VISIONS OF MIND

Architectures for Cognition and Affect

DARRYL N. DAVIS

Visions of Mind:
Architectures for Cognition and Affect

Darryl N. Davis
University of Hull, Kingston-upon-Hull, UK

INFOSCI
Information Science Publishing
Hershey • London • Melbourne • Singapore

Acquisition Editor:	Mehdi Khosrow-Pour
Senior Managing Editor:	Jan Travers
Managing Editor:	Amanda Appicello
Development Editor:	Michele Rossi
Copy Editor:	Lori Eby
Typesetter:	Jennifer Wetzel
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Publishing (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.idea-group.com

and in the United Kingdom by
Information Science Publishing (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 3313
Web site: http://www.eurospan.co.uk

Library of Congress Cataloging-in-Publication Data

Visions of mind : architectures for cognition and affect / Daryl N. Davis, editor.
p. cm.
Includes bibliographical references and index.
ISBN 1-59140-482-7 (hard) -- ISBN 1-59140-483-5 (pbk.) -- ISBN 1-59140-484-3 (ebook)
1. Artificial intelligence. 2. Cognitive science. I. Davis, Daryl N., 1955-
Q335.V57 2004
006.3--dc22

2004022148

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The view expressed in this book are those of the authors, but not necessarily of the publisher.

Preface: Architectures for Cognition and Affect

Introduction

In Spring 2000, the annual symposium of *The Society for the Study of Artificial Intelligence and the Simulation of Behaviour* (AISB) was held at the University of Birmingham. One of the principal workshops ran under the title "How to Design a Functioning Mind." This was organized by Professor Aaron Sloman and attracted a set of international speakers. The following section summarizes the aim, objective, and purpose of that symposium.

This text has similar objectives. It draws on those authors present at the workshop plus others to provide updates on their perspectives for this collected text. The main objective of this collection, nearly 50 years from the (in)famous Dartmouth Conference (McCarthy et al., 1956) where the term "Artificial Intelligence" was first used, is to present an overview of where the research area of artificial minds is at the start of the 21st century. As such, it draws on current and prospective works from pivotal researchers in the area. It includes perspectives from philosophy, psychology, social studies, cognitive science, and artificial intelligence. Rather than present a unified model of mind, it proffers a diverse collection that mirrors the academic discipline. The objectives and mission of the text are, in part, a rephrasing of those associated with the original workshop.

This text brings together research from academics in Europe and the United States interested in building bridges between various kinds of partial studies, with the long-term goal of understanding, at least in principle, how to build a complete mind. This topic is important to artificial intelligence and cognitive science, and also to the academic disciplines that they draw on and feed from, for instance, philosophy, computer science, and psychology. The collection of chapters follows in the tradition of *Computers and Thought* (Feigenbaum & Feldman, 1963), *Mind Design* (Haugeland, 1981), *Architectures for Intelligence* (VanLehn, 1988), and *Android Epistemology* (Ford, Glymour, & Hayes, 1989).

Symposium: How to Design a Functioning Mind

The objective of the two-day symposium was to adopt a multidisciplinary approach to the long-term problem of designing a human-like mind, whether for the scientific purpose of understanding human minds or for some engineering purpose. The following summary of the purpose of the symposium is an edited and abridged version of the original call, which can be found on the Internet (Sloman, 2000).

Purpose of the Symposium

The symposium was intended to bring together people interested in building bridges between various kinds of partial studies (in artificial intelligence, biology, computer science, engineering, ethology, philosophy, and psychology), with the long-term goal of understanding, at least in principle, how to build a complete mind.

Researchers in any discipline were invited to address this issue, whether in a speculative fashion or by reporting firm results that directly contribute to the long-term task. Examples of topics proposed included the following:

- Architectures to accommodate multiple aspects of human mental functioning
- Analyses of requirements for such architectures
- Critiques of existing architectures on the bases of their functional limitations or inconsistent empirical evidence
- Discussions of how important aspects of human minds might have evolved
- Analyses of the problems of designing an adult mind versus designing an infant mind that develops into an adult mind
- Comparisons between capabilities of different animals, which provides evidence for architectural differences
- Overviews of major results in neuroscience that have implications for the virtual machine architecture of a mind (for example, evidence from brain-damaged patients indicating what sorts of separable functional modules exist)
- Philosophical posters presenting familiar arguments to prove that the task is impossible were not particularly welcome, whereas philosophical arguments highlighting some of the difficulties to be overcome were.

Setting the Scene

This text is concerned with the study of questions such as the following:

- What is mind?
- What are theories of mind?
- What are computationally plausible theories of mind?
- What are computationally plausible designs based on these theories of mind?
- What are computationally plausible architectures and systems to support theories of mind?
- What are computationally implemented architectures and systems based on theories of mind?
- What kinds of tools can we use in producing and implementing such designs?
- How do we know when we are successful in producing artificial minds?
- Why do we want to produce artificial or synthetic minds?

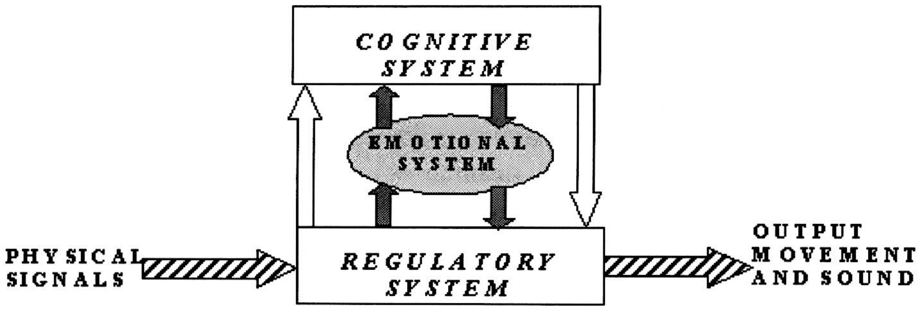
Such questions have been addressed throughout the history of western civilization, from the philosophers of ancient Greece (see Popkin & Stroll, 1993, and Bechtel, 1988, for an overview), through the Age of Enlightenment, for example, Descartes (Bechtel, 1988), the industrial revolution, and Babbage's designs for the Difference Engine (Swage, 2001), to the advent of the computer (von Neumann, 1963; Turing, 1950) and, since its inception as an academic field in its own right, in artificial intelligence (McCarthy et al., 1956).

Just over 20 years ago, Donald Norman (1980) set an agenda of important issues for cognitive science: belief systems, consciousness, development, emotion, interaction, language, learning, memory, perception, performance, skill, and thought. This nonexhaustive list of topics has played an important role in determining where researchers have focused their work.

Such study is arguably the core of the field of artificial intelligence (Barr & Feigenbaum, 1981; Franklin, 1995; Minsky, 1987; Nilsson, 1998; Russell & Norvig, 2003; Sharples et al., 1989; Winston, 1992) and cognitive science (Bechtel, 1988; Bechtel & Abrahamsen, 2002; Simon, 1979; Wilson & Keil, 1999; Winograd & Flores, 1986).

An important aspect of the agenda set by Norman is the use of an architectural perspective. One such example architecture is given in Figure 1. Five interacting processes are identified: the reception of incoming signals (perception), the

Figure 1. One architecture for an artificial mind (based on Norman, 1980)



generation of output (action selection), a reactive or regulatory system, a deliberative or cognitive system, and an emotional or affect system. Norman suggested that this is the type of architecture needed to address the 12 topics included in his agenda for cognitive science. The 15 chapters in this text address these and many related issues.

Chapter Overview

In Chapter 1, Andrew Adamatzky portrays the mind as an imaginary chemical reactor, where discrete entities of emotions and beliefs diffuse and react as molecules. He presents two models of mind using the computational chemistry metaphor: doxastic solution where quasi-chemical species represent knowledge, ignorance, delusion, doubt, and misbelief; and affective solution, where reaction mixtures include happiness, anger, confusion, and fear. Using numerical and cellular-automaton techniques, he presents a rich spectrum of nontrivial phenomena in the spatiotemporal dynamic of the affective and doxastic mixtures. This paradigm of nonlinear medium-based mind is to be used in future studies in developing intelligent robotic systems, designs of artificial organic creatures with liquid brains, and the diffusive intelligence of agent collectives.

In Chapter 2, Michel Aubé proposes a model of emotions relying upon an analysis of the requirements that are to be met by individuals of nurturing species, so as to adapt themselves to their social environments. It closely reflects the structures of other motivational systems which consist of control structures dedicated to the management of resources critical for survival. The particular resources emotional systems seem to handle have to do with social bonding and collaborative behaviors. These are referred to as second-order resources. They are made available to other agents, and they are captured in the model through the concept of commitments. Emotions thus appear as computational control systems that handle the variation of commitments lying at the root of interactive and collaborative behaviors. Some critical consequences of the model for the implementation of emotions in artificial systems are drawn at the end of the chapter.

In Chapter 3, John Barnden speculatively addresses the nature and effects of metaphorical views that a mind can intermittently use in thinking about itself and other minds, such as the view of mind as a physical space in which ideas have physical locations. Although such views are subjective, the chapter argues that they are nevertheless part of the real nature of the conscious and unconscious mind. In particular, it is conjectured that if a mind entertains a particular (metaphorical) view at a given time, then this activity could of itself cause that mind to become more similar in the short term to how it is portrayed by the view. Hence, the views are, to an extent, self-fulfilling prophecies. In these ways, metaphorical self-reflection, even when distorting and inaccurate, is speculatively an important aspect of the true nature of mind. The chapter also outlines a theoretical approach and related implemented system (ATT-Meta) that were designed for the understanding of metaphorical discourse but that have principles that could be at the core of metaphorical self-reflection in people or future artificial agents.

In Chapter 4, Joanna Bryson presents an analysis of the modularity of mind. Many architectures of mind assume some form of modularity, but what is meant by the term "module"? This chapter creates a framework for understanding current modularity research in three subdisciplines of cognitive science — psychology, artificial intelligence, and neuroscience. The framework starts from the distinction between horizontal modules that support all expressed behaviors versus vertical modules that support individual domain-specific capacities. The framework is used to discuss innateness, automaticity, compositionality, representations, massive modularity, behavior-based and multiagent artificial intelligence systems, and correspondence to physiological neurosystems. There is also a brief discussion of the relevance of modularity to conscious experience.

In Chapter 5, William Clocksin explores issues in memory and affect in connection with possible architectures for artificial cognition. Because memory and emotion are evolutionarily and developmentally rooted in social interdependence, a new understanding of these issues is necessary for the principled design of true intelligent systems. Emotion is not treated as an optional extra or as a brief episode of feeling but as the underlying substrate enabling the formation of social relationships essential for the construction of cognition. Memory is not treated as the storage and retrieval of immutable data but as a continuous process contingent on experience and never fully fixed or immutable. Three converging areas of research are identified that hold some promise for further research: social constructionism, reconsolidation memory theory, and memory models based on the nonlinear dynamics of unstable periodic orbits. He argues that the combination of these ideas offers a potentially more substantive approach to understanding the cognitive architecture.

In Chapter 6, Bruce Edmonds describes free will in terms of the useful properties that it could confer, explaining why it might have been selected over the course of evolution. These properties are exterior unpredictability, interior rationality, and social accountability. A process is described that might bring it about when deployed in a suitable social context. It is suggested that this process could be of an evolutionary nature, that free will might "evolve" in the brain during development. This mental evolution effectively separates the internal and external contexts, while retaining the coherency between individuals' public accounts of their actions. This is supported by the properties of evolutionary algorithms and possesses the three desired properties. Some objections to the possibility of free will are dealt with by pointing out the *prima facie* evidence and showing how an assumption that everything must be either deterministic or random can result from an unsupported assumption of universalism.

In Chapter 7, John Fox presents argumentation about the concept of mind. The idea of "mind" did not spring fully formed into human consciousness. On the contrary, it has been articulated slowly through the millennia, drawing upon countless metaphors and images in different cultures at different times. In the last 50 years, the concepts of conventional science and technology provided the primary images that we employ in discussing mental processes, though there are presently many competing perspectives. Each of these images is incomplete when it comes to explaining mental phenomena, and many are inconsistent. In this chapter, a few of the prominent perspectives that influenced cognitive science in the last half century or so, from information processing psychology to artificial intelligence, are reviewed. The conclusion is that a unified theory of mind will need insights from multiple viewpoints. The challenge to the field is to avoid disputes over different positions and look for ways to bring them together.

In Chapter 8, Stan Franklin describes an architecture for an autonomous software agent designed to model a broad spectrum of human cognitive and affective functioning. In addition to featuring "consciousness," the architecture accommodates perception, several forms of memory, emotions, action-selection, deliberation, ersatz language generation, several forms of learning, and metacognition. One such software agent, IDA, embodying much of this architecture, is up and running. IDA's "consciousness" module is based on global workspace theory, allowing it to select relevant resources with which to deal flexibly with exogenous and endogenous stimuli. Within this architecture, emotions implement IDA's drives, her primary motivations. Offering one possible architecture for a fully functioning artificial mind, IDA constitutes an early attempt at the exploration of design space and niche space. The design of the IDA architecture spawns hypotheses concerning human cognition and affect that can serve to guide the research of cognitive scientists and neuroscientists.

In Chapter 9, David Glasspool discusses how clues to the way behaviour is integrated and controlled in the human mind emerged from cognitive psychology and neuroscience. The emerging picture mirrors solutions (driven primarily by engineering concerns) to similar problems in the rather different domains of mobile robotics and intelligent agents in artificial intelligence. Details are presented about the relationship between a psychological theory of willed and automatic control of behaviour, the Norman and Shallice framework, and three types of engineering-based theory in artificial intelligence. As well as being a promising basis for a large-scale model of cognition, the Norman and Shallice framework presents an interesting example of apparent theoretical convergence between artificial intelligence and empirical psychology, and of the ways in which theoretical work in both fields can benefit from interaction between them.

In Chapter 10, Fernand Gobet and Peter C.R. Lane provide an introduction to the CHREST architecture of cognition and show how this architecture can help develop a full theory of mind. After describing the main components and mechanisms of the architecture, they discuss several domains in which it was already successfully applied, such as in the psychology of expert behaviour, the acquisition of language by children, and the learning of multiple representations in physics. The highlighted characteristics of CHREST that enable it to account for empirical data include self-organisation, an emphasis on cognitive limitations, the presence of a perception-learning cycle, and the use of naturalistic data as input for learning. They argue that some of these characteristics can help shed light on the hard questions facing theorists developing a full theory of mind, such as intuition, the acquisition and use of concepts, the link between cognition and emotions, and the role of embodiment.

In Chapter 11, Elizabeth Gordon and Brian Logan address the key problem for agents of responding in a timely and appropriate way to multiple, often conflicting goals in a complex, dynamic environment. They propose a novel goal-processing architecture that allows an agent to arbitrate between multiple conflicting goals. Building on the teleo-reactive programming framework originally developed in robotics, they introduce the notion of a *resource*, which represents a condition that must be true for the safe concurrent execution of a durative action. They briefly outline a goal arbitration architecture for teleo-reactive programs with resources, which allows an agent to respond flexibly to multiple competing goals with conflicting resource requirements.

In Chapter 12, Pentti Haikonen considers the following fundamental issues of artificial minds and conscious machines: the representation and symbolic processing of information with meaning and significance in the human sense; the perception process; a neural cognitive architecture; system reactions and emotions; consciousness in the machine; and artificial minds as a content-level phenomenon. Solutions are proposed for related problems, and a cognitive machine is outlined. An artificial mind within this machine that eventually controls the machine, is seen to arise via learning and experience as higher-level content is constructed.

In Chapter 13, Colin Johnson considers how new and emerging computational models relate to the arguments surrounding dualism and embodiment. In recent years, the idea that somatic processes are intimately involved in actions traditionally considered to be purely mental has come to the fore. These arguments have revolved around the concept of *somatic markers*, i.e., bodily states generated by the mind and then reperceived and acted upon. In this chapter, the somatic marker hypothesis and related ideas from the point of view of *postclassical computation*, i.e., the view that computing can be seen as a property of things-in-the-world rather than of an abstract class of mathematical machines, are considered. From this perspective, a number of ideas are discussed: the idea of somatic markers extending into the environment, an analogy with hardware interlocks in complex computer-driven systems, and connections with the idea of "just-do-it" computation.

In Chapter 14, Matthias Scheutz introduces an *architecture framework* called *APOC* (for "*Activating-Processing-Observing-Components*") for the analysis, evaluation, and design of complex agents. APOC provides a unified framework for the specification of agent architectures at different levels of abstraction. As such, it permits intermediary levels of architectural specification between high-level functional descriptions and low-level mechanistic descriptions that can be used to systematically connect these two levels.

In Chapter 15, Push Singh and Marvin Minsky consider how to build systems with "common sense," the thinking skills that every ordinary person takes for granted (?). To build systems as resourceful and adaptive as people, we must develop cognitive architectures that support great procedural and representational diversity. No single technique is powerful enough to deal with the broad range of domains every ordinary person can understand — even as children, we can effortlessly think about complex problems involving temporal, spatial, physical, bodily, psychological, and social dimensions. In their chapter, they describe a multiagent cognitive architecture that aims for such flexibility. Rather than seek a best way to organize agents, our architecture supports multiple "ways to think," each a different architectural configuration of agents. Each agent may use a different way to represent and reason with knowledge, and there are special "panalogy" mechanisms that link agents that represent similar ideas in different ways. At the highest level, the architecture is arranged as a matrix of agents: Vertically the architecture divides into a tower of reflection, including the reactive, deliberative, reflective, self-reflective, and self-conscious levels; horizontally the architecture divides along "mental realms," including the temporal, spatial, physical, bodily, social, and psychological realms.

References

AISB. (n.d.). *The society for the study of artificial intelligence and the simulation of behaviour.* Retrieved from the World Wide Web: *http://www.aisb.org.uk/*

Barr, A., & Feigenbaum, E. A. (1981). *The handbook of artificial intelligence* (vol. 1). San Francisco: Morgan Kaufmann.

Bechtel, W. (1988). *Philosophy of mind: An overview of cognitive science.* Mahweh, NJ: Lawrence Erlbaum Associates.

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the mind* (2nd ed.). Oxford: Blackwell Scientific.

Feigenbaum, E. A., & Feldman, J. (Eds.). (1963). *Computers and thought.* New York: McGraw-Hill.

Ford, K. M., Glymour, C., & Hayes, P. J. (Eds.). (1989). *Android epistemology.* Cambridge, MA: MIT Press.

Franklin, S. P. (1995). *Artificial minds.* Cambridge, MA: MIT Press.

Haugeland, J. (ed.). (1981). *Mind design.* Cambridge, MA: MIT Press.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). *A proposal for the Dartmouth summer research project on artificial intelligence.* Retrieved from the World Wide Web: *http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html*

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

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Nilsson, N. J. (1998). *Artificial intelligence: A new synthesis.* San Francisco: Morgan Kaufmann.

Norman, D. A. (1980). Twelve issues for cognitive science. *Cognitive Science, 4*, 1-33.

Popkin, R. H., & Stroll, A. (1993). *Philosophy* (3rd ed.). London: Reed Elsevier.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). New York: Prentice Hall.

Sharples, M., Hogg, D., Hutchinson, C., Torrance, S., & Young, D. (1980). *Computers and thought.* Cambridge, MA: MIT Press.

Simon, H. A. (1979). *Models of thought.* New Haven, CT: Yale University Press.

Sloman, A. (2000). *AISB 2000: The cognition and affect project.* University of Birmingham, UK. Retrieved from the World Wide Web: *http://www.cs.bham.ac.uk/research/cogaff/cogaff.html*

Swage, D. (2001). *The cogwheel brain: Charles Babbage and the quest to build the first computer.* Kent, UK: Abacus Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433-460.

VanLehn, K. (Ed.). (1988). *Architectures for intelligence.* Mahweh, NJ: Lawrence Erlbaum Associates.

von Newmann, J. (1963, originally published in 1937). General and logical theory of automata. In A. H. Taub (Ed.), *J. von Newmann, Collected works.* London: Pegamon/Elsevier.

Wilson, R. A., & Keil, F. C. (Eds.). (1999). *The MIT Encyclopedia of the cognitive sciences.* Cambridge, MA: MIT Press.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition.* Reading, MA: Addison-Wesley.

Winston, P. H. (1992). *Artificial intelligence* (3rd ed.). Reading, MA: Addison-Wesley.

Acknowledgments

Foremost on the list of acknowledgments, with their endless stream of helpful questions, are the chapter authors who responded so well to the editor's polite tyranny. Thanks also go to Keith Oatley, who on determining he would be unable to contribute directly, suggested contacting Michel Aubé for a psychological perspective.

Thanks are due to the publishing team at Idea Group Inc. In particular to Jan Travers and Michele Rossi, who responded to all my e-mail questions on the nature and format of the book once the project was underway, and to Mehdi Khosrow-Pour, who enthusiastically took the project on board when I was seeking an outlet.

Thanks are due to my department, and in particular, Derek Wills, for granting me the latitude to pursue this project, and those staff members involved in the endless discussions aspects of this text have provoked. Finally, thanks to Aaron Sloman who allowed me to take this little baby off his hands.

Darryl N. Davis, May 2004