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.

Table of Contents

PART III

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.