions define access mechanisms to the Gridland data structures, the Gridland object classes, and provide a high-level interface to the Motif widget set for menu and window creation.

The Gridland World

The Gridland world is an arbitrary cellular structure which maps on to a window in the Gridland simulated environment: (a) the cell size and grid dimensions are user definable, allowing a one-to-one or one-to-many mapping between cells and pixels; (b) agents can occupy one or more cells allowing irregularly shaped objects to be moved as single entities; and (c) cells can contain one or more objects – a foreground object and any number of background objects (allowing Abbott to share a cell with a ditch; or the Nursemaid to share a cell with the recharger).

As building 'world models' should remain in the domain of the agent architecture, only the physical characteristics of agents are stored in Gridland. Gridland cells are thus described by the vector {*Occupancy*, *Brightness*, *Hardness*, *Organic*, *Agent_id*}. This ensures that perception, and indeed misperception, are still interesting problems as agents do not come ready labelled (the *agent_id* field is only used to identify the target agent that has been eaten or pushed etc.). Finally arbitrarily sized Gridland worlds are supported by providing two methods of storing Gridland cells: (i) an array for efficient indexing; or (ii) in sparsely populated large worlds, cells are stored local to each agent – in which case all agents must be tested to determine the occupants of a particular cell.

Object Classes and The Graphical Interface

The Gridland environment makes extensive use of the Motif widget set and **rc_graphic.p** (Relative Co-ordinate Graphic) library. A number of high-level routines are included to aid the creation of popup and cascading pulldown menus, as well as access to the trace and debug scrollable windows. Object classes, mixins, and methods, for the physical attributes of the Gridland environment as well as the Motif widget set are stored in the **gl_agent.p** library. Classes, mixins, and methods, specific to the Abbott architecture (the *Society of Mind* model as well as support for physiological variables such as blood_pressure, heart_rate) are stored in the **gl_abbott.p** library.

Batch Operation

Commands can be entered through the graphical user interface, or as a parameter list passed to the main start routine. This latter batch mode, allows a series of experiments to be defined in advance (including adding and/or removing agents from the Abbott architecture) and then executed sequentially. Further speed gains can be made by running the batch mode without any graphical support – any interesting results can be repeated and analysed later in the graphical mode.

A.4 The Gridland Scheduler and Environment

The scheduler and rules for the scenario/environment form the final part of the Gridland toolkit. The Gridland scheduler hooks into the standard SIM_AGENT toolkit (see Figure A.4-1) to provide the toolkit with a powerful mouse-driven graphical interface. A simple command queue is used to synchronise mouse and menu events with the scheduler.

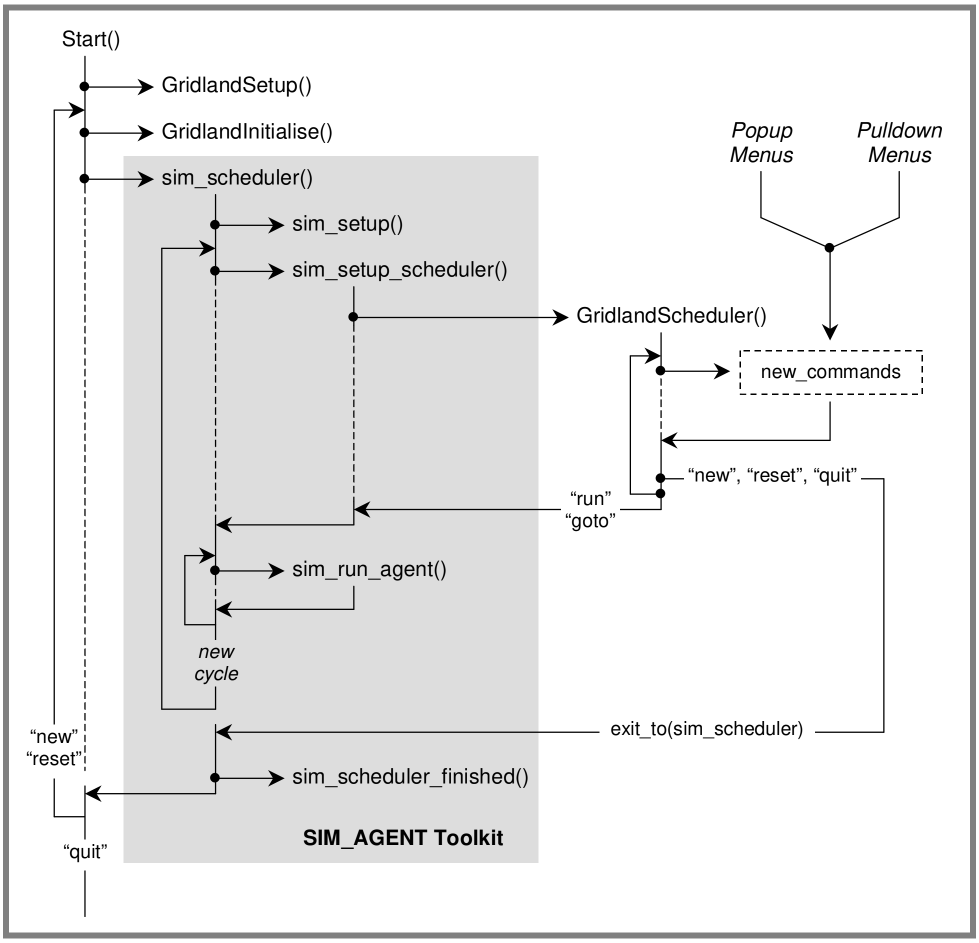


Figure A.4-1 Gridland Scheduler and the SIM_AGENT toolkit

Commands and Menus

The Gridland commands fall into three basic categories: (i) file commands associated with resetting, loading and saving experiments (Figure A.4-2); (ii) run commands concerned with pausing, single stepping and running the simulation (Figure A.4-3); and (iii) system commands concerned with setting cycle times and saving trace and debug results to disk (Figure A.4-4). In addition to these basic commands, other menu options can be used to select between different status windows (Figure A.4-5), display a list of all agents, or provide simple help facilities.

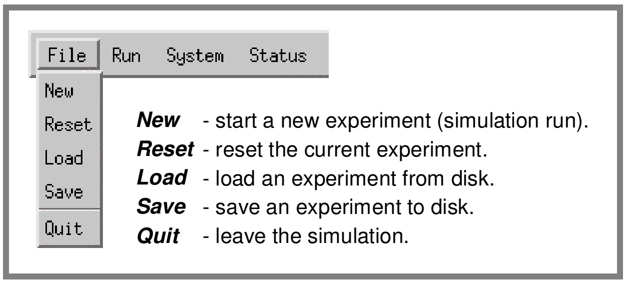


Figure A.4-2 Gridland File Menu

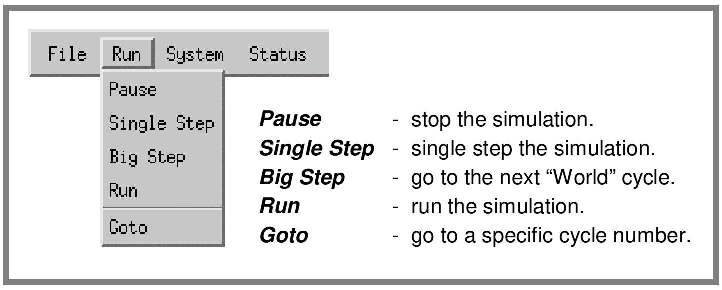


Figure A.4-3 Gridland Run Menu

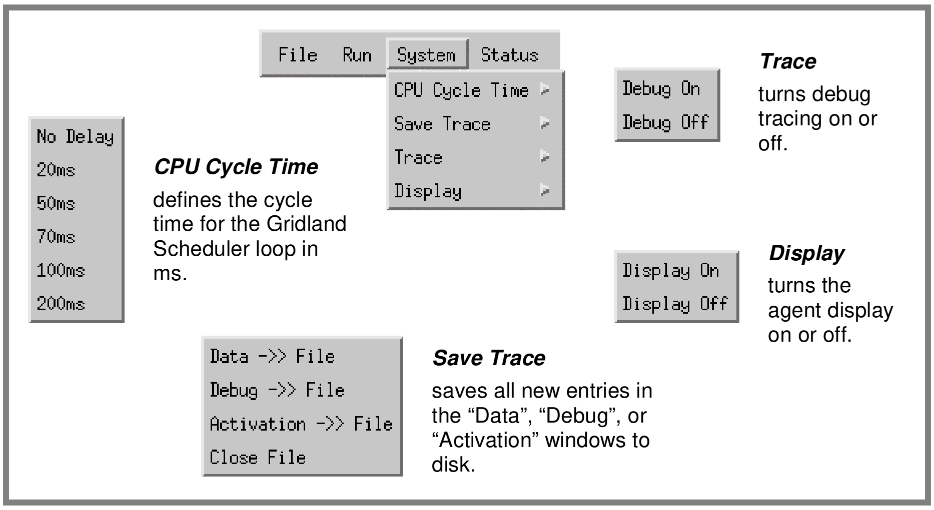


Figure A.4-4 Gridland System Menu

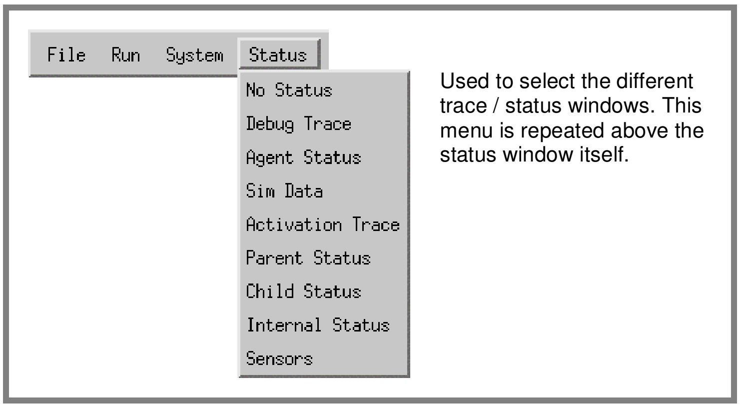


Figure A.4-5 Gridland Status Menu

Trace and Debug Features

The graphical interface also gives us the chance to select and interactively set debug and trace flags for individual agents (Figure A.4-6 and Figure A.4-7).

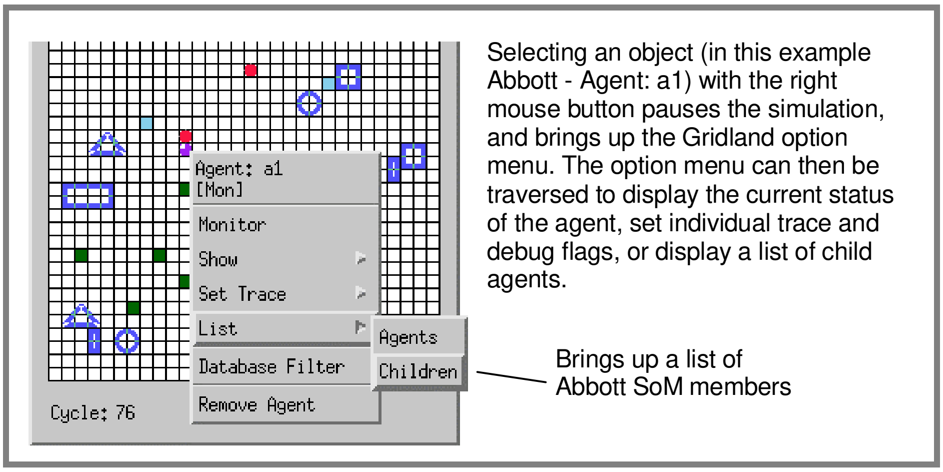


Figure A.4-6 Accessing Abbott's Society of Mind (SoM) members

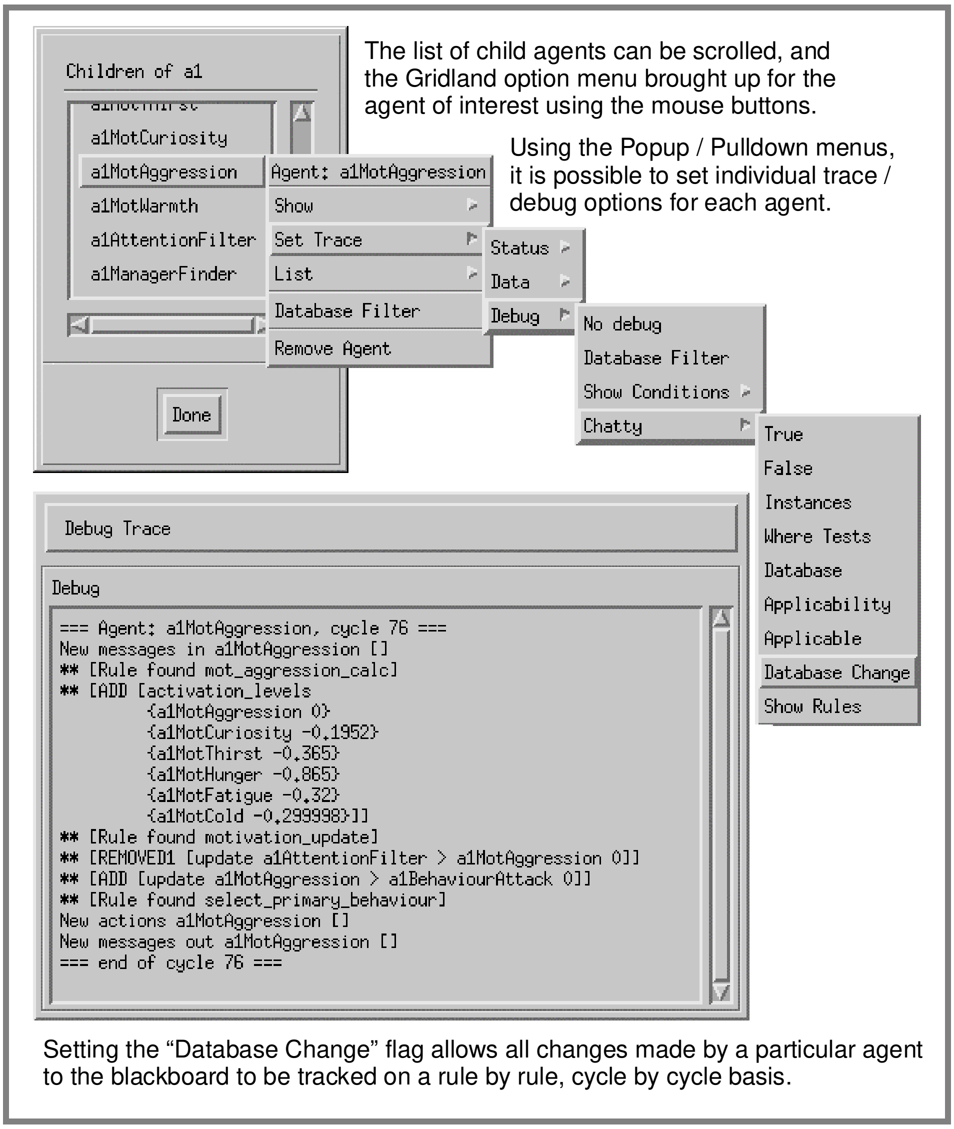


Figure A.4-7 Accessing SIM_AGENT's extensive debug features

A.5 Supporting Different Agent Architectures

Agent architectures can be represented in the SIM_AGENT toolkit in a number of different ways: (a) as a single agent – with sub-mechanisms implemented as separate rule families; (b) as a distributed agent network communicating through message passing; and (c) as a compound object such as the *Society of Mind* (SoM) model. The key to this flexibility lies in the object orientated nature of the toolkit, allowing different methods to be defined to support different agent classes.

Passive objects such as walls, reactive agents such as food and water sources, and simple active agents such as enemies, are all represented in the Gridland world as single agents. On every pass of the scheduler each agent is run for a single time-slice. By defining a separate method for the gl_abbott class agent (Figure A.5-1) we can easily extend the process to ensure that all child agents of the Abbott SoM are also executed whenever the parent Abbott is run.

```
define :method vars sim_run_agent(agent:gl_abbott, agents);
    lvars agent, agents, children, child;

    if sim_status(agent) == "dead" then return endif;

    /* run rulesysyem for parent */
    call_next_method(agent, agents);

    /* run rulesystem for children */
    dlocal sim_parent = agent;
    for children in gl_children_slots do
        for child in children(agent) do
            child -> sim_myself;
            "xBB_" <> sim_name(child) -> local_BB;       /* local blackboard */
            sim_run_agent(child, []);
            sim_do_actions(child, [], sim_cycle_number);
            []->(sim_data(child))("new_sense_data");
        endfor;
    endfor;
enddefine;
```

Figure A.5-1 SoM Compound Agent

B Abbott Experiments

The Gridland Scenario requires Pop11, the SIM_AGENT toolkit, and OpenMotif to run. The experiments themselves (section 6.2; section 8.1.3; and Figure A.5-1) were performed on a 433MHz Celeron PC running a SUSE version of LINUX. The source code and instructions for running the Abbott experiments can be found at:

http://www.cs.bham.ac.uk/research/poplog/abbott

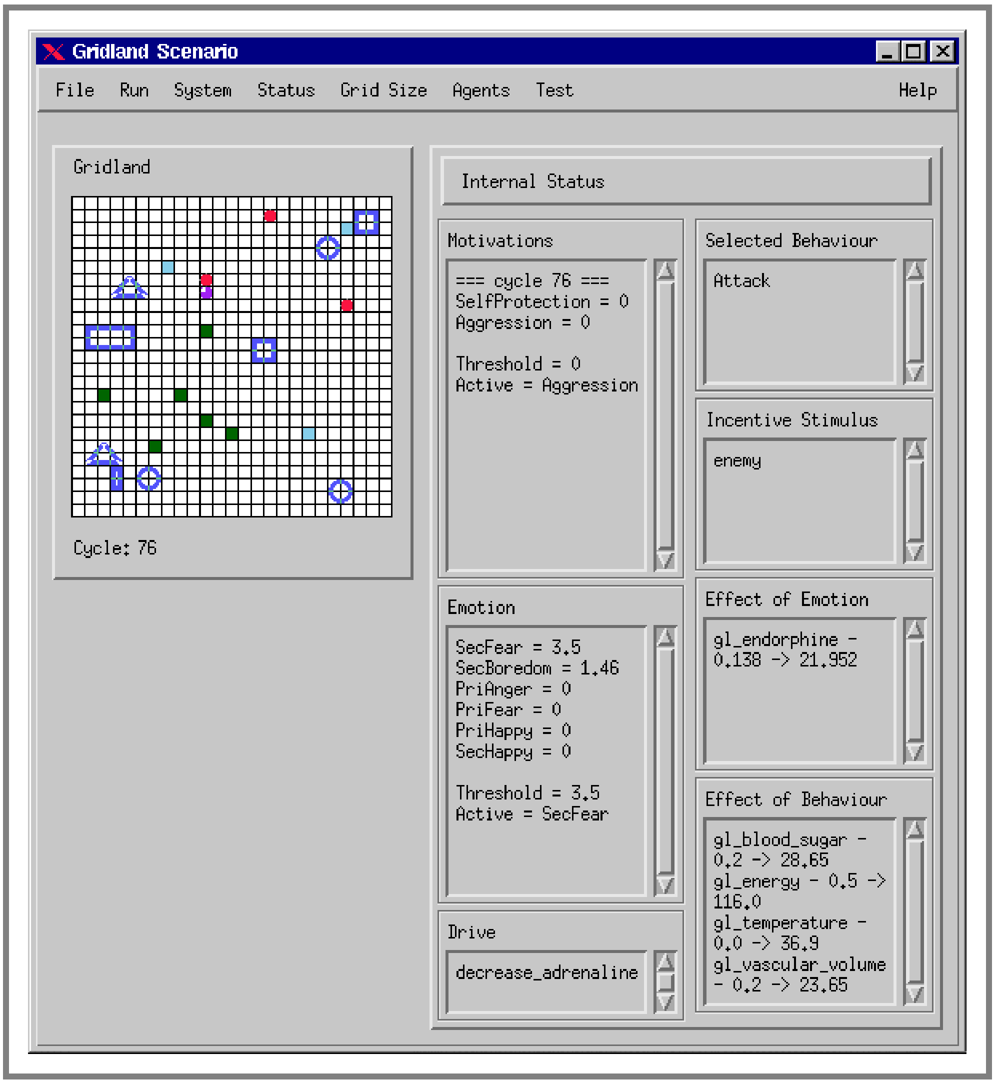


Figure A.5-1 Graphical Shell for the Gridland Scenario

C A Quick Tour of Brain Anatomy

For our ongoing discussion of the neurological basis of emotions, it is useful to have a general overview of the prominent regions of the brain involved in reasoning and emotion. Figure A.5-1a shows the low and high roads to the amygdala of LeDoux's fear system (see section 5.2.4).

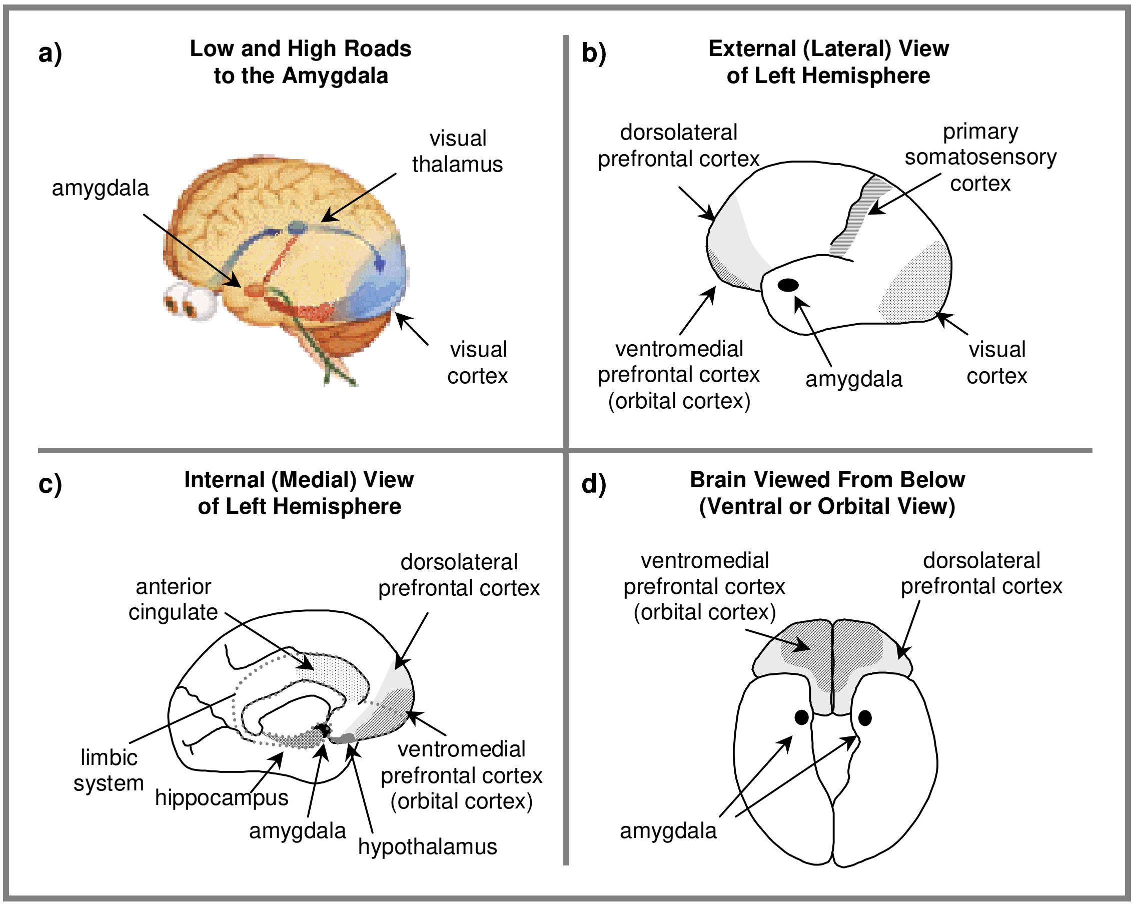


Figure A.5-1 Brain Regions involved in Emotion and Reasoning [Damasio 96; LeDoux 96]

Figure A.5-1b shows the lateral (or external) view of the left brain hemisphere – with the ventromedial prefrontal cortex, the dorsolateral prefrontal cortex, and the primary somatosensory cortex highlighted. Damage to these structures has a number of serious effects on cognition and emotion:

1) Damage to the ventromedial prefrontal cortices results in both compromised reasoning/decision making and emotion/feeling (especially in the personal and social domains) – this is the region Damasio [94] identifies as critical for *somatic marker* expression.

2) Damage to the dorsolateral region compromises decision making but "Either the defect is far more sweeping, compromising intellectual operations over all domains, or the defect is more selective, compromising operations on words, numbers, objects or space, more so than operations in the personal and social domain." [Damasio 96, page 70]

3) Damage to the right primary somatosensory cortex (the somatosensory cortex is right hemisphere dominant – "left hemisphere representations are probably partial and not integrated" [Damasio 96, page 67]) disrupts the process of basic body signalling. Significantly, such damage also compromises reasoning/decision making and emotion/feeling, thus providing further evidence for the influence of *somatic markers* in decision making.

Figure A.5-1c shows the medial (or internal) view of the left brain hemisphere, and the region normally associated with the limbic system. The term "limbic system" does not actually refer to an anatomically unique region of the brain, but it is generally defined to include the hippocampus, the cingulate, the amygdala, and parts of the ventromedial prefrontal/orbital cortex. These areas have specific tasks in reasoning and emotion/feeling:

1) The hippocampus creates a representation of the context that contains not individual stimuli but relations between stimuli. In bringing together all the actors, the hippocampal system is able to generate explicit memories about emotional situations.

2) The anterior cingulate (along with the lateral prefrontal cortex to which it is interconnected) forms part of the frontal lobe attention network, "a cognitive system involved in selective attention, mental resource allocation, decision making processes, and voluntary movement control." [LeDoux 96, page 277] Damasio describes this region as the "fountain head" region where "systems concerned with emotion/feeling, attention, and working memory interact so intimately that they constitute the source for the energy of both external action (movement) and internal action (thought animation, reasoning)." [Damasio 96, page 71]

3) The amygdala has been identified as being central to the fear emotion system, but its role in other emotion systems is still unclear. The amygdala does not appear to play a significant role in positive emotions. "All around the two amygdala, the brain of patient S was perfectly normal. But the amount of calcium deposition was such within the amygdala that it was immediately apparent that little or no normal function of the neurons within the amygdala could still take place." [Damasio 99, page 62] Although her sensory perceptions, her movements, her language, and her basic intelligence were normal, Damasio goes on to describe how "her social behaviour demonstrated a consistent skewing of her prevailing emotional tone. It was as if negative emotions such as fear and anger had been removed from her affective vocabulary, allowing the positive emotions to dominate her life." [Damasio 99, page 65]

D Definition of Hormones

There is some confusion in the computer science literature as to the definition and use of hormones. Cañamero [97] uses the term hormone to refer to the chemical messengers (dopamine, endorphine, and adrenaline) released by emotion and behaviour agents to modify perception and the activation level of the current motivation – see section 6.1.4; whereas Kitano [95] uses the term hormone to include serotonin, histamine, nor-epinephrine (also known as nor-adrenaline), epinephrine (adrenaline), and dopamine. For the purposes of this thesis we will adopt the following definition of hormones: Chemicals secreted by the endocrine glands (pituitary, thyroid, pancreas, adrenal, ovaries, testes) and transported via the bloodstream to certain target organs where they cause specific effects which are vital in regulating and co-ordinating body activities.

Strictly speaking, neurochemical messengers fall into three functional families, which merge into each other [Oatley and Jenkins 96, pages 152-153]:

1) *Dopamine* belongs to the family of *neurotransmitters which are released by nerve impulses at the end of a nerve cell's axon and rapidly diffuse across the tiny synaptic gaps between cells to excite or inhibit the receiving nerve cell or muscle fibre*. There are three classes of neurotransmitters:

    a) *amino acid transmitters* – glutamate and aspartate (excitatory), gamma-aminobutyric acid (GABA) and glycine (inhibitory).

    b) *aminergic transmitters* – acetylcholine, epinephrine, nor-epinephrine, dopamine, serotonin, and histamine.

    c) *peptides* – which, depending on their mode of operation, are sometimes classified as neuromodulators [Bloom 95, page 1064], or even hormones.

    In humans the mesolimbic dopamine system is implicated in incentive motivational processes through which evaluative processing is translated into action, and the stratial dopamine system is implicated in response preparation such as motor readiness [Robbins and Everitt 95, pages 712-713]. Kitano [95] associates dopamine, nor-epinephrine, and epinephrine with pleasure, anger, and fear respectively.

2) *Adrenaline* belongs to the family of *hormones which are carried around the body by the blood to affect organs that are sensitive to them*. Adrenaline molecules are too big to pass though the blood-brain barrier, but are thought to modulate implicit emotional memories and explicit memories of emotion through action on the amygdala and hippocampus [LeDoux 96, page 208].

3) *Endorphine* belongs to the family of *neuromodulators which diffuse some distance to affect many thousands of other nerve cells* – implicated in modulating the pain system.

E Survival Machines

> *"We are all survival machines for the same kind of replicator – molecules called DNA – but there are many different ways of making a living in the world, and the replicators have built a vast range of machines to exploit them. A monkey is a machine that preserves genes up trees, a fish is a machine that preserves genes in the water; there is even a small worm that preserves genes in German beer mats. DNA works in mysterious ways."*
>
> – Dawkins, *The Selfish Gene* (page 21)

E.1 The Selfish Gene

The simple tautology that 'Replicators Replicate' contains the gem of a very powerful idea. A replicator needs no reason, or strives towards no goal, when it replicates. Replication is simply what it does. By the same token, our genes need no grand design, no intentionality, they simply work together in order to get replicated. From a biological perspective, we are the creation of our "selfish" genes, replicators whose only interest lies in replication.

Strictly speaking a gene is not a replicator, it is a label given to *any portion of chromosomal material that potentially lasts for enough generations to serve as a unit of natural selection*. As each useful sequence of DNA is labelled as a gene, we can make the simplifying abstraction of referring to DNA as a community of genes – the junk DNA still plays a useful role by providing raw material for fortuitous mutations. This community of genes needs a vehicle, and so they build survival machines.

The survival machines our genes build are not purposefully designed, they evolve through Darwinian natural selection. Genes that contribute to successful survival machines increase in the population, and genes that find themselves in unsuccessful survival machines die out. The individual genes are oblivious to this fact, they simply code for a particular phenotypic effect.

> "The technical word phenotype is used for the bodily manifestation of a gene, the effect that a gene, in comparison with its alleles, has on the body, via development. The phenotypic effect of some particular gene might be, say green eye colour. In practice most genes have more than one phenotypic effect, say green eye colour and curly hair. Natural selection favours some genes rather than others not because of the nature of the genes themselves, but because of their consequences – their phenotypic effects." – [Dawkins 89, page 235]

Survival machines are not static, they are subject to continuous development throughout their lifetime. Most survival machines start as a single cell which, when it divides, makes identical copies of its genes. Although each cell contains copies of the same genes, it runs a slightly different version of the genetic program. Cells specialise to form the distinctive shapes of limbs and major organs, and it is this development process that is the secret of the gene's success. Changing the timing of an event (possibly through a single genetic mutation) can have dramatic consequences for the final phenotypic effect. Genes code for phenotypic effect via development.

The community of genes does not exist in isolation, but must interact with the environment. This interaction forms the manifest extended phenotype that is actually selected for by natural selection. The phenotypic effect of having 'sharp claws' in itself has no survival value, but the enhanced ability to defend oneself does.

> "Once the genes have provided their survival machines with brains capable of rapid imitation, the memes will automatically take over. We do not even have to posit a genetic advantage in imitation, though that would certainly help. All that is necessary is that the brain should be capable of imitation: memes will then evolve that exploit the capacity to the full." – [Dawkins 89, page 200]

Evolution teaches us to look at nature from the point of view of our genes. We, and every other living creature, are the survival machines of selfish genes. But the story does not end there. The survival machines created by our genes have developed very sophisticated control systems capable of learning and adapting to the environment. Our brain may be the creation of our genes, but our human mind is also the creation of another kind of replicator – that of the meme.

E.2 The Selfish Meme

> "The invasion of human brains by culture, in the form of memes, has created human minds, which alone among animal minds can conceive of things distant and future, and formulate alternative goals. The prospects for elaborating a rigorous science of memetics are doubtful, but the concept provides a valuable perspective from which to investigate the complex relationship between cultural and genetic heritage. In particular, it is the shaping of our minds by memes that gives us the autonomy to transcend our selfish genes." – [Dennett 95, page 369]

A meme is defined as "a unit of cultural transmission, or a unit of imitation." [Dawkins 89, page 192]. A human mind today is vastly different to a human mind of only a few centuries ago, our genes have not had time to change, but our culture has. From a cultural perspective, our minds are shaped by our "selfish" memes, replicators whose only interest lies in being replicated.

> "Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation." – [Dawkins 89, page 192]

Memes are living structures comprising of networks of images, sounds, feelings, tastes, and other memes. A memetic network can have an almost infinite number of variations, and still be recognisable (we probably all have very different images of the meme 'rock and roll' – depending on whether we were teenagers in the 1960's, liked the music, have seen television clips of that era, etc). The meme network for a wheel defines its 'wheelness', a function (in the mathematical sense) of its purpose, its shape, its size, its weight, etc. The meme 'wheel' has many different 'phenotypic' representations – bicycle wheels, pram wheels, cart wheels – and many more personal interpretations, but its essential 'wheelness' can still be captured by the single meme 'wheel'.

> "[…] memes should be regarded as living structures, not just metaphorically but technically. When you plant a fertile meme in my mind you practically parasitize my brain, turning it into a vehicle for the meme's propagation in just the way that a virus may parasitize the genetic mechanism of a host cell. And this isn't just a way of talking – the meme for, say, "belief in life after death" is actually realized physically, millions of times over, as a structure in the nervous systems of individual[s] […] the world over." – N. K. Humphrey, in [Dawkins 89, page 192]

Memes are not restricted to human minds, chimpanzees use sticks to 'fish' [Goodall 91; Dennett 95] for termites. A chimpanzee must select a suitable stick and feed it through the entrance of the termite mound. The stick is then 'attacked' by the soldier termites and must be carefully withdrawn without dislodging the attackers, allowing the termites to be picked off the stick and eaten. Gauging when and how to withdraw the stick requires a skill that has to be learnt.

The 'termite fishing' meme can form part of the culture of one troop, but be absent from the culture of a neighbouring troop, despite both troops having overlapping home ranges. The reason for this is that the 'termite fishing' meme has to be passed on from mother to infant through imitation, and without language a meme lives a precarious life. For storage, a meme needs the mind of living survival machines, and for transmission it must rely on imitation. Chimpanzee troops are small, and so it is no wonder that the 'termite fishing' meme can be lost when a troop fractures. The value of age comes indirectly through the benefits of culture and the need to store memes.

The memes that define human culture are very different to the memes that define the culture of a chimpanzee. We have the ability to express and store our memes symbolically. Our language and symbolic representations allow our memes to be transmitted over large distances and through time. Our memes are no longer dependent on the single mind of a living survival machine. The network of human minds effectively forms a highly connected 'meme-space' (not unlike the 'cyberspace' of the Internet) which gives memes a far greater chance of combining with complementary memes and evolving into ever more powerful meme complexes. As our memes evolve, so too do our minds.

In the same way that a word-processor can be thought of as a virtual machine running on the hardware of a computer, our minds are virtual machines running on the hardware of the brain. A mind has a physical representation in the brain, just as a computer program is represented by charge trapped in a silicon circuit. A mind may be limited by the architecture of the brain, but it is also free to evolve independently of that architecture through the acquisition of culture in the form of memes.

E.3 Evolution of Mind

A mind is the virtual machine that runs on the hardware of the brain. Dennett [95, chapter 13] has proposed a scheme for partitioning the evolution of brains (and minds), based on the idea of a *tower of generate-and-test*.

> "I want to propose a framework in which we can place the various design options for brains, to see where their power comes from. It is an outrageously oversimplified structure, but idealization is the price one should often be willing to pay for synoptic insight. I call it the Tower of Generate-and Test; as each new floor of the Tower gets constructed, it empowers the organisms at that level to find better and better moves, and find them more efficiently." – [Dennett 95, page 373]

The tower is built with the aid of cranes standing on the achievements of the floor below. Cranes are devices for lifting an organism on to a greater level of complexity. Sex, reinforcement learning, foresight, and language are all cranes which enable organisms to climb the *tower of generate-and-test*. Dennett proposes five key milestones in the development of a brain – each milestone is represented by a floor of the *tower*.

Darwinian Creatures

Darwinian creatures are hard-wired organisms, incapable of learning. They are blindly selected for by the environment and represent the ground floor of the *tower*. Our base level Abbott3a architecture (Figure 7.1-2; section 7.1.1) has reached the first stage of Darwinian development.

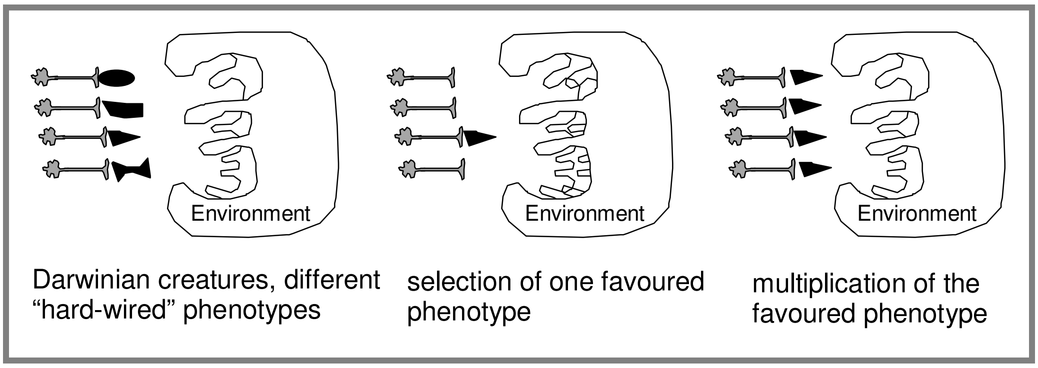


Figure E.3-1 Darwinian Creatures [Dennett 95, page 374]

Skinnerian Creatures

At some point in evolution, Darwinian creatures developed the property of phenotypic plasticity. These organisms are capable of reinforcement learning, and named after the behaviourist B. F. Skinner. Abbott3b (Figure 7.1-3; section 7.1.1), simple classifier systems, and neural networks have reached the stage of Skinnerian development.

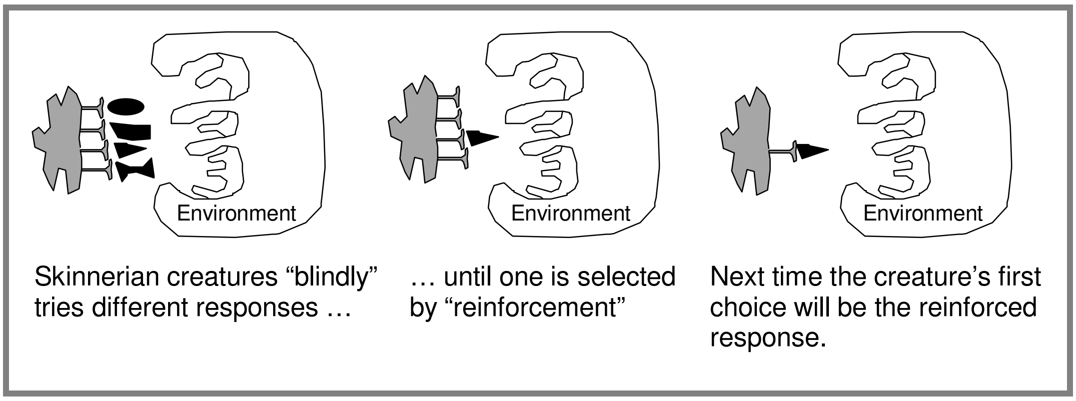


Figure E.3-2 Skinnerian Creatures [Dennett 95, page 375]

Popperian Creatures

Popperian creatures have made the next major advance by building an inner selective environment to preview candidate acts before they meet with the external environment. This inner environment can simply be the result of natural selection, or more complex deliberative reasoning – Abbott3c (Figure 7.1-4; section 7.1.1) has managed to join the ranks of most birds and mammals in reaching the stage of Popperian development.

> "We may call the beneficiaries of this third story in the Tower Popperian creatures, since, as Sir Karl Popper once elegantly put it, this design enhancement 'permits our hypotheses to die in our stead.'" – [Dennett 95, page 375]

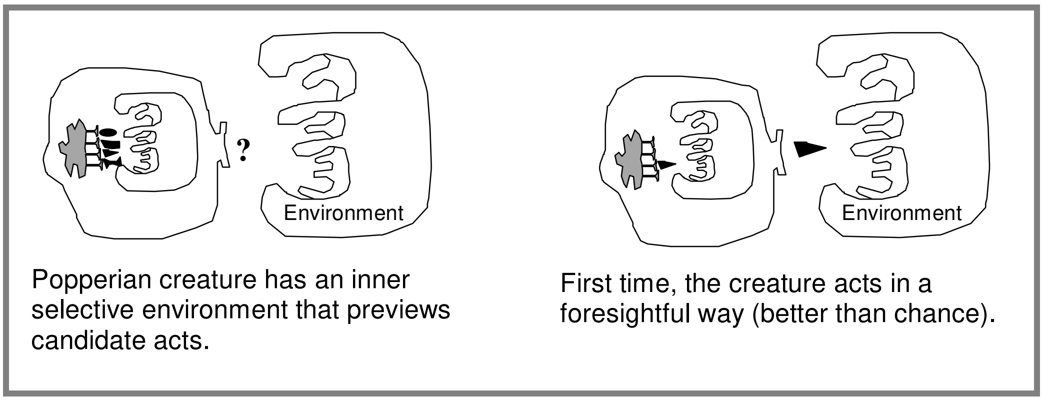


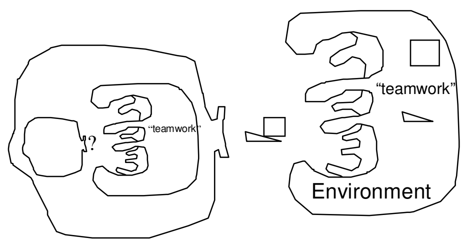
Figure E.3-3 Popperian Creatures [Dennett 95, page 375]

Only certain aspects of the environment can be modelled internally, Popperian creatures still need to take hard knocks to learn from the real environment.

Gregorian Creatures

Gregorian creatures have taken another major leap – the acquisition of mind-tools or memes. Memes not only allow the organism to adapt to the environment, they also allow it to build better internal models of that environment (from the hard knocks taken by others). We would like to think that our Abbott3d (Figure 7.1-5; section 7.1.1) architecture has almost reached the stage of Gregorian development (there is still more work needed to provide the right types of hooks to support the acquisition of memes – a task for future research).

> "The successors to mere Popperian creatures are those whose inner environments are informed by the designed portions of the outer environment. We may call this sub-sub-subset of Darwinian creatures Gregorian creatures, since the British psychologist Richard Gregory is to my mind the pre-eminent theorist of the role of information […] in the creation of Smart Moves." – [Dennett 95, page 377]



Gregorian creature imports mind-tools from the (cultural) environment; these improve both the generators and the testers.

Figure E.3-4 Gregorian Creatures [Dennett 95, page 378]

Scientific Creatures

The top floor of the tower is inhabited by Scientific creatures. Science is perhaps the ultimate form of generate-and-test. Scientific creatures reach the top floor of the Tower by developing complex memes to improve both the generators and testers of their internal model of the environment.

Science advances through the sharing of memes, and needs a rich medium through which scientific memes can be expressed and transmitted. Language provides that medium. Without language we would find ourselves continually reinventing the wheel, or more likely, unable to simplify the problem enough to invent anything. To date we are the only creatures to have developed language and, using it, climbed to the top of the Tower of Generate-and-Test.

E.4 Conclusion

> "The study of the deaf shows us that much of what is distinctively human in us – our capacities for language, for thought, for communication, and culture – do not develop automatically in us, are not just biological functions, but are, equally, social and historical in origin; that they are a gift – the most wonderful of gifts – from one generation to another. We see that Culture is as crucial as Nature." – [Sacks 90, page xiii]

A human infant starts out life as a Popperian creature, relying on primary affects to make behavioural selections. With the acquisition of simple memes the infant can move up the tower to join the ranks of Gregorian creatures capable of modifying their responses to match their environment. Finally language allows the infant to develop more complex memes and join the exclusive club of Scientific creatures. All this happens within a single lifetime.

Without language we would be little more than just another Gregorian primate, and without memes no further advanced than Popperian pigeons. Our position as Scientific creatures is the product of genetic and cultural evolution. Memes are as necessary as genes in the development of a human mind, but the story of memes has one further twist.

Dennett [95] proposes that it is highly probable that language and complex memes can only develop in animals that are self-aware – i.e. humans, chimpanzees, gorillas and orang-utans [Denton 93]. However, there are many areas of activity in which a conflict of interests can arise between the needs of our genes and the needs of a meme-infested self-aware mind. These conflicts seem almost inevitable when the high costs of reproduction are considered – reproduction is necessary for the germ-line replication of genes, but of little use for a meme (aside from the impact on the well-being of the host).

In reality our genes still replicate, and our drive for reproduction is strong enough to influence our culture. The reason behind this lies deep within our Popperian past. No matter where our "selfish" memes try and take us, our mind is built around the machinery of primary emotions – genetically predisposed to survival and *reproduction*. Before we even become aware of an event, that event is likely to have been coloured by our primary affect mechanism. A mechanism in the hands of our genes.

As we strive to build autonomous agents capable of living and working alongside us in our complex world, it is worth pausing for a second to ask ourselves what a self-aware artificial mind would really be like, free of the checks and balances imposed by our genes.