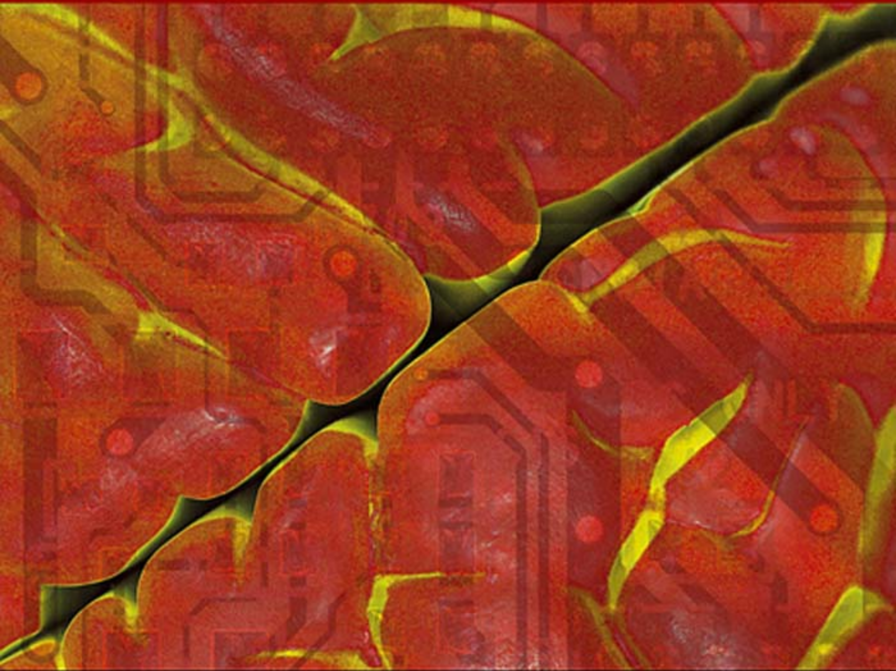
ELSEVIER
ACADEMIC
PRESS

Cognitive Systems

Information Processing Meets Brain Science



RICHARD MORRIS • LIONEL TARASSENKO
MICHAEL KENWARD

COGNITIVE SYSTEMS:

INFORMATION PROCESSING MEETS BRAIN SCIENCE

This page intentionally left blank

Cognitive Systems: Information Processing Meets Brain Science

Scientific Coordinators

RICHARD MORRIS FRS, FRSE
Division of Neuroscience, University of Edinburgh, UK

and

LIONEL TARASSENKO FREng
Department of Engineering Science
University of Oxford, UK

Editorial Consultant

MICHAEL KENWARD OBE
Grange Cottage, Staplefield, Haywards Heath, UK

ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS •SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

This book is printed on acid-free paper

Elsevier Academic Press
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.elsevier.com

Elsevier Academic Press
84 Theobald's Road, London WC1X 8RR, UK
http://www.elsevier.com

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

Library of Congress Catalog Number: 2005923102

ISBN 0-12-088566-2

Typeset by Charon Tec Pvt. Ltd, Chennai, India
www.charontec.com

Printed and bound in Great Britain
05 06 07 08 09 9 8 7 6 5 4 3 2 1

Contents

This page intentionally left blank

Contributors

Bashir Ahmed
Department of Physiology
The University of Oxford, UK

Jim Austin
Advanced Computer Architectures Group
Department of Computer Science
University of York, UK

Philip Barnard
MRC Cognition and Brain Sciences Unit
Cambridge, UK

Sarah-Jayne Blakemore
Institute of Cognitive Neuroscience
London, UK

Tim Bussey
Department of Experimental Psychology
University of Cambridge, UK

Dave Cliff
Semantic & Adaptive Systems Department
Hewlett-Packard Laboratories
Bristol, UK

Anthony G. Cohn
School of Computing
University of Leeds, UK

Mike Denham
Centre for Theoretical and Computational
Neuroscience
University of Plymouth, UK

Jon Driver
Institute for Cognitive Neuroscience
University College London, UK

Maria Fox
Department of Computer Science and
Information Systems
University of Strathclyde
Glasgow, UK

Karl Friston
Functional Imaging Laboratory
Welcome Department of Imaging
Neuroscience
Institute of Neurology
University College London, UK

Uta Frith
Institute of Cognitive Neuroscience
London, UK

Robert Ghanea-Hercock
Future Technologies Group
BTexact Technologies
Ipswich, UK

Kim Graham
Memory and Knowledge Group MRC
Cognition and Brain Sciences Unit
Cambridge, UK

Wendy Hall
Intelligence Agents Multimedia Group
Department of Electronics and
Computer Science
University of Southampton, UK

Graham Hitch
Department of Psychology
University of York, UK

Nicholas R. Jennings
Intelligence Agents Multimedia Group
School of Electronics and Computer Science
University of Southampton, UK

Michael Kenward
Grange Cottage, Staplefield
Haywards Heath, UK

Simon Laughlin
Department of Zoology
University of Cambridge, UK

Derek Long
Department of Computer Science and Information Systems
University of Strathclyde
Glasgow, UK

Michael Luck
Intelligence Agents Multimedia Group
School of Electronics and Computer Science
University of Southampton, UK

William D. Marslen-Wilson
MRC Cognition and Brain Sciences Unit
Cambridge, UK

Andrew Matus
Friedrich-Mischer Institute
Basle, Switzerland

Danius T. Michaelides
Intelligence Agents Multimedia Group
School of Electronics and Computer Science
University of Southampton, UK

Richard Morris
Division of Neuroscience
University of Edinburgh, UK

Steve Munroe
Intelligence Agents Multimedia Group
School of Electronics and Computer Science
University of Southampton, UK

Lynn Nadel
Department of Psychology and Division of Neural Systems and Aging
University of Arizona, USA

Kieron O'Hara
Intelligence Agents Multimedia Group
Department of Electronics and Computer Science
University of Southampton, UK

Peter Redgrave
Department of Psychology
University of Sheffield, UK

Edmund Rolls
Department of Experimental Psychology
University of Oxford, UK

Nigel Shadbolt
Intelligence Agents Multimedia Group
Department of Electronics and Computer Science
University of Southampton, UK

Bill Sharpe
The Appliance Studio Ltd
Bristol, UK

Lionel Tarassenko
Department of Engineering Science
University of Oxford, UK

Sir John Taylor
Director of the Foresight Cognitive Systems Project
UK

Keith van Rijsbergen
Department of Computing Science
University of Glasgow, UK

Vincent Walsh
Institute of Cognitive Neuroscience
University College London, UK

Mark J. Weal
Intelligence Agents Multimedia Group
School of Electronics and Computer Science
University of Southampton, UK

David Willshaw
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh, UK

Andy Wright
BAE Systems Advanced Technology Centre
Filton, Bristol, UK

Preface

Lynn Nadel

Department of Psychology and Division of Neural Systems, Memory and Aging, University of Arizona, Tucson

Cognitive systems include both natural and artificial information processing systems responsible for perception, learning, reasoning, communication, action and more. Although we can trace the emergence of a field of study concerned with such systems back to the nineteenth century, its modern focus on information processing is more readily traced to a series of critical advances in the twentieth century. Among these must be included work by Alan Turing on computability, by Frederick Bartlett on the role of schemas, by Kenneth Craik on the importance of internal 'models', by Warren McCulloch and Walter Pitts on the use of idealized neurons as logical devices, by John von Neumann on game theory and the development of computers, and by Warren Weaver and Claude Shannon on information theory.

Much of this work unfolded in the context of critical wartime needs, and reflected productive interactions between life scientists, physical scientists and social scientists. As exemplified by a series of extraordinary meetings convened by the Josiah Macy Foundation from 1942 through to the early 1950s, these interactions seemed destined to lead to rapid understanding of how natural and artificial processing devices could solve even the most complex problems.

Among the participants in these and related meetings held both in the UK and the United States were brain scientists, mathematicians, anthropologists, physicists and psychiatrists. A heady mix. During these years a number of critical advances in understanding the brain emerged, along with the development of conceptual frameworks that offered the promise of deep insights into how mental functions could be realized in biological, and hence physical, form.

New methods in the study of the nervous system such as the development of electroencephalographic (EEG) recording and intracellular recording in model systems, such as the squid giant axon, were particularly important. The former led to the discovery of the 'reticular activating system' by Giuseppe Moruzzi and Horace Magoun, a key step in understanding the underlying patterns of activity in the brain, and how they are affected by arousal and attention. The latter led to the development of detailed models of axonal transmission, synaptic communication and neuronal function that have shaped the neuroscientific research agenda for decades.

On the conceptual front perhaps the most important contribution was that by Donald Hebb, whose neuropsychological speculations about cell assemblies and phase sequences provided the first serious attempt at explaining, in terms capable of being simulated in artificial circuitry, how brain processing might embody the mind. Indeed, attempts to instantiate Hebb's insights in early computer programs followed almost immediately, through informal communications and meetings between Hebb's group in Montreal, and the group connected to Frank Rosenblatt, the 'inventor' of the Perceptron, at Cornell.

However, as the field of cognitive science crystallized in the mid-1950s it lost much of its fervor for interactions between life scientists and physical scientists. There were many reasons for this. John von Neumann himself pointed out the several ways in which brains were not really like computers, at least not the computers of that era. Advances in understanding information processing, the structure of language, and the representation and manipulation of mental images were all made with little reference to their underlying biological substrates. At the same time, the attempts to model Hebb's neuropsychological concepts ran up against serious obstacles, including the limits of both Hebb's models (the earliest of which had not included neuronal inhibition), and the computational devices available at that time.

The few decades during which cognitive science, neuroscience and physical science interacted minimally were not a complete loss – considerable understanding was generated in isolated subsystems. Experimental psychologists unpacked many of the mysteries of attention, perception, vision, memory and action. Analyses of language, by Noam Chomsky and others, brought great insights into the formal properties of this uniquely human cognitive system. At the same time, we gradually came to understand a great deal more about the wetware of the nervous system. Through developments such as single-neuron recording in the intact mammalian brain (in the 1950s and 1960s), and more recently ensemble recording (in which multiple single-units are recorded simultaneously), we have accumulated considerable data on how neurons work both individually and collectively in the service of adaptive behavior. Given this progress, the time is ripe for renewed attempts to characterize cognitive systems in broader biological and physical terms.

Among recent attempts to reinvigorate such interactions is the Foresight Project on Cognitive Systems, which led to the material published in this book. Starting from the premise that many issues confronting society in the present and future will be greatly influenced by how well both humans and machines process information, the Foresight Project sought to spell out the achievements and limits of current science, to identify the challenges to further progress, and to speculate on the opportunities that would arise should these problems be solved. At the heart of this enterprise is the notion that re-connecting physical scientists and life scientists will be critical to generating solutions. The Foresight Project itself was a test case – could communities of scientists whose recent history has kept them apart come together in ways that might promise real advance?

On the evidence of the discussions held in workshops during the project and the contributions presented in this volume, there is reason for optimism. Similarly, the contributions to a recent *Encyclopedia of Cognitive Science* (Nadel, 2003) illustrate the extent of overlap of research interests right across this diverse community of scientists. In a variety of domains this and related projects have connected scientists across the gulf between biological and physical science. In each case, the connection has generated new ideas, new approaches and even new experiments.

Will all this pan out? It is obviously too early to tell. But the excitement is palpable, and readers of this book have the opportunity to see where the future might lie in the sciences of the mind and brain.

Reference

Nadel, L. (ed.) (2003) *The Encyclopedia of Cognitive Science*. London: Macmillan.

Introduction
Brain Science and Information Technology – Do They Add Up?

Sir John Taylor OBE, FRS, FREng
Director of the Foresight Cognitive Systems Project, UK

False dawns are not new to scholars of the relationship between brain science and information technology. The two have tried to exchange ideas before. However, such was the disappointment surrounding lack of progress in artificial intelligence (AI), for example, that the term has all but disappeared from the scientific lexicon, along with funding for research in the area.

There were good reasons behind the earlier failure to deliver the promised 'breakthroughs' in computing that would, the proponents of AI claimed, build on our knowledge of how living systems achieve their marvels. To begin with, our understanding of living cognitive systems was too primitive to offer enough leads for IT researchers. At the time, in the 1980s, computer power was paltry in comparison with today's systems.

Two decades on, we have been through 10 generations of 'Moore's Law', the self-fulfilling premise that computing power, usually measured as the density of the transistors on an integrated circuit, will double every 18 months or so. More important, the cost of computer power has plummeted, bringing it to many more scientific domains. Research in natural and artificial cognitive systems has also made much progress in the past 20 years. There have, however, been obstacles to the exchange of ideas between the two areas of scientific endeavour. Not least has been the ever-greater specialization that is almost an inevitable consequence of deeper knowledge. Even within the life sciences, for example, individual researchers find it ever harder to maintain an intimate knowledge of what is happening in areas of science close to their own domain, let alone in the seemingly alien world of IT.

It is, then, understandable that throughout the 1990s, while some scientists tried to straddle the disciplines, there was little systematic dialogue between researchers in IT and brain scientists. The Foresight Cognitive Systems Project was perhaps the first sustained initiative to bring the two fields together to see if they have anything productive to say to one another. Cognitive systems – natural and artificial – sense, act, think, feel, communicate, learn and evolve. We see these capabilities in many forms in living organisms. The natural world shows us how systems as different as a colony of ants or a human brain achieve sophisticated adaptive behaviours. Among the more significant outcomes of the Foresight Cognitive Systems Project was the conclusion that research into natural and artificial cognitive systems is indeed at an exciting stage. The researchers who took part in the project also agree that there could be great benefits in bringing together the life sciences and physical sciences to consider how they can collectively accelerate progress in cognitive systems.

When the project began, there was only a limited exchange between the research communities looking at cognitive systems from their different perspectives. The premise for the project was to consider if it was timely for these two communities to work together. Would there be any common ground? In the event, the timing may have been perfect. Perhaps encouraged by the Foresight Project, made possible by the Office of Science and Technology, a part of the UK's Department of Trade and Industry, a number of agencies around the world have started to investigate the territory afresh. To pick just two examples: in Europe, the European Commission has identified cognitive systems as one of the priorities for the new Sixth Framework Programme; in the USA, the Defense Advanced Research Projects Agency (DARPA) has launched an initiative in cognitive systems, to 'develop the next generation of computational systems with radically new capabilities, "systems" that know what they're doing'.

The Right Time

There were two underpinning reasons for thinking that it would be timely to review the conjunction of brain science and IT. First, there was the recognition of the importance of cognition and the fact that new tools are helping us to study living systems. Secondly, we could see that in some areas artificial cognitive systems are hitting a wall using a strictly engineering approach to problem solving.

On the first of these reasons, techniques from the physical sciences are making a growing contribution to the life sciences. Hardly a day goes by without yet more newspaper coverage of the contribution of functional magnetic resonance imaging (fMRI) to the study of brain activity. A chapter in this collection describes some of the excitement that comes from our growing use of new techniques in brain science. This is but one of a series of research reviews that the OST commissioned from some of the leading researchers in their subjects. Our brief to them was to describe what is happening in their fields of science in terms that communicate their excitement to people working in different disciplines. So we asked life scientists to write for physical scientists, and vice versa.

Challenges for the Future

While this is in itself a significant goal, perhaps more important was our request that the authors report on challenges where it is not quite clear where the research is going. As we put it to them, what are the open questions in your area of science? The research reviews in this volume are, therefore, more than an account of the state of play in two of the most active areas of modern research. They are, in effect, manifestos for the future of research in brain science and IT.

The research reviews grew out of a series of meetings at the beginning of the 18-month project. The first task of these meetings was to see if there really was enough common ground to warrant a continued engagement. Having concluded that this was the case, the scientists then identified the areas where they thought exchange was most likely to be fruitful. It was at this stage that the project commissioned this series of research reviews, each of which would describe an area of particular interest. The idea was that the reviews would explain the excitement of important areas of research in cognitive systems in a language that would be accessible to experts in other disciplines. This need for communication became clear during the first meetings of the researchers who took part in the project. Their first hurdle was to get to grips with the specialist language of the different disciplines, especially the ways in which they use the same terms in different ways.

Choosing the subjects that we should cover in these Research Reviews provided plenty of opportunities for lively discussion. But the remarkable thing we found was that it was possible to agree on the key areas of research in our diverse disciplines. Much more important, as several researchers pointed out, was the remarkable extent to which the subjects mapped on to one another across what now looks like an increasingly artificial divide between the life sciences and physical sciences. Thus several research reviews consider subjects from the different perspectives of brain science and IT. For example, there are two papers on aspects of speech and language, one on human speech the other on automatic speech recognition.

The Broader Remit

While one reason for collecting the reviews is to alert more scientists to the new possibilities at this particular interface between the life sciences and the physical sciences, by showing the richness of the dialogue between scientists, we also hope to ease some of the obstacles that they face if they want to work together across traditional disciplinary boundaries. For example, peer review is a key part in the assessment of most grant applications. One of the problems the scientists who participated in the project highlighted was the difficulty of getting effective peer review and funding decisions in areas of science that draw on many disciplines. As a result, the project has set up its own peer review pool and is exploring other ways of enabling scientists to cross-fertilise.

Our hope is that in a small way this volume will also give potential reviewers an appreciation of possibilities that could come from supporting interdisciplinary projects in cognitive science.

A New Generation

The participants in the project identified the supply of young researchers as another important issue. If research in cognitive systems is to make the most of the progress throughout science, it will need contributions from researchers who are comfortable to work across traditional disciplinary boundaries while still being experts in their own field. This could be encouraged through 'cross discipline' PhD students, with supervisors from the life sciences and physical sciences. For established researchers, fellowships could provide opportunities for life scientists to acquire knowledge of the physical sciences, and vice versa. Here too, our hope is that this volume of research reviews will prompt young researchers to ask themselves how they can become a part of this increasingly lively research area.

The project did not set out to solve all of the problems of brain science or IT research. Its value has been in encouraging discussions that simply could not happen without the involvement of experts from both domains. Plenty of research is in the same boat, and would benefit from an 'expertise transplant'. Thanks to the project, the research community is now much happier to accept that input.

Public Debate

A further issue assumed increasing importance as the project progressed. Research in natural and artificial cognitive systems has enormous social implications. If society is to appreciate the possibilities, to accept novel applications and technologies and to influence their development, it is important to debate the issues in advance. For many researchers, the Foresight Project was their first opportunity to discuss the implications of their work in a public forum. The present volume now allows a wider audience to begin to debate those issues. In particular, the review of research in social cognition, the discussion of advanced neuroscience techniques and the paper on the possible applications of artificial cognitive systems highlight many of the issues that will certainly provoke a lively public debate, one that could well match the current discussions on such issues as genetically modified foods and nanotechnology.

A New Dawn?

The respected 'grand old man' of modern molecular biology, Professor Sydney Brenner, recently recounted how, in the 1960s, an eminent computer expert rang him to ask if they could meet up to share ideas. The computer person wanted to know if he could borrow any ideas from biology in his own work. Professor Brenner, for his part, said that he was interested in ideas flowing the other way.

Sydney reports that, apart from having a nice lunch, and an interesting conversation, the two researchers concluded that there really wasn't very much that they could profitably borrow in the way of inspiration. Over the past couple of years the project has laid on a few nice lunches for biologists and computer experts. This time we would like to think the exchange has been more profitable. Indeed, we already know that scientists from 'across the divide' are formulating joint proposals for research projects. The funding agencies have expressed their willingness to support these endeavours. It may take time for this to bear fruit. But it seems most unlikely that a future Sidney Brenner will make similar observations about encounters of a discipline-bending nature.

SECTION

1

How to Design a Cognitive System

The growing complexity of computer systems is a catalyst for their designers to look to nature for ideas.

Section Contents

This page intentionally left blank

How to Design a Cognitive System: Introduction

Lionel Tarassenko and Richard Morris

As computer systems become more complex, the likelihood of system failure increases accordingly. The designers of tomorrow's computer systems are starting to include the ability of the system to self-repair at the top of their list of desirable characteristics. Scientists at IBM have recently come up with a list of characteristics for the next generation of computers which not only includes the ability to self-repair but also the ability to self-organize, the ability to adapt to changing environments or workload, the ability to interact with other systems in a dynamic way and the ability to anticipate users' actions.

The chapter 'Large-scale, small-scale systems', written by Jim Austin, Dave Cliff, Robert Ghanea-Hercock and Andy Wright, sets out to present a biology-inspired view of what these complex adaptive systems might be. As with neurobiology, they consider systems made up of large numbers of relatively simple components, for example ultra-massive parallel processors. The components of these systems may interact in non-linear ways. These can then give rise to large-scale behaviour which cannot necessarily be predicted from knowledge of the characteristics of the individual components and their small-scale local interactions. This phenomenon has sometimes been described, perhaps unhelpfully, as emergent behaviour or computation.

Cliff and Wright take the reader on a whistle-stop tour of artificial intelligence (AI), at the beginning of which they elegantly describe the engineering approach as 'seeking simply to create artificial systems that reliably exhibit some desired level of cognitive performance or behaviour'. They point out that, for much of its history, research in AI largely ignored biology. This is changing, prompted in part by the realization that the increasingly large computing systems being designed today are becoming more difficult to build and control. (It is no accident, however, that the Foresight Project adopted the all-inclusive banner of cognitive systems, rather than that of artificial intelligence.)

Biology also strongly influences a recent development in AI, autonomous agents. Cliff and colleagues define autonomous agents as 'entities that are capable of coordinating perception and action, for extended periods of time, and without human intervention, in the pursuit of some set of goals'. They include in their review both physical autonomous agents, such as robots, and agents with no physical embodiment, such as software agents that exist purely in virtual environments.

The interest in gathering insights from biology has been fuelled by the increasing availability of data concerning the properties and behaviour of the elements of complex biological systems at the individual level, be they genes, proteins or cells. David Willshaw, in Chapter 1, argues that one unifying principle of organization is self-organization,

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

which is found throughout the biological and physical world.

Many of the intricate patterns seen in nature, such as the patterns of zebra stripes, the paths formed by social insects, cloud convection and snow-flake patterns, are examples of self-organization. Willshaw defines self-organization as 'those aspects of organization that result from interactions between the elements of the system as well as external influences that do not themselves provide ordering information'. He identifies three forms of self-organization: self-organization in development, self-organization as a complement to experiential changes and self-organization as a complement to damage. More than half of the chapter is devoted to the first of these.

During development, self-organization relieves the genome of much of the burden of specifying the exact numbers and positioning of nerve cells and the connections that they make. The internal, self-organizing dynamics combine with external influences, such as random activity in the participating nerve cells.

There is much less that can be said about how self-organization operates during *cognitive* development, within the processes of memory storage and retrieval and as a response to insult, in all cases acting against a background of continual neural change. Willshaw in Chapter 1 and Austin and colleagues in Chapter 2 agree that knowledge about how the nervous system continually self-organizes in response to change will be relevant to the design of artificial cognitive systems.

One issue to be faced in the design of the large distributed systems described by Wright is how to organize large amounts of data for efficient storage and rapid retrieval. Willshaw suggests that these large-scale systems may need to rely on software agents that independently harvest information for integration and self-organize to maximize their utility to the overall system.

Cliff and colleagues in Chapter 2 paint a picture of a future in which federated networks of computing facilities will house tens of thousands of servers, all connected on an ultra-high bandwidth network and providing computing on demand. These facilities, which could come on-stream within the next five years, will use techniques inspired by biology to provide self-healing resilience to load fluctuations, component failures and attack by computer viruses and worms.

Willshaw speculates that the self-organizing capabilities of complex biological systems could help to create a new generation of hardware devices that dynamically and organically reconfigure themselves. This is echoed in the 20-year 'vision' sketched out by Cliff and colleagues where silicon is no longer the dominant substrate for computing devices, being replaced instead by genetically engineered organic substrates. However, Cliff and Ghanea-Hercock also point out that this vision of the future is threatened by the pace of developments in quantum computing. Does self-organization play a part at the quantum level?

CHAPTER

1

Self-organization in the Nervous System

David Willshaw

1 INTRODUCTION

The term self-organization is commonly held to describe the process by which individuals organize their communal behaviour to create global order by interactions amongst themselves rather than through external intervention or instruction. Despite this term receiving only scant mention in dictionaries, it has been used to describe many different types of activities. The clouds formed by birds in the sky, the coordinated movement of schools of fish or the paths formed by ants, as well as the intricate patterns seen in snowflakes are all the results of self-organization. Other complex examples of spatial patterns are the many man-made or natural crystal structures.

In physics, the simplest examples are closed systems, where the system acts independently of external influences. The future state of the system is then controlled by its constitutive elements. Crucially, the emergence of a global pattern of order requires interactions between elements. Cooperative interactions will iron out local variations whereas competitive interactions will exaggerate them.

In the visually stunning Belouzov–Zhabotinsky reaction, two chemicals inhibit each other's autocatalysis, resulting in striking periodic changes in colour, as indicated

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

by an appropriate dye. In magnetic materials, it is energetically favourable for the dipoles of neighbouring atoms to co-align, resulting in a global magnetic field. An example where there is a simple external influence is found in a laser. At low levels of excitation, individual atoms emit their light independently to produce incoherent light; at higher levels, the emission of light from all the atoms becomes highly coordinated through local interactions, producing coherent light.

Many systems exhibit both competition and cooperation. A well-analysed example of a temporal pattern of self-organization in biology is found in the statistics of the populations of hares and their predators, lynxes, as recorded by the Hudson Bay Trading Company in Canada between 1849 and 1930 (Murray, 1993). Analysis of the number of pelts collected suggests the following pattern of events: a large fluctuation in one population can upset equilibrium states, in which the rates of reproduction and death of both species balance out. For example, a decrease in the number of prey will cause a corresponding decrease in the number of predators, who will have less food. The presence of fewer predators will then increase the number of prey and consequently will increase the number of predators, until finally the preys will decrease in number again. This pattern of events will repeat over and over again, yielding the cyclical variation in both prey and predator numbers over time that is seen in the records. Clearly this behaviour emerges from interactions between lynxes and hares and thus is an example of self-organization.

1.1 Self-organization in the Nervous System

As the words suggest, order in a self-organizing system emerges through local interactions between individuals in the absence of any external influence. As a highly complex and dynamic system involving many different elements interacting with each other, the nervous system displays many features of self-organization. However, there will be very few, if any, examples of true self-organization within the nervous system.

It is very likely that the organization of regions of the nervous system depends on external influences, either from other regions of the nervous system of the body or under the influence of external stimuli, such as sensory stimulation from the outside world. The resulting organization will be the result of interactions between the elements of the system itself as constrained by the particular boundary conditions that are in force, together with ongoing external influences.

1.2 Outline of the Chapter

I take the term self-organization to refer to those aspects of organization that result from interactions between the elements of the system as well as with external influences that do not themselves provide ordering information. I identify three forms of neural self-organization, which I shall discuss in turn. These are:

- *Self-organization in development* Since a key challenge in our understanding of the nervous system is to comprehend how such a highly structured yet complex system can emerge from a single fertilized egg, many phenomena displaying self-organization are concerned with how the nervous system develops. Many of these developmental processes are a result of interactions within the system itself. External influences exist but they can be regarded as initial constraints or boundary conditions acting on the system.
- *Self-organization as a complement to experiential changes* This refers to later stages in development, when self-organization plays a role along with other mechanisms such as those involving external signals arising from the sensory environment. I examine the effects of external influences only when these do not contain any patterning information. Therefore I do not discuss the neurobiology of learning and memory, where specific patterns of activity are required to be stored in or recalled from the system.

- *Self-organization as a complement to damage* The adult nervous system can respond to surgical or accidental damage. The facility for damaged brain to regenerate is either minimal or non-existent, which implies that the brain can self-organize, allowing healthy regions to take over functions previously carried out by other regions.

Section 2 considers development. I introduce some concepts of development at the genetic and molecular level. I then describe self-organization in the formation of pattern within collections of cells (section 2.1), in producing the correct numbers of cells (section 2.2) and in the formation of ordered nerve connections (section 2.3).

In section 3, I look at the role of self-organization in experiential changes. Section 3.1 describes the self-organization of patterns of feature selectivity in the cortex and section 3.2 provides a brief introduction to the self-organization of cognitive function.

Section 4 is concerned with self-organization as a response to injury, principally in the adult.

Finally, in section 5 I discuss some open questions that are relevant to the subject of this essay. Section 6 gives a short reading list.

2 SELF-ORGANIZATION IN DEVELOPMENT

Generating nerve cells of the right type, in the right numbers, in the right places and with the right connections is a formidable task. It involves cell division, cell migration, cell death and the formation and withdrawal of synapses. The essential steps of embryonic development are reviewed in many books. Wolpert (1991) provides a simple readable introduction: Price and Willshaw (2000) discuss mammalian neural development.

Every organism is defined by the sets of genes in its genome. This contains the initial instructions from which development proceeds. The set of three-letter 'words' obtained by reading the sequence of bases along the DNA defines a sequence of amino acids. Proteins are made out of amino acids and cells are made out of proteins.

There has been considerable progress in our understanding of how genes control development. The fruit fly, *Drosophila melanogaster*, has been used intensively in genetic research for many decades. It is small, has a short life cycle of two weeks, and large numbers of mutants have been identified and studied. The combination of the extensive knowledge of mutants and experimental embryological and molecular biological techniques has provided a profound understanding of the genetic regulation (i.e. control) of development in this species.

Remarkably, not only have many of the control mechanisms that operate in *Drosophila* been conserved in mammals, but so have many of the genes themselves. It is now commonplace to use information obtained from studies of *Drosophila* to search for specific regulatory genes in higher species and to formulate hypotheses regarding the general principles that underlie development in all organisms. In particular, work on *Drosophila* has provided a comprehensive understanding of how different regions of a developing organism can develop regional specificity. For example, certain morphogens – molecules that control the development of form, or morphogenesis, a term coined by Turing (1952) – are distributed in gradients in the early *Drosophila* embryo. They evoke different cellular responses at different concentrations, specifying the expression patterns of other genes that themselves regulate later-expressed genes. In this way, complex patterns of later-expressed genes emerge to confer positional identity on cells at each position in the embryo. The combined action of the specific cocktail of regulatory genes that each cell expresses is essential for conferring on each cell a particular phenotype appropriate for its position.

Many groups have shown that vertebrates have genes that are similar to those of *Drosophila*. Researchers have found vertebrate homologues for *Drosophila* genes that act within cells to regulate the expression of other genes (transcription factors) or that signal between cells to control processes such as axonal guidance. A good example of transcription factors is the large family of Hox genes (members of the homeotic clusters of genes) in mouse which have homology to the genes of the Antennapedia complex of *Drosophila* and which regulate the identity of segments of the *Drosophila* body. Another good example of conservation of developmental mechanisms is in the guidance of axons. Many of the receptor systems that have been implicated in this process are highly conserved between *Drosophila*, other invertebrate species and mammals.

It might be thought therefore that the genome could hold a coordinate-by-coordinate blueprint of the nervous system, specifying where each nerve cell is to be situated, what its functional properties are to be, and which other cells are to be contacted. If homeotic genes control the production of gross anatomical structure and cell differentiation, is it not possible that subordinate gene families subsequently control the remaining development processes? This is extremely unlikely given the large numbers of nerve cells and the many more connections that they make compared to the relatively small size of the genome.

If the 10^{14} connections between the 10^{10} neurons of the human neocortex were made at random, this would require at least 10^{15} bits of information compared to the 10^{9} bits in the human genome. It is more likely that the genome contains the 'rules of development'. For example, it is well known that the connectivity of the brain is highly structured, with topographic maps found between many sensory structures and neocortex. The rules of development would specify the general features of the mapping and the fine details could be arranged through interactions between the constituent parts. Many of the features of the system are, so to speak, arranged by the nervous system itself, or self-organized. The next three subsections describe different parts of the developmental process where mechanisms of self-organization make an important contribution.

2.1 Self-organization and Pattern Formation

A central aim of developmental biology is to understand how cells in different positions develop differently; i.e. how regional specification comes about. This is as true for the development of a structure as complex as the cerebral cortex – where each point in the dorsal telencephalic wall (the precursor of the cortex) acquires a unique functional property, with relative invariance in the layout of these properties between the individuals of the same species – as it is for the development of the five distinct digits of the hand or the patterns of markings on seashells, zebras or leopards. Many of the principles that govern the ways in which regional specification arises during development have been identified in studies of early embryogenesis. We are relatively ignorant of the mechanisms that control brain regionalization, which makes it all the more important to generate hypotheses with knowledge of principles deduced from studies of earlier developing systems.

The key questions are:

- How does a mass of developing cells acquire differences one from another?
- How is this information used to determine their different fates?

These two questions are interlinked. For example, does each cell acquire its own identity independently from the instructions in the genome? Or do certain cells act as organizers and instruct their neighbours, which happens in limb morphogenesis?

Turing (1952) investigated one possibility theoretically. He analysed the emergence of pattern in a collection of cells and showed that, starting from a uniform concentration of morphogens, which interact with each other and diffuse between cells (giving rise to the term reaction-diffusion), the patterns of molecular concentration produced over the cell population had peaks and troughs defining characteristic periodicities. Turing argued that in a developing organism, a set of chemicals could be used in this way to set up a prepattern that will determine specific features of the organism. The patterns are not preprogrammed and emerge through self-organization.

In the brain, generation of regional differences has to occur at different levels. How are particular areas or regions within a brain nucleus or system distinguished one from another? How is cellular identity within a given region determined? I now discuss regional specification within the forebrain and within the neocortex. I discuss specification within a brain region in section 2.3.

2.1.1 Pattern Formation in the Specification of Forebrain

Much of our understanding derives from work on the amphibian *Xenopus laevis*. Two important terms used to describe complementary mechanisms that can generate regional specification are mosaicism and regulation, i.e. whether the development of an individual cell is independent of or dependent on the development of other cells.

Regulative behaviour is commonly found among cells undergoing regional specification in the developing mammalian nervous system, which indicates that the mechanism of specification involves intercellular signalling. In early embryogenesis, major sources of such signals are well defined, and include the so-called Spemann organizer (Spemann, 1938). The signals that affect the developmental pathway are termed inductive signals as the process involves transfer of information from mesoderm to ectoderm, two of the three germ layers formed very early in development. Although inductive signalling is almost certainly a widespread mechanism in the later stages of cortical regionalization, its clearest roles are in the early stages of forebrain development.

The very early regionalization of the developing forebrain can be detected by morphological criteria and by analysis of the discrete domains of expression of regulatory genes. It is possible that the many genes known to be involved give each region of the developing forebrain a unique identity, probably through combinatorial actions. They may do this by controlling the expression of numerous other genes required for the characteristic morphological differentiation of that region. Amongst the molecules known to be involved are the diffusible proteins notch and delta, the wnt family of glycoproteins and the hedgehog family of proteins first identified in *Drosophila*.

Regional specificity of gene expression in the telencephalon, from which the forebrain develops, is likely to control regional differences in morphological characteristics, through actions on the cellular processes of proliferation, migration and differentiation. How different regions come to express different genes in the first place is a subject of speculation. One simple possibility is that a small number of genes distributed over the neural plate and very early neural tube, the forerunners of the nervous system, generate gradients of molecules.

Through transport and inter-cellular exchange, molecules at different levels of concentration would become localized in different cells. This can create domains of gene expression with sharp boundaries. This type of process is known to generate regionalized domains of gene expression in the early embryo of *Drosophila*. Although there are homologues of these genes in the mammalian forebrain, drawing close parallels between mammalian forebrain and *Drosophila* development may be dangerous given the differences between them at a cellular level. Nonetheless, the principle that continuous molecular gradients may be read out to create domains of expression of other genes distributed with discrete levels is well established. There are various models for how this can be done. These models are usually formulated according to the concept of positional information and are constrained by the regulatory phenomena often seen in embryogenesis.

Positional information Evidence from classical embryological experiments on a mass of cells after the removal of some cells or the transposition of cells to a new position resulted in the proposal that the fate of a cell is determined by its position within the morphogenetic field of cells. A particular set of cells that makes a single field can form its own organ when transplanted to a foreign site and cells within the field can regulate to take over the function of other cells that are removed from it. How is each cell within the field instructed or, as expressed by Wolpert (1969), how does the cell acquire its positional information? As already discussed, one fundamental way in which information is supplied in development is through inducing signals supplied through extracellular means and various ways of assigning differences amongst cells by means of morphogens have been considered. In the simplest case of a one-dimensional field of cells that specifies the digits of the hand, for example, a gradient of morphogen would enable different parts of the field to be distinguished; specifying particular threshold values of morphogen would determine which cells would develop into which digit.

Simple source/sink models Various different ways of producing spatially varying profiles of a putative morphogen have been considered. For a single dimension, morphogen flows from a single source to a single sink to set up a graded variant of morphogen down the line of cells. Alternatively there could be a single source and all cells acts as sinks through leakage and other forms of loss. These models have been found to be unsuitable. In particular, they do not adapt in the required fashion following perturbations such as the removal of a substantial number of cells.

Reaction–diffusion model Gierer and Meinhardt proposed a model of the reaction–diffusion type in which there are two molecules with different properties: an activator, which stimulates its own production, and an inhibitor, which diffuses at a faster rate (for a review, see Meinhardt, 1982). The activator stimulates production of the inhibitor but the inhibitor represses the production of the activator. A small local increase in the amount of activator will result in more activator being produced, thus giving rise to a local source of this molecule. The inhibitor produced as a result will spread out more quickly than the activator and so a sink for activator will be established nearby. In this way, spatial patterns of activator and inhibitor become distributed across the array of cells. In these reaction–diffusion models, a crucial parameter is the size of the morphogenetic field over which the pattern is being formed compared with the diffusion lengths of the two molecules. If the field is very much smaller than the diffusion lengths, periodically repeating patterns will be produced; if the field is comparable in size to these diffusion lengths, a single gradient of morphogen results. Imposing a weak gradient of activator production to determine polarity yields a single gradient. Diminution of field size causes the full gradient to be restored, up to a limit. This is important as an explanation of the findings in developmental biology that in some animals structures can regenerate from partial structures.

Reaction–diffusion mechanisms have been applied to the generation of many different patterns, such as stripes, spots and other markings that appear on animal coats, and to other naturally occurring patterns, such as those on butterfly wings or those on sea-shells (Meinhardt, 1982; Murray, 1993). There is a close relationship between these mechanisms, involving different types of non-linear interactions, and the self-organizing systems studied in physics.

Role of gradients The primary role to be fulfilled by systems of gradients is to provide a way for cells to be distinguished from one another. The reaction–diffusion scheme at least provides a way of doing this which is resistant (within limits) to changes in morphogenetic field size. It is assumed that a separate mechanism translates an amount of morphogen into an instruction to build a cellular structure. In some cases, patterns of morphogens are required to specify the coordinate systems of developing organs. It is natural, although not necessary, to assume that the axes of the morphogens will match those of the required coordinate system. For example, a rectangular coordinate system might be provided by two morphogens, each identified with one axis. In cases where there is no such requirement, as long as cells can be distinguished one from another, the pattern of morphogens can be arbitrary.

2.1.2 Pattern Formation in the Specification of Neocortex

The neocortex, the uniquely mammalian structure which has evolved rapidly and extensively in primates, is thought to be the source of our highly developed cognitive functions. We can subdivide it into distinct areas, according to anatomical and functional criteria. There has been much discussion of how these distinct areas of neocortex develop from the early cortical plate, with a relatively homogeneous appearance. It has been suggested that neocortical organization is determined by:

- its afferents (inputs)
- the significant amount of information preprogrammed into the neocortex
- interactions within the developing neocortex, independently of its inputs.

In the adult mammal, the cerebral cortex has been divided into areas according to their histological appearance. This is illustrated in Fig. 1.1, which shows the cytoarchitectonic fields of the human brain as defined almost

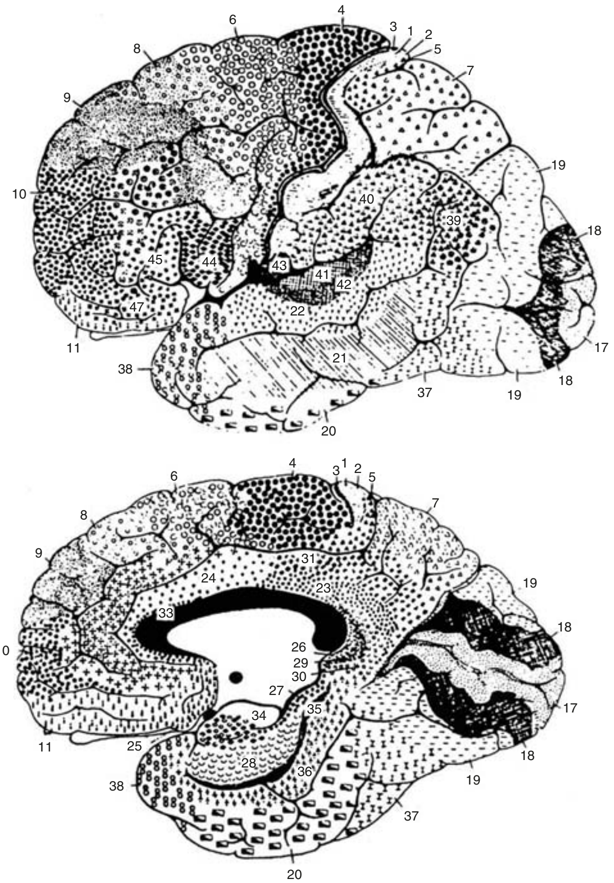


FIGURE 1.1 Brodmann's cytoarchitectonic fields of the human brain marked on the left cerebral hemisphere, seen from the lateral (*top*) and medial (*bottom*) aspects. (After Brodmann, 1909)

a century ago by Brodmann (1909). These distinct fields were defined according to the relative thickness of cell and fibre layers and Brodmann was able to delineate sharp borders between neighbouring areas. This analysis of the human brain has been extended to other species, showing that their cortices can also be subdivided into cytoarchitectonic fields, with equivalent fields occupying relatively similar positions. Most importantly, these anatomically defined regions have different functional specializations.

The question of how much of the regionalization of the cortex is imposed on it by the order of its afferents (i.e. inputs) and how much is specified before innervation has become a major preoccupation. One extreme view is that the cortex is a naive sheet of cells whose identities are determined by the nature, the order and the detailed connectivity of the afferents that they receive. Another view is that the cells of the cortex do have at least some regional identity before they become innervated.

There is now considerable electrophysiological evidence from the results of embryonic and neonatal transplantations that differences between distinct areas in the neocortex are induced by the afferent thalamocortical axons that innervate the cortex. However, some area-specific differences are detectable before any cortical innervation has taken place. There is evidence for region-specific differences in the rates of proliferation and expression of molecules in the cerebral cortex prior to innervation. In some cases these molecular differences are not altered when expressing regions are transplanted to non-expressing sites, suggesting that region-specific differences may be determined (i.e., irreversible) prior to innervation. The abolition of the activity in cortical afferents does not prevent the development in the cortex of characteristic region-specific distributions of molecules. In addition, there are several reports of cortical area-specific gene expression that begins before or is independent of afferent innervation.

In mouse, a specific gene is expressed in the somatosensory cortex from before the time of afferent innervation. The gene is still expressed if the developing somatosensory cortex is transplanted to an ectopic location but is not expressed by other regions of developing cortex even if they are transplanted to the somatosensory cortex. Most recently, the arrangement of cortical areas in two different mouse mutants that lack thalamocortical connections have been found to be abnormal. These results argue in favour of cortical cells having some positional identity without innervation, although how much is far from clear.

Other experiments have addressed this issue by studying the properties of different cortical regions either in culture or after transplantation. Tissue culture experiments to investigate the specificity of axons from different thalamic regions for different cortical areas showed that axons from thalamic explants exhibit no preference for the area of neocortex with which they were cultured. They grew equally well on their normal target areas as on other non-target areas of the neocortex.

Other experiments have involved the transplantation of regions of the developing cortex to abnormal sites. In transplants between neocortical regions, the donor tissue was found to develop attributes of the new host region rather than retaining its normal attributes. Pieces of visual cortex grafted into motor cortex developed persistent projections to the spinal cord, as does normal motor cortex but unlike normal visual cortex. When pieces of motor cortex were grafted into the visual cortex, they developed persistent projections to the superior colliculus, as does normal visual cortex, but unlike normal motor cortex.

Sur *et al.* (1988) carried out experiments on the regeneration of connections in ferrets where target structures of sensory fibres were removed, leading to the fibres being diverted to other cortical structures. Removal of lateral geniculate nucleus (a relay station between retina and cortex) and visual cortex led to visual afferents innervating medial geniculate nucleus (the destination of auditory fibres). The result was that cells of the auditory cortex became responsive to visual stimuli and acquired the functional characteristics of cells of the visual cortex.

All of these experiments indicate that different regions of the embryonic and neonatal neocortex have a low level of commitment to their specific regional fates. Although it is widely accepted that the fates of embryonic cortical regions are not determined by birth (i.e. they are not irreversible), the degree to which they are specified remains uncertain. The results of the experiments described above suggest that it is easy to deflect developing neocortical regions from their normal developmental pathways. However, recent transplant experiments similar to those outlined above have come up with opposite results with no clear explanation for the difference. Furthermore, it may be harder to alter the fates of embryonic tissue transplanted between the neocortex and other cortical areas, such as the limbic cortex.

2.1.3 Self-organization in Forebrain and Neocortical Development

To summarize section 2.1.1 and 2.1.2, there is much evidence at the genetic, molecular and neural levels for the roles of self-organizing influences in the development of regional specificity in these systems. We know most about the early stages of development of the forebrain, where inductive effects are important. In the development of cortical regionalization, the influences of specification prior to innervation and cortical afferents contribute. Several mathematical and computer models have been developed but at present they lack specific application.

2.1.4 Pattern Formation in the Positioning of Cells

For nerve cells to function correctly, they have to be placed in the correct position. To investigate how this is done, we need to look at systems where it is clear what it means to place cells correctly. This is most easily accomplished in parts of the nervous system where there is a high degree of order.

Nerve cells within invertebrates are well ordered, while the vertebrate nervous system is less highly ordered. The degree of order varies greatly between different parts of the nervous system. In mammals, hippocampus, olfactory cortex, cerebellar cortex and retina are examples of relatively ordered brain structures.

In the cerebellar cortex, the Purkinje cells (the main output cells) form, with other cell types, a regular three-dimensional lattice. These cells define regular, typically hexagonal, neighbourhoods with intercellular spacings that grow steadily in size during the first few postnatal weeks.

In the retina, neurons form regular arrays, called mosaics. These exist in many species, being more regular in invertebrates than vertebrates. Several cell types are distributed regularly in the two-dimensional plane of the retina, giving a constant cell-spacing between adjacent cells. In insects, receptor cells are highly ordered, forming precise hexagonal patterns.

It is important for cells to be arranged regularly across the retina. The vertebrate retina is organized in such a way that many local circuits can analyse each part of the image. This allows it to handle the vast information flux of a complex, ever-changing visual scene using only slow, noisy neurons. The circuits work in parallel to assess different static or dynamic aspects of colour, brightness and contrast. As it is the regular repetition of these circuits across the retina that gives rise to mosaics, each mosaic can be assumed to embody a unique function in sensory processing.

There are different types of mosaics, such as those involving cholinergic cells (amacrine cells with acetylcholine as transmitter), horizontal cells, different classes of ganglion cells and the different types of photoreceptor cells. The mosaics appear early on in development. They form while the cell membership is still forming, by the processes of neurogenesis, cell death and cell migration. For example, in developing cholinergic mosaics, as new cells enter the array, neurons move sideways to preserve a constant inter-cell spacing.

It is thought that the rudiments of the pattern are determined at a very coarse scale by molecular markers, derived from genes such as the pax family of homeotic genes, in the embryonic neuroepithelium, which forms the eye. The ordered spacing can be simulated by a very simple local exclusion rule, applied to cells of the same type, which specifies the minimum distance between cells.

Recent research suggests how this rule can be implemented at the molecular level. A computer simulation study has shown that regular, advancing arrays such as cone mosaics can emerge from simple cellular-automaton rules that are applied to initially random arrangements, but also that many different sets of these rules converge on the same simple patterns.

Although information must be supplied to position cells in the correct general region, this solution has the advantage that there is no need for a means of pre-specifying the positioning of nerve cells precisely. There are other advantages. Since interactions are short range, neither local errors nor the introduction of new cells at the periphery of the array disturb the pattern. If the local interactions are restricted to cells of the same type, introduction of an array of cells of a different type will not perturb the arrangements in the preexisting arrays. Finally, the ability of cells to recognize others within a limited range could lay the basis for the formation of the topographically ordered maps of connections that exist in the retina and which subserve visual processing (see section 2.3).

2.2 Making the Correct Numbers of Cells: Cell Death

Of all the mechanisms involved in the formation and maintenance of the nervous system, cell death is the best understood, especially at the level of how genetic instructions can bring about cellular self-destruction. It has long been realized that cell death can be a physiological as well as a pathological process, i.e., that many cells die even during normal brain development.

Cell death during animal development was first observed by Vogt in 1842, in amphibia. The basic findings are:

- there is substantial motor neuron death in normal development
- removal of a developing limb bud from chick embryos causes increased death of the motor neurons
- some of the motor neurons that would have died during normal development can be rescued by grafting in an extra limb bud.

Researchers have reported similar findings in *Xenopus*. Subsequent studies have shown that cell death is a normal occurrence amongst many neuronal populations in the developing vertebrate nervous system, taking place when axons begin to reach and activate their targets. The number of neurons in a given population first rises then declines towards the constant adult number. The proportion of cells that dies varies among different neuronal populations, ranging from the removal of only a small number of the neurons in some regions to more than half the population in others. In the retina, results from a range of mammalian species indicate that 50–90% of the retinal ganglion cell population will die during development.

That so much cell death occurs during normal development is counterintuitive. Thinking anthropomorphically, it appears wasteful. It is not clear why evolution should have selected such a developmental process.

Blocking normal neuronal death – by making transgenic mice with mutations of genes that regulate cell death, thereby increasing the numbers of neurons – does not affect lifespan. None the less, cell death is a significant developmental process that demands an explanation both in terms of its role in development and the molecular mechanisms that control it.

Researchers have advanced various explanations for nerve cell death, amongst them being:

- failure of neurons to find their target
- failure to make the correct connections

- the elimination of entire structures that may act as transient scaffolds; examples of this are the subplate, a layer of cells that is critical for the development of cerebral cortex, and the Rohan–Beard cells in amphibian embryos, which are temporary sensory neurons
- removal of transient branches of the tree of lineage; this seems to be the case in invertebrates – in the nematode *C. elegans*, around 20% of the 300 nerve cells generated by cell division are preprogrammed to die
- lack of adequate innervation, as is the case in insect optic lobe.

All these explanations seem to apply in special cases.

A commonly held view is that in many cases this substantial amount of nerve-cell death results from the action of a mechanism that matches the number of presynaptic cells to the number of postsynaptic cells. One hypothesis for this is that neurons compete for a supply of one or more so-called 'neurotrophic' factors that are known to be produced by their target cells. According to this neurotrophic hypothesis, insufficient neurotrophic factor is produced to support the excessive numbers of neurons generated and those that are unsuccessful in the competition die.

Additional support for the idea of matching presynaptic and postsynaptic cell numbers has come from experiments in which chick lumbosacral cords were transplanted into quails and vice versa before the limbs became innervated. Chicks are larger than quails and have bigger muscles with more muscle fibres. More quail motor neurons survived in the chick than in the native quail, and fewer chick motor neurons survived in the quail than in their own environment. There was a correlation between the number of motor neurons surviving and the number of muscle fibres available for innervation.

The nervous system contains control mechanisms by which the numbers of presynaptic and postsynaptic cells are matched apparently automatically. This involves the apparently destructive side-effect of cell death. This self-organizing mechanism has great benefit, making it unnecessary to have ultra-precise controls over the generation of neuron numbers. Whilst qualitatively the effect is well established, the underlying mechanisms and their quantitative implications represent exciting challenges for the future.

2.3 Development of Connections

Once nerve cell axons have found their correct target within the nervous system, neuron numbers have been adjusted to their adult levels and neurons have been correctly positioned, the appropriate connections have to be made. It is generally thought that this happens in two stages. Initially, the axonal terminals are distributed across the target relatively diffusely. Subsequently, there is a rearrangement or refinement of connections. This picture has emerged from many different animal preparations, principally from experiments on the innervation of skeletal muscles and autonomic ganglia of neonatal rats, and on the visual pathways of *Xenopus*, kittens, ferrets and infant monkeys.

There are good reasons why some sort of refinement of connections is essential. First, far too many neurons exist for the positioning of each to be controlled by a genetically determined programme. Secondly, it is difficult to see how such a programme can determine the fine details of the connections between independently specified sets of neurons. Finally, as a consequence of the continual growth of some animals there has to be a continual remapping of connections during development to accommodate the generation of new cells and consequently new connections.

The two stages are thought to involve mechanisms that guide axons to their initial, approximate destination to generate an initial pattern of connections, followed by a stage during which connections are remodelled to form the adult configuration, which involves the loss of existing connections and the generation of new ones. Many people think of the first stage as being programmed genetically and the second one driven by neural activity so that the refinement of connections is sculpted to fit the uses to which the neural system has to be put. The precise division between these mechanisms is not clear, particularly in that it is not clear how much specificity of connection is imparted during each stage.

I will describe the role of self-organization at two levels. I will first discuss the network level, as exemplified by the development of ordered maps of connections between the vertebrate retina and optic tectum in lower vertebrates (equivalent to the superior colliculus in mammals). I will then consider the single-cell level, as exemplified by the elimination of connections from the developing neuromuscular junction.

2.3.1 Map Formation

A striking feature of many of the connection patterns between collections of nerve cells is that they are highly ordered. Evidence for this comes mainly from two types of experiment. First, in electrophysiological experiments, stimulation of a small region in one structure, such as a sensory surface or a nucleus, leads to activation of cells in a small region of its target. As the stimulus is moved systematically across the structure, the region that responds shifts in a corresponding fashion. The region in stimulus space that produces a response at a particular target position is called the 'receptive field'.

Secondly, the mapping between two points in different structures can often be established in anatomical experiments using axonal tracers (molecules that can be injected at discrete points to label axons running to and from those points). Tracers placed at one point in one structure typically label a small, circumscribed area in the target, the spatial layout of points of administration (in different animals) being reflected in the layout of points to which the tracers go in the target. Such ordered anatomical layouts of connections provide the substrates for the ordering observed in electrophysiological experiments.

Many neural maps are effectively projections of one two-dimensional surface onto another. For example, axons from each small cluster of ganglion cells in the mammalian retina project, via the lateral geniculate nucleus (LGN) of the thalamus, onto a small area of visual cortex, with the result that a map of the retina is spread over the surface of its target structure (Fig. 1.2*a*). In amphibia and fish the retina projects directly to the optic tectum where, once again, an orderly map of the retina is found. Auditory cortex contains an example of a one-dimensional map of frequency (Fig. 1.3*a*).

Another striking example is the existence of precise maps of connections in somatosensory cortex. Rodents make much more extensive use of tactile information than of visual information. Correspondingly, their somatosensory cortex is relatively large whereas their visual cortex is relatively small and simple. A large area of primary somatosensory cortex is occupied by an ordered representation of the facial whisker pad (Fig. 1.2*b*).

Anatomical investigations have revealed a set of barrel-like structures, with a one-to-one relationship between the arrangement of whiskers and the arrangement of barrels: where the muzzle contains an extra whisker, there is an extra barrel in the topographically equivalent place and vice versa. Neurophysiological recordings have established that activation of an individual whisker excites the cells in the corresponding barrel. There is also a topographical representation of the whiskers in each of the two nuclei which form the relay stations linking the sensors with the somatosensory cortex. This ordered map is not preserved throughout the length of the pathway from sensorium to cortex but rather is recreated at each individual relay station.

There is much evidence for plasticity in the whisker-to-barrel pathway in rodents. An intact sensory periphery is required during a certain critical period of development for the normal map to develop. When rows

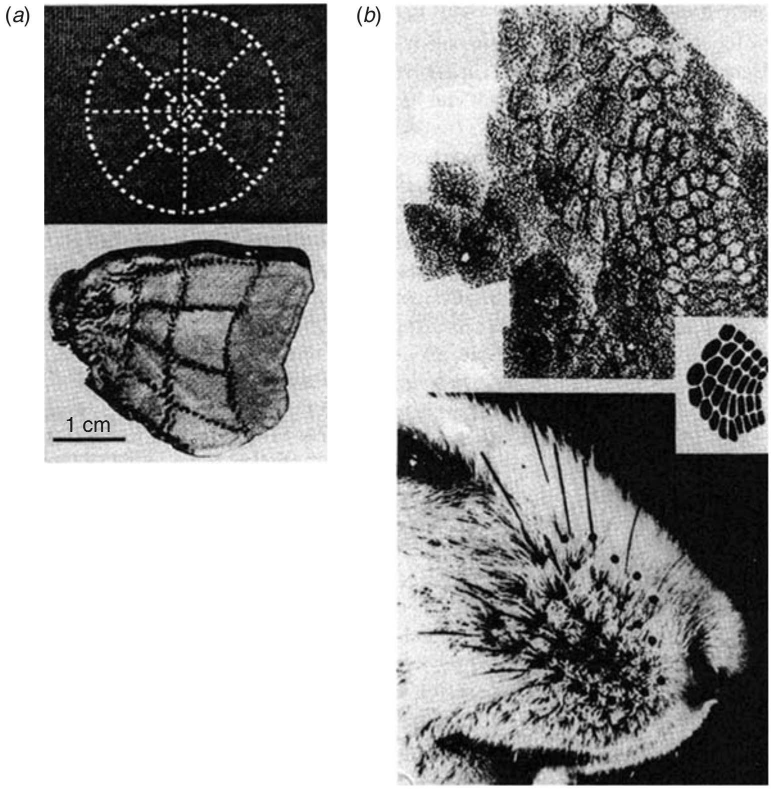


FIGURE 1.2 (*a*) Showing the ordered projection of the retina onto the visual cortex of monkey. The animal was exposed to the image (*upper*) and activity in the visual cortex was revealed by processing for deoxyglucose uptake (*lower*). Note the ordered projection and non-linearities. (Reproduced from Tootell *et al.*, 1982) (*b*) The pattern of whiskers on the adult rodent snout (*lower*) and the pattern of barrels in somatosensory cortex (*upper, schematized in inset*) to which the whiskers project in a one-to-one fashion. The pattern of whiskers is almost invariant from one animal to another. (Reproduced from Woolsey and van der Loos, 1970; Copyright acknowledged)

of follicles are injured at birth, before barrels form, the corresponding row of barrels is absent in the adult, and is replaced by a small barrel-less territory. Barrels develop in the first postnatal week and their morphology can be manipulated by the selective lesioning of the whisker follicles. The earlier the follicles are removed, the more extensive the resulting morphological aberration.

Generally, where axonal projections are through intermediate structures, the intermediate maps are themselves well-ordered. Ordered projections are the rule, although there are a few apparently disordered projections such as that of visual space onto the pyramidal cells of mammalian hippocampus, which in rats respond to specific locations in the animal's environment, and in the direct projection between cortex and striatum.

The mapping of the elements in one structure onto another is studied at the coarse-grained level, as in the case of the mapping of thalamic nuclei onto the cortex; or at a fine-grained level, as in the case of the mapping of cells within a particular thalamic nucleus onto a particular cortical region. Such maps are readily understood on the basis of zone-to-zone or point-to-point connections between the structures,

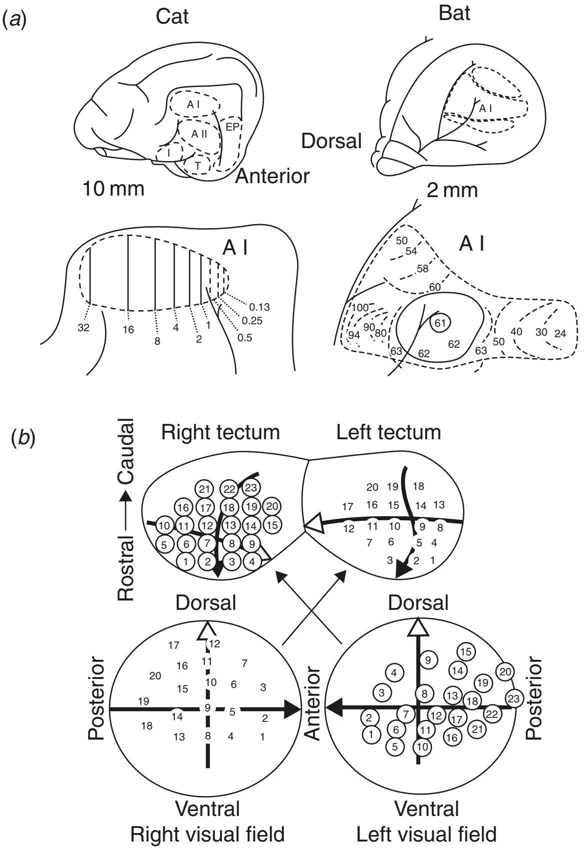


FIGURE 1.3 (*a*) The one-dimensional map of frequency in auditory cortex. (AI and AII, primary and secondary auditory cortex; EP, posterior ectosylvian gyrus; I, insula; T, temporal field; numbers 0.13–100 kHz.) There is a regular tonotopic representation in cat but a distortion of this regularity in bat by a large representation of 61–62 kHz, the frequency of its echolocating signal (based on Suga, 1978 and Shepherd, 1994). (*b*) The ordered projection from retina to contralateral tectum in adult Xenopus laevis. The numbers indicate where in the visual field a small point stimulus evoked maximal response at the correspondingly numbered tectal position. (Reproduced from Jacobson, 1967)

mediated either by bundles of axons (in the case of coarse-grained mapping) or by individual axons (in the case of fine-grained mapping).

In a complex structure such as the cerebral cortex, more complex properties of the sensory environment than position in the field are detected by specific nerve cells. The properties of the stimulus that produces maximal excitation varies over the cortex, defining 'feature maps' (see section 3.1).

2.3.2 Topographic Map Formation in the Visual System

Most work on topographic maps has been carried out on the vertebrate visual system. Here I describe the direct projection of retina onto optic tectum (the analogue of the superior colliculus in mammals) in amphibia and fish. The first crude maps were constructed from the results of axon degeneration studies but the first maps with any precision were made by extracellular recording from the optic tecta of goldfish, and of the frog, *Rana*, and *Xenopus* (Fig. 1.3*b*). The connections are not precise at the cell-to-cell level (as in invertebrates) but in the retinotectal map in *Xenopus*, for example, at least 50 recording positions can be distinguished, all arranged in topographic order. The other important attribute of such maps is that they always have a specific orientation. All retinotectal maps are arranged so that temporal retina projects to rostral tectum and dorsal retina to medial tectum (Fig. 1.3*b*).

The problem of understanding how maps are formed was originally formulated in terms of the establishment of a functional relationship between the cells in the two sets of cells. With the advent of powerful methods for tracing patterns of connections, this problem has become that of specifying the connections themselves.

One powerful set of results, generated from electrophysiological recordings carried out in the 1970s, involves experiments on the regeneration of connections in the retinotectal system of adult goldfish. The paradigm experiment involves removing part of the retina, allowing connections to regenerate and then making a map of the remaining part of the retina onto the tectum. After several months, retinal fibres had expanded to innervate the entire tectum. Complementary experiments revealed that after removal of half the tectum from goldfish, the projection from the entire retina eventually becomes compressed, in order, on to the surviving half-tectum. The basic result in these so-called 'mismatch' experiments is that the retina and the tectum match together as a system, independently of the actual sizes of both structures.

Although these results were found in the regeneration of connections, exactly the same type of phenomenon occurs during development. In *Xenopus laevis*, the retina and tectum grow in different ways. New retinal cells are added on in rings to the outside of the retina whereas tectal growth predominantly involves addition of cells to the back. None the less, there is an ordered projection of retina onto tectum from a very early stage. This implies a gradual shifting of connections. For example, throughout development, central retina always projects to central tectum, the position of which moves progressively backwards as more tectal cells are added. This inference was first made from extracellular recordings and then confirmed by electron microscopy studies which demonstrated the degeneration of synapses as axons shifted their positions during development. Later work on frog tadpoles that combined electrophysiological and electron microscopy confirmed this by showing that retinal ganglion cell axons move across the tectum during development, continually changing their tectal partners as they do so.

These results demonstrate that connections cannot be made by means of a simple set of instructions specifying which cell connects to which other cell; more likely, the two populations of cells self-organize their connections so as to ensure the correct overall pattern. Several different hypotheses for the formation of nerve connections in this paradigm system have been made. The two main contenders are:

1. As first proposed by the Nobel laureate Roger Sperry (1963) in his doctrine of chemospecificity, the two populations of nerve cells, one in the retina and one in the tectum, are labelled separately by sets of molecular markers. By some means, the correspondence between the two sets of markers is communicated to the participating cells. Each retinal axon then uses this information to find its correct tectal partner. In addition, there is some means of regulating the constitution of the molecular labels, when required, to account for the lability of connections during development and regeneration. This proposal has received renewed interest recently due to the discovery of a type of receptor located in the retina, the Eph receptor, and the associated neurotrophins located in the tectum, the ephrins, which bind to these receptors. The Ephs and the ephrins could be the markers that label the

two sets of cells. The origin of the markers themselves has not yet been linked to the generation of regional specificity discussed in section 2.1.

2. Initially, a roughly ordered map of connections is made with subsequent refinement of the map through electrical activity.

At present, the status of both contending proposals is unclear. They lack experimental verification at the mechanistic level. For (1), it could be that the tectum acquires its markers from the retina (Willshaw and von der Malsburg, 1979), thereby ensuring that the two sets of markers are properly coordinated. For (2), a Hebbian type synaptic modification mechanism[1] might operate. By reinforcing the contacts made by neighbouring retinal cells to neighbouring tectal cells at the expense of the connections made by non-neighbouring cells (Willshaw and von der Malsburg, 1976), each pair of neighbouring presynaptic cells comes to connect to postsynaptic cells that also are neighbours, resulting in a topographically ordered map.

2.3.3 *The Elimination of Superinnervation from Developing Muscle*

The second part of the two-stage process thought to underlie the development of nerve connections is illustrated by cases where the development of the connections on individual targets has been monitored, as in many vertebrate skeletal muscles. In the adult each muscle fibre is innervated at its endplate by a single motor neuron. This pattern of innervation arises from an initial state in which there is innervation from a number of different axons at a single endplate.

During early development, contacts are withdrawn until the adult configuration is reached (Fig. 1.4). The same pattern of events takes place in the adult after transection of the motor nerve. In the initial stages of reinnervation, muscle fibres are superinnervated and this pattern is transformed into one of single innervation after a few weeks. Muscles vary in size. The soleus muscle is one of the larger muscles in rat with around 3500 fibres innervated by some 25 motor neurons, so that in the adult each motor neuron innervates on average 140 muscle fibres; the rat lumbrical muscle has about 100 muscle fibres and 10 motor neurons.

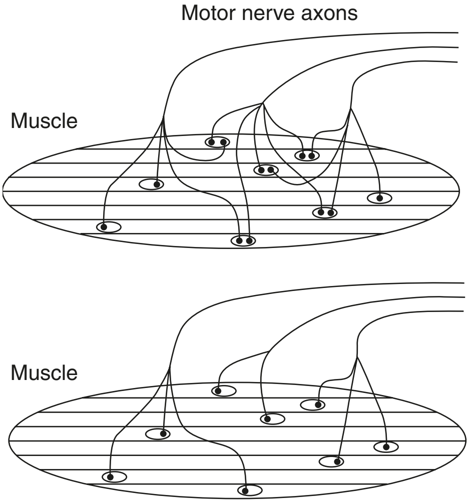


FIGURE 1.4 Schematic of the states of innervation of mammalian skeletal muscle: (*top*) the initial pattern of superinnervation; (*bottom*) the adult state where each muscle fibre has contact from a single axon. (After Rasmussen and Willshaw, 1993)

This developmental loss of synaptic contacts has been observed at both central and peripheral sites, in systems as diverse as the neuromuscular junction of invertebrates and the cerebral cortex of primates. The precise time course over which synapse elimination occurs and the proportion of afferents lost varies greatly between areas of the nervous system, even within a single species. In neonatal rat skeletal muscles, there are on average four to five contacts per fibre initially. This reduces to exactly one per muscle fibre over the next two weeks.

[1] Referring to the hypothesis due to Hebb (1949) that synapses are strengthened by conjoint activity in the presynaptic and the postsynaptic cells.

In rat cerebellum, the elimination of climbing fibre synapses onto Purkinje cells occurs during the second postnatal week, about the same time course as at the neuromuscular junction. In contrast, the elimination of preganglionic synapses onto neurons of the rat submandibular ganglion occurs over at least five postnatal weeks, far longer than is required for elimination at the neuromuscular junction.

The number of cochlear nerve synapses on neurons of the chick cochlear nucleus declines rapidly, from about four to two afferents and reaches a mature state even before hatching. Therefore, synapse elimination appears to be a widespread phenomenon, although there are no general rules about the percentage of afferents that is lost or the duration of time required.

There is a long and established history of the role of neural activity in the development and regeneration of nerve connections in the neuromuscular system. Tenotomy (cutting the tendon) delays the withdrawal of superinnervation by a moderate amount. Muscle paralysis results in increased, long-lasting levels of polyneuronal innervation, as does application of tetrodotoxin (TTX), a neural activity blocker, to the motor nerve. Chronic muscle stimulation accelerates the elimination of synapses during development.

Researchers can compare the effects of activity with inactivity by applying nerve blocking agents during the reinnervation of neuromuscular connections to a rat muscle with two separate branches of the motor nerve that can be manipulated independently. After cutting or crushing both nerve branches, motor axons regenerate, they superinnervate the muscle fibres and then gradually the pattern of single innervation is reestablished. By blocking one of the nerves with TTX during reinnervation, a competitive advantage can be given to the active synapses as against the inactive synapses (i.e., those made by the axons from the nerve blocked by TTX). It turns out that active synapses have an advantage over inactive synapses. For example, the ability of a regenerating nerve to regain its territory is enhanced if the other nerve is blocked and is diminished if its own nerve is blocked.

The final pattern of one contact per muscle fibre is both clear and unequivocal and several possible causes of the elimination process have been considered. Here is an assessment of them.

- It is unlikely that withdrawal of connections is random as this would leave many fibres uninnervated, contrary to observation.
- Nerve-cell death cannot provide the appropriate reduction in contacts as there is no cell death during this stage of development.
- Some terminals might withdraw if they were misdirected to the wrong region or the wrong fibre type. This is also unlikely as the muscles are almost homogeneous in fibre type and the somatotopic ordering of motor neurons across the muscle is very low.
- Another possibility is that synapses are preprogrammed to die. This is also unlikely as when a proportion of the motor neurons are removed before connections have been established, the surviving motor neurons make more contacts than normal.

This last point provides strong evidence that the identity of the surviving contacts depends on which other contacts are present; i.e., there is competition amongst synaptic contacts for survival.

There are various formal models for this competitive process. According to some models, there is competition for a fixed amount of synaptic strength possessed by the motor axons and shared amongst its terminals; in other models, the making of synapses is thought of in terms of the binding of neurotrophins onto the receptors on different axons, with competition amongst the receptors for the neurotrophins. How the various effects of activity can be interpreted is not yet clear. It is interesting that the mathematics underlying these models bears a strong family resemblance to the mathematics underlying many self-organizing phenomena studied in physics and the competitive interactions of the predator-prey type studied in population biology and other disciplines.

In summary, the production of the precise connection pattern of one contact per muscle fibre seems to be a task that is not fulfilled by the genome but instead is the responsibility of a competitive, 'self-organizing', mechanism involving interactions at the synaptic level.

3 THE ROLE OF SELF-ORGANIZATION IN EXPERIENTIAL CHANGE

I now examine situations where the putative self-organizing mechanisms are more strongly influenced by the signals impinging on the system from outside.

3.1 Feature Maps

In mammalian cerebral cortex, convergence of inputs onto neocortex causes cortical cells to respond to complex properties of the input. For example, in certain areas of the visual cortex, there are cells that are sensitive to the orientation of a stimulus or its direction of movement as well as its position in the visual field. Such attributes of the external environment can be detected by means of the neural circuitry and the connectivity of the central nervous system. The properties of the stimulus that produces maximal excitation in each small area changes over the cortical surface, defining 'feature maps'.

3.1.1 *Ocular Dominance, Orientation and Direction Selectivity*

Visual cortical cells receive innervation from both eyes, but with differences in the strength of the innervation from each eye that vary systematically across the surface of the cortex (Fig. 1.5). The Nobel laureates Hubel and Wiesel (Hubel *et al.*, 1977) discovered that cells in monkey binocular visual cortex vary in their responsiveness to the two eyes. Similar ocularity preferences extend down to layer IV of neocortex – one of the six layers in neocortex defined on anatomical criteria – where axons from the lateral geniculate nucleus terminate, and thus the concept of 'ocular dominance columns' arose.

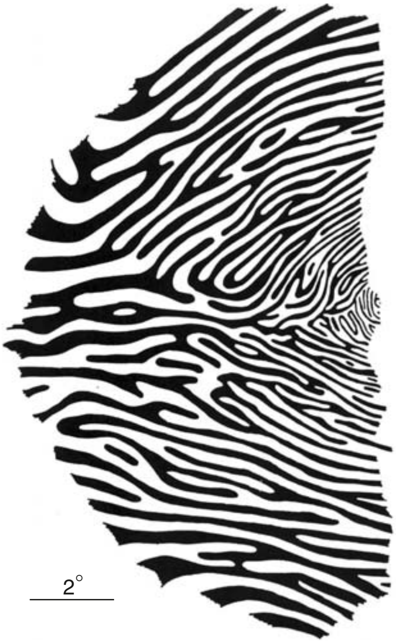


FIGURE 1.5 Computer reconstruction of the pattern of ocular dominance columns in layer IVc of area 17 of a macaque monkey, produced by reduced silver staining, translated into the visual field. (Reproduced from Hubel and Wiesel, 1977)

These systematic variations in ocular dominance are superimposed on the basic retinotopic map (Hubel *et al.*, 1977). Subsequently, existence of such columns was confirmed anatomically; the map of ocularity specificity across the entire surface of binocular cortex resembles a pattern of zebra stripes.

In cat and monkey, segregation begins at or around birth and is complete about six weeks later. Ocular dominance columns seem to be the result of a competitive process between the axons from the two eyes. They result from an initially overlapping distribution of innervation originating from the left

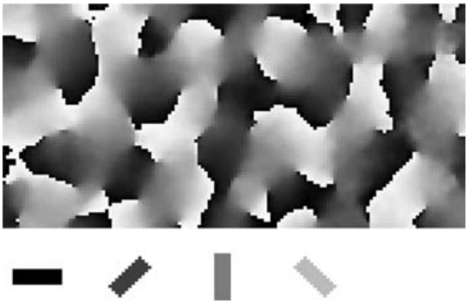


FIGURE 1.6 Orientation preference map for area 17 of the cat visual cortex, constructed by optical recording. The orientation at each point is indicated on a grey-scale. The key below shows how orientations are assigned on the scale. The length of the bars corresponds to 0.5 mm. This type of picture is best seen in colour (e.g., Hubener *et al.*, 1997).

and right eyes. In another preparation, a similar pattern of stripes can be produced through competition within the axons from a single eye. In *Xenoptis*, so-called 'compound eyes', constructed from two matching half eye rudiments, develop a single optic nerve which produces a striped projection on the optic tectum.

Cells in certain visual areas of mammalian neocortex are orientation selective; i.e., each cell is responsive to a small bar of light when presented in a particular orientation at a particular position in the visual field (Fig. 1.6). The existence of orientation maps was established by extracellular recording and, more recently, the method of optical recording has been used to produce detailed orientation maps over the entire surface of the visual cortex. The maps produced are complex and have a number of features, such as periodically repeating patterns, and more complicated features, such as saddle points and singularities (points on the cortex around which orientation domains are clustered in a pinwheel fashion). This type of data has provided an irresistible challenge to modellers.

The magnitude of the response from some cells elicited by a moving bar stimulus in a particular orientation, may depend on the direction of movement (at right-angles to the orientation of the bar). This forms a directionally selective map.

3.1.2 Relations Between the Different Feature Maps

The different types of map are interrelated. The ocularity map effectively interrupts the retinotopic map; i.e., if all the pieces of cortex innervated by one of the eyes were removed and the remaining pieces, innervated by the other eye, were pushed together, then a completely ordered retinotopic map would result. In cat visual cortex, pinwheel centres in orientation maps are mainly located in the middle of ocular dominance columns. In addition, according to recent optical recording experiments in cat, the orientation domains tend to intersect at right angles the borders of ocular dominance stripes. In the classic model of Hubel *et al.* (1977), developed for the cat, iso-orientation columns run in straight lines at right-angles to the inter digitating ocular dominance columns.

3.1.3 The Effect of Neural Activity

It is well established that lack of normal visual experience early in postnatal life prevents the normal development of the visual cortex. This suggests that some information required to generate the normal visual system derives from interactions of the developing system with the external environment. Clearly, many aspects of cortical development occur prenatally and are immune to the effects of postnatal functional deprivation.

Studies in both cats and primates have shown that before the neonate has received any visual experience, geniculocortical fibres find their main target. In layer IV, they converge to generate immature orientation selective cells clustered into rudimentary orientation columns, form a retinotopic map and, at least in the primate, begin to segregate to form ocular dominance columns.

Visual deprivation early in life, during the so-called critical period, does not abolish these features of early organization. The continued refinement of cortical connections is strongly influenced by patterned neural activity. Visual deprivation can have a devastating effect on the development of the detailed circuitry required for the normal functional properties of visual cortical neurons. For example, in areas 17 and 18 of the cat – these are the primary (18) and secondary (17) areas of cat visual cortex – the normal appearance of a large proportion of orientation selective cells is prevented by dark-rearing or by binocular eyelid suture.

A variety of results shows how experimental interference can affect the development of ocular dominance columns. Most of these experiments involve the manipulation of activity levels by: monocular deprivation; rearing animals in the dark; distorting the retinal input by artificially inducing strabismus, and removal of spontaneous retinal activity by administering the neural activity blocker TTX. The results of most of these experiments indicate the important role of neural activity in the formation of ocular dominance columns. Thus, deprivation prevents the emergence of the full richness of functional architecture and receptive field properties of the normal adult visual cortex.

The most striking effects of deprivation occur when vision through only one eye is impaired during the critical period. This results in expansion of the cortical territory of the projection serving the normal eye relative to that of the projection serving the deprived eye. Neuronal activity plays a crucial role in this organization, and it appears that monocular deprivation places the geniculocortical afferents from the deprived eye at a competitive disadvantage. The effects of monocular deprivation can be reversed by opening the deprived eye and closing the other before the end of the critical period (reverse suture). These changes involve the sprouting and/or trimming of geniculocortical arbors. It appears that, as the developing geniculocortical fibres elaborate on their initial framework of immature inputs, it is especially important that the projections serving one eye should be as active as those serving the other eye. A balance of activity is required to ensure that the growth of terminals from each eye is restricted within its own cortical territory.

3.1.4 Self-organization and the Formation of Feature Maps

Self-organization plays a role, in combination with external signals, in determining the response properties of the individual cells in feature maps as well as the pattern of response properties distributed over the map itself. In the development of ocularity maps, determination of the ocular preference of an individual cell is influenced heavily by the nature of the afferent activity, whereas the pattern of ocular dominance is the result of an interaction between activity and mechanisms of self-organization.

In the development of orientation maps, self-organization may be involved in specifying both single cell properties and the overall pattern. Von der Malsburg (1973) published the first paper demonstrating that the pattern of orientation specificity in visual cortex could be developed in a self-organizing model under the instruction of simulated orientated bar stimuli. Since then, research has established that patterned stimuli are not needed to develop the individual cell's response properties. Radially symmetric patterns of activity drive the production of orientated receptive fields by a process of symmetry breaking. There has been much recent theoretical work developing models of all types of feature map and the relation between them. In these models, changes in the external conditions (simulating, for example visual deprivation) lead to the model successfully self-organizing to adapt to the new conditions.

3.2 Self-organization and the Acquisition of Cognitive Function

All the examples of self-organization in the nervous system we have discussed so far concerned the development of the nervous system in cases where we can relate the results of the self-organization to structural and functional changes at the cellular and subcellular level. I now discuss self-organization and the acquisition of cognitive function.

Perhaps because of the lack of a strong structural basis, less can be said about the precise form of the self-organizing mechanisms than in the examples of neural self-organization discussed already. It will be seen that many of the themes already discussed are echoed, but with different vocabulary to place an emphasis on the cognitive and psychological levels.

3.2.1 Cognitive Self-organization

The term self-organization has been used to understand how new cognitive behaviours are acquired. As noted earlier, these behaviours could be derived from built-in knowledge (in cognitive science terms, nativism, attributed in this context to Fodor) or through learning (attributed in this context to Piaget). However, it has been pointed out that through self-organization, order can emerge from interactions within the system rather than by explicit instruction.

According to Karmiloff-Smith (1992), children's brains are genetically prewired to contain modules (or domains) of knowledge which grow and interact during development. She proposes that in each domain, the child's development is subject to constraints that initially shape the way that information in different domains is processed and represented. The child's initial knowledge in each domain is then progressively redescribed according to existing knowledge and the child's experiences. The level of redescription in each domain thus depends on a complex interaction between the current state of knowledge in the domain and the child's experience. Along with mastering new behaviours, the child also learns to introspect about what he/she has done and ultimately constructs his/her own theories about how the world works.

While accounting for a large body of knowledge, these theories of re-representation are difficult to interpret insofar that no details are given about the mechanistic basis of re-representation. An attempt to make a bridge between the cognitive and neural levels of brain function has been supplied under the name of neural constructivism. This has been developed from Piaget's constructivism, a term that reflects his view that there is an active interaction between the developing system and the environment in which it develops. Neural constructivism emphasizes the dynamic interaction between the developing system and the developmental 'task' to be accomplished and in this respect it can be regarded as a form of self-organization.

Neural constructivism lies between the extreme versions of the theories of chemospecificity (the entire blueprint of the nervous system is specified in the genome) and a tabula rasa theory (nothing is prespecified). It has been contrasted with the doctrine of selectionism which, according to some authors, is the idea that neural development proceeds by initial overproduction of neural structure followed by the selective pruning away of the inappropriate parts. We can view selectionism as an extreme form of the two-stage process thought to underlie the development of connectivity, mentioned in Section 2.3. According to selectionism, the first stage is involved in the formation of connections and the second stage in the breaking of connections.

Neural constructivism is concerned with how neural activity guides the development of connections, the patterns of activity being generated from the external environment rather than through, for example, spontaneous activity. It is suggested that, as neocortex evolved in mammals, there was a progression toward more flexible representational structures, rather than there being an increase in the number and complexity of innate, specialized circuits. Different types of evidence are cited in favour of neural constructivism, some of which were reviewed earlier in this chapter:

- *Changes in synaptic number* During development in primates the number of synapses rises and eventually falls but there is no decrease until post-puberty.

- *Axonal growth* In the visual system during the period of activity-dependent development there is evidence for axonal growth, contrary to the doctrine of selectionism.
- *Dendritic* growth Dendrites grow at a much slower rate than axons, over a longer time scale which stretches over the period of time prescribed by selectionism. Dendritic morphology can be moulded by the patterned neural activity incident upon it.
- *The extent of cortical development* Human cortical postnatal development is also more extensive and protracted than generally supposed, suggesting that the neocortex has evolved so as to maximize the capacity of environmental structure to shape its structure and function through constructive learning.
- *The power of constructive neural networks* Finally, arguments based on the computational power of artificial neural networks, used to simulate the phenomena observed, are used to support neural constructivism, particularly the view that networks that develop their structure whilst they are being trained are more powerful than fixed architecture neural networks.

3.2.2 Neural Constructivism and Neural Networks

Neural networks are collections of highly interconnected simple computing units modelled loosely on the architecture of the brain. In such a system, a computing unit is a stylized representation of a nerve cell. The networks are required to learn specific input/output relationships (Hertz *et al.*, 1991; Bishop, 1995) by selective adjustment of connection strengths between neurons. Most applications of neutral networks are to problems in pattern recognition, classification and prediction. Their behaviour is usually investigated through computer simulation.

Neural networks can be trained by two main methods. In supervised learning, many examples of the input/output pairings to be learnt are presented to the network, in the forms of patterns of activity over the computing units, which can be likened to patterns of neural activity. As a result, the strengths of the connections (weights) between individual computing units changes. Ultimately, the weights in the network become set so that the network, when tested, gives the required output for any input presented during learning. Hopefully, it generalizes its behaviour to give the appropriate response to an input that it has not seen before.

In unsupervised learning, there is no teaching signal, in the form of the required output being supplied for each input. Presentation of every input generates an output and the network modifies its weight strengths 'by itself', with the result that in the initial stages the output generated for each input may change. After many presentations, a stable output comes to be associated with each input.

Neural networks have been said to be self-organizing in that, in both learning paradigms, learning depends critically on the structure of the network and the interactions between computing units. In supervised learning, both the pattern of weight strengths that emerge in learning a given mapping, and the ability of the network to respond to novel inputs, is self-organized by the network itself. In unsupervised learning, the nature of the input/output mapping produced depends on the interaction between network structure and the nature of the input patterns.

How a given task is learnt, or whether it will be learnt at all, depends on the structure of the network. Using recently developed methods, the structure of the neural network can be built up whilst the task is being learnt, rather than training a network with a fixed architecture. The network structure becomes tailored to the specific problem and in many cases it is claimed to yield better performance than networks with a fixed architecture. By analogy, neural constructivism describes the idea that the development of structures such as the neocortex involves a dynamic interaction between the mechanisms of growth and environmentally driven neural activity. How information is represented in the cortex depends on the particular computational task being learnt.

Neural network models have been used to simulate human performance on specific cognitive tasks. This area of research is known as connectionism. Typically, the modeller prescribes a specific network structure. The strengths (weights) of the connections between 'neurons' are chosen randomly, so that the model does not contain any preprogrammed knowledge. One early and influential example of the use of neural network models in this context was Rumelhart and McLelland's model of past tense learning (Rumelhart and McClelland, 1986a). They showed that a neural network could learn the mapping between the stem form of a verb and the actual past tense without the need for the rule that prescribes such transformations to be specified externally. In addition, the U-shaped pattern of learning characteristic of a child's learning is produced automatically by application of this model. Whilst this model has been subject to much criticism, its development and application did demonstrate that there are alternative ways to viewing language acquisition than the application of preprogrammed rules.

4 SELF-ORGANIZATION AS A RESPONSE TO DAMAGE

It has long been known that the effects of brain injury early in life are much less severe than similar effects in the adult. Until relatively recently, the commonly held view has been that this is because the brain has a capacity for plasticity during development that can be brought into play as a response to injury. It is now clear that the adult brain has a much higher capability for reorganization than previously thought. Recent experimental studies illustrate that the adult mammalian nervous system does have substantial capacity to reorganize itself functionally. Plasticity should be thought as a property of both developing and adult systems.

4.1 Self-reorganization

A landmark study was made by Raisman in the early 1970s on experiments in rats. This involved partial denervation of the septal nuclei, which are part of the limbic system that includes the hippocampal formation and the amygdala. The lateral septum has two main inputs. When either one of these was cut, there was a large temporary reduction in the number of synapses on the septal cells. Over the following two weeks, the number of synapses returned to normal levels but all the synapses were now the type characteristic of the intact input. The same result was obtained whichever septal input was cut. This study is important as it provided the first solid evidence of plasticity at the cellular level in the adult mammalian central nervous system.

Since that time, there have been several studies showing substantial plasticity in mammalian neocortex. Many sensory and motor modalities have multiple representations on the neocortex. Originally it was thought that these maps would be relatively permanent once they have formed. However, it is now established that the neocortex retains a large degree of plasticity throughout adult life.

In the 1980s, Merzenich and colleagues found in monkeys that, following peripheral nerve injury, the ordered mapping of the body surface onto primary somatosensory cortex becomes reorganized substantially. Electrophysiological recordings were made to examine the responsiveness of the areas that had responded to stimulation of the cut sensory nerve. When recordings were made soon after surgery, it was found that a large part of each of these areas was responsive to cells from neighbouring areas of skin, which normally projected to neighbouring cortical areas. If the area was large, there was a region in the middle from which no response to sensory stimulation could be obtained. Over the next month or two, the cortical representation of nearby somatic areas gradually expanded into the unresponsive area until the entire somatosensory region of the cortex responded to sensory stimulation, making a map that was a reorganized and distorted version of the original one.

Another manipulation that has been tried is removing part of the body instead of denervating it. For example, removal of whole digits caused the affected area of cortex to come under the control of the digits from either side of the ablated ones. Most likely, this reorganization involved two different mechanisms. The gradual spreading of the somatosensory projection seen in the long term is likely to involve an anatomical rearrangement involving the regrowth of connections. However, this type of change is relatively slow and so cannot also account for the changes seen after surgery. The most likely explanation is that surgery triggers an unmasking of a population of silent synapses, agreeing with inferences drawn from earlier experiments on the cat spinal cord.

Similar effects have been found in both primary auditory and visual cortex in monkey. The auditory cortex contains a one-dimensional map of frequency, with high frequency tones represented in caudal regions and low frequencies more rostrally. Destruction of nerve fibres in the cochlea which are responsive to high frequencies caused, a few months later, a reorganization of this tonotopic map with low frequencies being represented rostrally and mid-range frequencies more caudally. In primary visual cortex, small retinal lesions initially produce an area of unresponsive visual cortex. Over the next few months, this cortical region gradually acquires new receptive fields from places in the retina near the site of the lesion.

These effects are also seen in humans. In the 1990s, Ramachandran and colleagues carried out a series of studies on sensory reorganization following limb amputation. On examination a few weeks after amputation of an arm, patients reported sensations from their phantom limb which were referred to regions of the face and the intact arm. In several cases it was possible to define fields of sensation across regions of the face in which the normal somatotopic organization of the digits of the hand was maintained. In other cases these same types of map were reported in the region of the operated arm above the level of the amputation.

The most likely explanation of these results is again that the sensory fields in cortex that were adjacent to the part of somatosensory cortex that has suffered deafferentation had invaded the deafferented region. It might be noted that in the normal somatosensory map, representation of the face is near to representations of the hand and the arm.

4.2 Can the Nervous System Regenerate After All?

The commonly held view is that damaged axons in the mammalian nervous system will regenerate in certain cases but nerve cells will not. Damage to major axon tracts or large areas of nervous tissue leads to permanent loss of function at the neuronal level as neither damaged nor killed axons will regenerate. In the mammalian peripheral nervous system, limited repair is possible as axons can regenerate, leading to the restoration of functional connections with other nerve cells and muscle fibres. In contrast, invertebrates and non-mammalian vertebrates have the capacity to regenerate axons and thereby nerve connections throughout their nervous system.

The view that nerve cells cannot regenerate is being challenged. Recent research shows that an important class of cells, called stem cells, exists in many parts of the adult mammalian nervous system. These cells can differentiate into all types of cell, including nerve cells.

Stem cells have been found in the dentate gyrus of the hippocampus and in the olfactory lobe. It may be that nerve cells can be generated from these cells, even during adult life, possibly to replace damaged nerve cells. One series of experiments observed continuous formation of cells (not identified as nerve cells) in adult mouse neocortex. Targeted destruction of nerve cells in a small region of neocortex then led to a population of new cells being generated, a small proportion being nerve cells. These new nerve cells formed connections with other neural structures, suggesting that indeed they were substituting for the set of destroyed cells.

Research into the potentialities of stem cells is a current hot topic and new results are reported frequently. For example, a recent paper (Mezey *et al.*, 2003) reports post mortem analysis of humans who had received bone marrow transplants as treatment for leukaemia. In the four patients studied, there was evidence for newly generated nerve cells in the brain, which had derived from the bone marrow transplants. These findings are still preliminary and controversial but may yet change the view that individual nerve cells cannot regenerate.

5 OPEN QUESTIONS

In attempting to understand a complex system such as the nervous system, it is important to identify important general principles of operation. 'Self-organization' is one such principle. It manifests itself during both development and the functioning of the nervous system. This section suggests key questions for activities that could arise out of the work reviewed here, both for the furtherance of research into neuroscience and for the construction of new types of computational ('cognitive') systems. It is worth restating the obvious point that these two activities have different goals. One is concerned with understanding a given system: the other is concerned with designing a system to achieve a particular computational task, where the nature of the internal workings of the system are dictated by the task to be accomplished rather than having to reflect any biological plausibility.

5.1 Questions for the Neurosciences

The term self-organization refers to a postulated mechanism rather than a collection of phenomena such as those embraced by, for example, perception or learning and memory. The nature and scope of other examples of self-organization remain to be determined. Therefore, questions such as 'What is the future for research into neural self-organization?' are premature. Instead I focus on those aspects of the life sciences which are related to the areas of research described in this chapter.

5.1.1 Levels of Analysis

This chapter describes how self-organization operates at the synaptic, cellular and network levels. At what other levels may self-organization apply?

- *The cognitive level* The role of self-organization within developmental cognition was described briefly in section 3.2. To make sense of self-organization at the cognitive level, it is important to be able to identify the nature of the elements that do self-organize. As this type of self-organization involves extensive regions of the brain, this will require the use of powerful methods for assaying whole brain activity to identify these elements. As mentioned already, modelling is a crucial experimental tool here. To apply computer modelling successfully, models must contain the correct degree of neurobiological realism.
- *The subcellular level* Although we now have at our disposal several completely sequenced genomes, we are still for the most part remarkably ignorant about how genes interact and regulate each other to develop the nervous system. A plausible picture of cellular regulation would involve networks of multiple interactions, with feedback between all layers. There is great scope for exploiting the parallels between maps of metabolic pathways and those of gene expression, both of which involve processes of global control by self-organization. More widely, the similarities between the patterns of connectivity over ecological, neural and biochemical networks are being compared (Lee *et al.*, 2002; Milo *et al.*, 2002).

Arguably, perhaps the only system of many interacting elements that has been characterized in detail is the artificial neural network. There must be a parallel, so far unexploited, between neural networks and subcellular networks.

- *Crossing the levels* Quite often, testing a postulated mechanism requires observation of phenomena at one level and the identification of mechanism at lower levels. More links between levels are required to achieve this. For example, there is still no solid evidence linking the nature of the nerve connections making up the topographically ordered maps and the underlying molecular mechanisms. This will require a cross-disciplinary approach. In this case, what is needed is: experimental evidence of the nature of the connections made at both the electrophysiological and anatomical levels; evidence for the distribution of signalling molecules or patterns of neural activity amongst nerve cells that carry the information enabling the correct connections to be made; and an explanatory framework to link the two levels. New imaging methods that obtain information at many different levels, particularly the protein, synaptic, cellular, network and whole brain levels will be crucial.

5.1.2 *The Use of Mathematical and Computer Models*

Often the consequence of any given hypothesis involving a large number of interacting elements can only be obtained by constructing and analysing the properties of computer and mathematical models. This approach is now recognized within neuroscience as an important means of formalizing ideas and concepts and testing them out for their self-consistency and against the large amount of neuroscientific data that is becoming available, and forms part of the field of neuroinformatics. Many areas of neuroscience described in this chapter have profited from the application of computer and mathematical models.

- *Neural simulators* There is an increasing trend towards standard powerful neural simulators that will enable researchers to share data, models and results.
- *New types of model* Completely new types of model will be required in some cases, such as those required to model the regeneration of nerve cells (if indeed this occurs) as described in Section 4.2.
- *Modelling of real nervous systems* The power of the present generation of computers is now sufficient to enable simulation of the complex geometry of the nervous system, which is an important constraint on function. For example, up till now, nerve cells have been often modelled as abstract entities rather than occupying particular positions in a complex three-dimensional environment.

5.1.3 *Evidence from Invertebrates*

This chapter has been restricted almost entirely to vertebrates, with very little discussion about the organization of invertebrate nervous systems. None the less, two general comments can be made.

- Lessons to be learnt from the evidence: Observation of the same type of phenomena in invertebrates may suggest completely different types of mechanism. For example, much of the evidence at the synaptic and cellular level suggests a very precise and inflexible organization. For example, in the neuromuscular system of the fruit fly, *Drosophila*, there is a precise and fixed relation between a motor neuron and the muscle fibre that it innervates. There are various possible reasons for this. In this case, for example, it may be that a more subtle form of self-organization operates; or it may be that for the smaller nervous system the genome can afford to specify precisely all the parameters values needed which have a smaller number of neurons. Knowledge of invertebrate nervous systems can provide another source of inspiration for the physical sciences.
- A complete explanation: To have a complete knowledge of how the underlying

neural substrate generates and controls the animal's behaviour it is necessary to understand it at many different levels, from genes through proteins, synapses, cells, networks to behaviour. In the chapter, emphasis has been on the properties of mammals and other vertebrates. However, it is more likely that the first complete understanding will be obtained for an invertebrate system. One prime example is *Drosophila*. It has a known genome, containing around 15 000 genes, with a variety of complex behaviours including those involving learning and memory. In addition, newly developed imaging techniques can be applied that enable visualization of nervous system activity at the subsynaptic, synaptic and cellular and network levels.

5.2 Inspiration for Other Sciences – 'Cognitive Systems'

Replication or inspiration? Whereas living and artificial systems are made out of the same basic set of elements, namely atoms, clearly their structures are different – silicon chips are different from pieces of brain. The nature of the substrate constrains the properties of the system. This means that whereas it may be possible to build a system that mimics the input/output relations of a living system, at the mechanistic level the systems will be different and will operate in different ways. As a consequence, results from neuroscience can at best only inspire the construction of new types of computing device rather than lead to the construction of exact replicas of living systems. To take a well-known example, neuroscience has inspired the growth of the field of artificial neural networks. This can offer many interesting ideas for the construction of new computing devices themselves rather than give a precise blueprint for such a system.

Three inspirations:

- As emphasized throughout this chapter, the mathematics developed to describe neural self-organization, itself influenced from studies of physical systems, has wide application. Applications of self-organization to social systems and to economic systems are those that have not been mentioned so far.
- Using principles of self-organization to solve problems of coordination among autonomously interacting agents, such as those that occur within e-communities, is an obvious specific application.
- Neural self-organization will serve as an inspiration to self-repair technology as an example of application to a hardware problem.

Acknowledgments

My very grateful thanks to Fiona Jamieson, Stephen Eglen, Kit Longden, David Price and Peter Dayan for reading and commenting on this chapter. Any errors and omissions are mine.

Further Reading

This chapter draws on a large body of literature. Instead of providing a long list of papers, I give below a number of texts that cover most but probably not all of the work I have discussed, together with descriptions of their various scopes. In addition, a few important papers are included which are referenced in the text.

Self-organization There is no book specifically about self-organization and the nervous system. Kauffman (1993) discusses theoretical aspects of self-organization in evolution and Camazine *et al.* (2003) is a recent book discussing self-organization in biological communities, typical examples being aggregates of bacteria, communities of fireflies, fish and ants.

Mathematical basis The book by Murray (1993) is a classic, concentrating on the mathematical basis of theoretical developmental biology; Edelstein-Keshet (1987) is a very good alternative.

Development Alberts *et al.* (1994) is a text at the molecular biology level and Wolpert

(1991) is a readable introduction to embryonic development for non-specialists.

Nervous system Nicholls *et al.* (2001) is a classic treatise on the nervous system, which was first published in 1977 and has undergone continual revision since then. Shepherd (1994) and Bear (2001) are recent research level text books.

Development of the nervous system Purves and Lichtman (1985), Sanes *et al.* (2000) and Brown *et al.* (2001) are research level texts on the development of the nervous system, the latter two being more recent. The recent research monograph by Price and Willshaw (2000) describes genetic, molecular, systems and modelling approaches to understanding neocortical development. Elman (1996) takes a connectionist approach to development.

Neural networks Excellent texts amongst the plethora available are those by Hertz *et al.* (1991) and Bishop (1995). Rumelhart and McClelland (1986a,b) are two collections of classic papers. Arbib (2003) is an encyclopaedic collection of short papers written by experts on many different aspects of theories of brain function, connectionism and artificial neural networks. As the name suggests, the treatment of the nervous system is generally at the cellular and network level.

Brain damage and repair Fawcett *et al.* (2001) is a collection of papers reviewing recent research in this field.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) *Molecular Biology of the Cell.* New York: Garland Publishing.

Arbib, M.A. (ed.) (2003) *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press.

Bear, M.F. (2001) *Neuroscience: Exploring the Brain*, 2nd edn. Baltimore, MD: Lippincott Williams and Wilkins.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford: The Clarendon Press.

Brodmann, K. (1909) *Vergleichende Lokalisationslehre der Groshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: J.A. Barth.

Brown, M., Keynes, R. and Lumsden, A. (2001) *The Developing Brain*. Oxford: Oxford University Press.

Camazine, S., Deneuborg, J.-L., Franks, N.R., Sneyd, J., Theraulaz, G. and Bonebeau, E. (2003) *Self-Organization in Biological Systems*. Princeton, NJ: Princeton University Press.

Edelstein-Keshet, L. (1987) *Mathematical Models in Biology*. Boston, McGraw Hill.

Elman, J.L. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fawcett, J.W., Rosser, A.E. and Dunnet, S.B. (2001) *Brain Damage, Brain Repair*. Oxford: Oxford University Press.

Hebb, D. (1949) *The Organization of Behavior.* New York: Wiley.

Hertz, J., Krogh, A. and Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation.* Reading, MA: Addison–Wesley.

Hubel, D.H., Wiesel, T.N. and LeVay, S. (1977) Plasticity of ocular dominance columns in monkey striate cortex. *Phil. Trans. R. Soc. Lond. Ser. B*, 278: 377–409.

Hubel, D.H. and Wiesel, T.N. (1977) Ferrier lecture. Functional architecture of the macaque monkey visual cortex. *Proc. R. Soc. London B*, 198: 1–59.

Hübener, M., Shoham, D., Grinvald, A. and Bonhöffer, T. (1997) Spatial relationships among three columnar systems in cat area 17. *J-Neurosci.*, 17: 9270–9284.

Jacobson, M. (1967) Retinal ganglion cells: specification of central connections in larval *Xenopus laevis*. *Science*, 155: 1106–1108.

Karmiloff-Smith, A. (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science.* Cambridge, MA: MIT Press.

Kauffman, S.A. (1993) *The Origins of Order*. Oxford: Oxford University Press.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T. *et al.* (2002) Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298: 799–804.

Meinhardt, H. (1982) *Models of Biological Pattern Formation*. New York: Academic Press.

Merzenich, M.M., Kaas, J.H., Wall, J.T., Nelson, R.J, Sur, M. and Felleman, D.J. (1983) Topographic reorganisation of somatosensory cortical areas 3B and 1 in adult monkeys following restricted deafferentation. *Neuroscience*, 8: 33–55.

Mezey, E., Key, S., Vogelsang, G., Szalayova, I., Lange, G. and Crain, B. (2003) Transplanted bone marrow generates new neurons in human brains. *Proc. Natl Acad. Sci. USA*, 100: 1364–1369.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, 298: 824–827.

Murray, J.D. (1993) *Mathematical Biology*, 2nd edn. Berlin: Springer-Verlag.

Nicholls, J.G., Fuchs, P.A., Martin, A.R. and Wallace, B.G. (2001) *From Neuron to Brain*, 4th edn. Sunderland, MA: Sinauer Associates.

Price, D.J. and Willshaw, D.J. (2000) *Mechanisms of Cortical Development*. Oxford: Oxford University Press.

Purves, D. and Lichtman, J.W. (1985) *Principles of Neural Development.* Sunderland, MA: Sinauer Associates.

Ramachandran, V.S., Rogers-Ramachandran, D. and Stewart, M. (1992) Perceptual correlates of massive cortical reorganisation. *Science*, 258: 1159–1160.

Rasmussen, C.E. and Willshaw, D. (1993) Presynaptic and postsynaptic competition in models for the development of neuromuscular connections. *Biol. Cybern.*, 68: 409–419.

Rumelhart, D.E. and McClelland, J.L. (1986a) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D.E. and McClelland, J.L. (1986b) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models.* Cambridge, MA: MIT Press.

Sanes, D.H., Reh, T.A. and Harris, W.A. (2000) *Development of the Nervous System.* New York: Academic Press.

Shepherd, G.M. (1994) *Neurobiology*, 3rd edn. Oxford: Oxford University Press.

Spemann, H. (1938) *Embryonic Development and Induction*. New Haven, CT: Yale University Press. (Reprinted, New York: Garland, 1988.)

Sperry, R.W. (1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc. Natl Acad. Sci. USA*, 50: 703–710.

Suga, N. (1978) Specialization of the auditory system for reception and processing of species-specific sounds. *Fed. Proc.*, 37: 2342–2354.

Sur, M., Garraghty, P.E. and Roe, A.W. (1988) Experimentally induced visual projections in auditory thalamus and cortex. *Science*, 242: 1434–1441.

Tootell, R.B.H., Silverman, M.S., Switkes, E. and De Valois, R.L. (1982) Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, 218: 902–904.

Turing, A.M. (1952) The chemical basis of morphogenesis. *Phil. Trans. R. Soc. Lond. Ser. B*, 237: 37–72.

Von der Malsburg, C. (1973) Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14: 85–100.

Willshaw, D.J. and von der Malsburg, C. (1976) How patterned neural connexions can be set up by self-organization. *Proc. R. Soc. Lond. B*, 194: 431–445.

Willshaw, D.J. and von der Malsburg, C. (1979) A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem. *Proc. R. Soc. Lond. B*, 287: 203–234.

Wolpert, L. (1969) Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.*, 25: 1–47.

Wolpert, L. (1991) *The Triumph of the Embryo.* Oxford: Oxford University Press.

Woolsey, T.A. and van der Loos, H. (1970) The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Res.*, 17: 205–242.

CHAPTER

2

Large-scale, Small-scale Systems

Jim Austin, Dave Cliff, Robert Ghanea-Hercock and Andy Wright*

1 INTRODUCTION

1.1 State of the Art

We present a selective review of research in computer science that is directed toward engineering artificial complex adaptive systems. Such systems typically consist of many individually simple components. These interact in relatively simple, but non-linear, ways. In these systems, compounded

* Authors are listed in alphabetical order. Authorship of specific sections is indicated in the text.

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

non-linearities in interactions between components can give rise to large-scale behaviours at the system level that are difficult or impossible to predict, even when the individual components and their small-scale local interactions are well understood. Many biological systems, including the nervous systems of animals, are complex adaptive systems. So we can see that research into artificial complex adaptive systems can be relevant to the study of cognitive systems.

Continuing falls in the real costs of computing over the past two decades have increased activity in the exploration and development of computational techniques and architectures that draw inspiration from biological complex adaptive systems. Thus, this chapter focuses on the relevance to the study of cognitive systems of biologically inspired complex adaptive systems (BICAS) approaches within computer science.

Examples of BICAS techniques in computer science include:

- use of biological metaphors from embryology/morphogenesis, and from Darwinian evolution, in the automated design and layout of electronic circuits
- creation of artificial immune systems for computer network security
- development of artificial neural networks that can learn distributed representations
- development of animal-like artificial autonomous agents, either as physical mobile robots for use in real-world scenarios or as virtual entities for use purely in software applications.

While we can consider a single artificial autonomous agent as a complex adaptive system in itself, at another level a single agent may be one member of a team of autonomous agents, acting cooperatively as a multi-agent system, or 'super-organism' to achieve some overall goal. In many applications of genuine and significant interest, the environment within which the system is to operate, that is the agent's 'niche', is uncertain, dynamically changing and noisy. Thus much research in BICAS is directed at creating adaptable systems, often using techniques inspired by habituation and learning in neural systems and by evolutionary adaptation in populations of organisms.

We discuss three related areas: the first two address issues of the underlying computer architectures on which future artificial cognitive systems may be based; the third explores current attempts to create complete creature-like cognitive autonomous agents.

- *Circuit-level computation* We consider the construction of ultra-massive parallel processors (UMPP) that may provide future generations of processing architectures. Although such systems exist in nature, in animal nervous systems for example, the construction of even relatively simple systems remains a challenge. Such processors should have many desirable attributes, including the abilities to self-organize, to adapt to changing environments and to self-repair.
- *Distributed and networked computing* Increased bandwidth and cheap and well-packaged computing have enabled the production of large-scale computing networks. Again the ability to self-organize, to adapt and to respond to changing demands and environments where, importantly, the control of the system is at the *local* level (and so decentralized) remains a core challenge in engineering such systems. Ideas derived from the study of social animals and collective organisms have a role to play in the derivation of novel 'bottom up' design methods, as does the development of machine learning ideas for multi-agent systems. It is envisaged that these will lead to the development of decentralized systems in large-scale computing networks (e.g. the Grid), telecommunications and defence.
- *Artificial autonomous agents* We examine BICAS approaches to the design of software agents and autonomous mobile robots. In both types of agent, there is a need for sensory-motor 'controller'

architectures that robustly and reliably coordinate perceptions and actions in the pursuit of that agent's goals, or of the goals of that agent's community or team. Inspiration and metaphors from biology have had a strong influence at many levels in the development of biologically inspired autonomous agents. Applications of such systems include areas as diverse as space exploration, telecommunications and computer games.

1.2 Key Open Questions

Each domain we discuss in this chapter reveals a number of key questions that need to be answered for research in BICAS to fulfil its potential. From these issues, we have derived three substantive open questions.

Development of a coherent theory Research in BICAS has developed independently in a number of related directions, yet in a number of somewhat disparate fields. The area is highly multidisciplinary. It draws on research in topics as diverse as evolutionary biology and animal behaviour through to control theory and machine learning. It requires mathematical tools from disciplines such as game theory and dynamical systems theory. These diverse disciplines do not always have a long tradition of interaction. The development of a coherent theoretical framework is needed to take effective advantage of the synergies between disciplines. Although there are frameworks, these are often specific in nature and do not draw the different fields together fully.

There is an open question around the nature of a general framework that will allow the design and production of biologically inspired complex adaptive systems that are adaptable, in predictable ways, and that can cope with the inherent uncertainty of the real world.

Development of robust adaptation methods Many biological complex adaptive systems can self-organize, self-regulate and self-heal. This adaptability makes them better-suited to their environment, or 'environmental niche', even when that environment is dynamically varying and/or unknown in advance. Computer scientists have not yet developed appropriate machine learning or adaptation techniques that could provide similar adaptability in artificial systems.

The derivation of appropriate, practical and well-understood methods of adaptation or machine-learning remains an open issue. This question needs to be addressed not only from a theoretical point of view but also through the empirical design and testing of such methods.

Demonstration of design practicality The practicality of the 'bottom up' design methodology we describe relies on two assumptions: first, that it can produce systems and collections of agents that can function to a desired level of competence; second, that the resulting systems are robust to the inherent disturbances and changes any real system will suffer. Thus, this bottom-up approach will have to answer the questions of *robustness, reliability, flexibility* and *scalability*. This is necessary to overcome the natural conservatism of the designer with a traditional 'top-down' view. Convincing demonstration of these new methods on industrial-scale safety-critical systems has not yet taken place.

The demonstration of BICAS approaches in applications where there are safety critical reliability issues – such as financial systems, telecommunications networks, defence applications – has yet to occur.

2 INTRODUCTION

Dave Cliff and Andy Wright

2.1 Biology in Cognitive Systems: Science and Engineering

Broadly speaking, the study of cognitive systems can serve the purposes of science or of engineering. This is a very coarse division but it can help to distinguish between the two purposes. A scientific study seeks to establish novel and falsifiable statements about naturally occurring cognitive systems, or perhaps about the space of *possible* naturally occurring cognitive systems. An engineering approach seeks simply to create artificial systems that reliably exhibit some desired level of cognitive performance or behaviour. Of course, the two approaches are not mutually exclusive, and indeed can inform each other.

As all known naturally occurring cognitive systems are biological in origin, scientific studies of cognitive systems ignore biological data at their peril. For engineering, however, there is no *a priori* requirement to pay attention to biology.

A significant element of research directed at the engineering approach to creating cognitive systems took place under the banner of artificial intelligence, a field strongly, but not uniquely, identified with computer science. For much of its history, research in artificial intelligence (AI) largely ignored biology.

From the outset, commonly identified as the 1956 Dartmouth Conference on AI (McCarthy *et al.*, 1955), up until the mid-1980s, the focus within AI research was almost exclusively on treating cognition as a process involving the manipulation of symbolic representations of facts or knowledge, using techniques inspired by mathematical logic and/or computer programming languages. This approach eclipsed other, biologically inspired, approaches, such as early studies in cybernetics that treated networks of neurons as logical devices (McCulloch and Pitts, 1943), Rosenblatt's 'perceptron' (1962) and the automated learning work of Widrow and Hoff (1960). We can trace the demonstration of minimal biologically inspired architectures for mobile robot controllers back to the cybernetics research of Walter's (1950, 1951) 'turtle' robots and Ashby's (1952) 'homeostat'. We can trace attempts to synthesize lifelike phenomena in mechanical automata back to the mid-1700s (for a historical review see Langton, 1989).

The 1980s, however, saw the renaissance of these biologically inspired ideas with:

- renewed interest in artificial neural networks and parallel distributed processing (PDP) architectures, inspired by the nervous systems of animals
- the development of behaviour-based autonomous agent systems, inspired in part by studies in ethology and behavioural ecology
- and the rapid growth of research interest in 'artificial life', seeking to engineer artificial systems that draw inspiration from a diverse range of natural systems, including developing embryos, the human immune system, evolving gene-pools and interacting groups of autonomous agents.

We explore these topics in more depth later in this chapter. However, what is common to all these new approaches is the observation that many naturally occurring systems can, at one level of analysis, be described as being built from components that are individually 'simple' and that interact with each other in relatively 'simple' ways. Yet at another level they exhibit some 'complex' overall behaviour that is not readily predictable from the individual components. Typically, the complex overall behaviour is the result of compounded non-linearities in the component interactions. Furthermore, many systems exhibit sophisticated adaptation responses over multiple timescales, and are resilient with respect to variations in component connectivity – often as a result of properties of self-organization. Almost all such naturally occurring systems are biological in origin. We use the phrase 'biologically inspired complex adaptive systems' (BICAS) to refer generically to all such artificial systems.[1] Some examples include:

- We can describe an individual nerve-cell, a neuron, at one level of analysis as a

[1] We note that a number of significant complex adaptive systems are non-biological: for example, simulated annealing (Kirkpatrick *et al.*, 1983) is a computerized optimization technique inspired by the cooling of paramagnetic materials. Non-biological natural complex systems are implicitly included here as we would not want to exclude them purely because they happen not to be biological in origin.

simple component that integrates electrical impulses received on its inputs and, if the sum of received impulses is sufficiently high, generates an output impulse. This sounds simple. Yet connect enough neurons together in the right way, and expose this tangle of neurons to the right environment for long enough, and the end-result could be an adult human brain that can think, learn and act.

- We can view an individual animal as a simple self-interested vehicle for propagating its genes. It just has to survive long enough to find a good mate and produce viable offspring, and its work is done. Yet, over a sufficiently long period, random genetic variation combined with evolutionary processes such as Darwinian survival-of-the-fittest selection can create 'designs' of animals that are exquisitely well-tailored to those animals' environmental niches.
- An individual trader in a marketplace can be viewed as a simple self-interested agent. Sellers try to trade at the highest price possible, while buyers try to trade at the lowest possible price. Yet, in the right conditions, this conflict between groups of traders acting out of naked self-interest can collectively form a market where transaction prices rapidly and repeatedly settle to the theoretical equilibrium price. This is the price that yields the most efficient allocation of scarce resources within the market. The traders in the market can do this without the presence of a central coordinating or synchronizing 'auctioneer', despite dynamic variations in the underlying supply and demand in the market.

As the last example makes clear, the interpretation of 'naturally occurring' extends to social systems, which are, at root, biological. Hopefully, these examples cast light on the somewhat enigmatic title of this chapter. In most complex adaptive systems, there are local 'small-scale' components and interactions that are relatively well understood and predictable. These components and their interactions compound to create global 'large-scale' system behaviours that are, hopefully, desirable, but generally hard to predict in advance from knowledge of the small-scale characteristics.

2.2 Applications of BICAS

The increased interest in computer science in complex adaptive systems is not only a result of research into cognitive systems. Interest is also driven by the realization in computer systems engineering that it is becoming progressively more difficult to build and control increasingly large systems. This is compounded by the challenge to build computer systems that meet the future needs of society.

Although difficult to predict, it seems likely that the increased ubiquity of computing devices will pose many challenges. In particular, networking of heterogeneous processors will provide greater functionality and access to remote data stores. The architectures of such future systems may range from large purpose-built computer systems, comprising many tightly connected elements, through to large amorphous systems of loosely connected distributed and decentralized computing elements, where the network connectivity is dynamically changing. Our lack of understanding of how to design and control such large systems, across a broad spectrum of possible system architectures, is a major challenge.

As such systems become larger, the spatial and temporal interactions between different elements can lead to unpredictable behaviours, particularly in response to unforeseen disturbances. Consequently, large-scale properties of the system may no longer be easily predictable from the small-scale properties of its constituent elements. This can give rise to serious problems in quality assurance, i.e. in proving that a finished system meets its design specification. A notable example of a network in which small-scale interactions led to unforeseen effects with highly undesirable large-scale properties was the collapse in 1996 of the electricity supply grid on the eastern seaboard of the USA.

Thus, from an engineering perspective, behaviours that arise through the interactions of many separate component elements are often unwanted, whereas biological systems appear to take positive advantage of these emergent behaviours. Consequently, some researchers of computer systems have looked at biological systems for inspiration, attempting to harness emergent properties in a positive way. The hope is that such systems will not only cope well with the inherent uncertainties of the real world, but also will be adaptable to new and unforeseen environments and will be capable of self-repair and self-configuration to make the best use of available resources.

2.2.1 Circuit-level Computation

In the past, research in biologically inspired computer systems focused on the circuit-level understanding of interactions between neurons. The hope was that this research would provide insights into how to build more sophisticated computer architectures capable of solving more complex problems. Such biologically inspired architectures differ radically from the traditional ‘von Neumann’ computer architecture, consisting of a single memory processor bus. More recent research has investigated how system level analogies from the nervous system could aid the design of new computers, such as loosely coupled distributed and networked computer systems.

At the circuit level, computer science has long wanted to match the processing capacity of animal brains to perform intelligent tasks. It has been hypothesized that the development of economically interesting computational systems will require computer systems consisting of hundreds of thousands, if not millions, of computational elements.

The advent of VLSI technology to create integrated circuits and microprocessors was clearly a significant step toward making such systems possible. However, it remains a challenge to build computer systems containing the 10^{12} elements of the brain. Even superficially mundane issues such as provision of power, and heat dissipation, become major issues. If we consider an individual neuron to have the computational ability of a small 8-bit microprocessor, then a computer system matching the intelligence and abilities of the human brain may need as many as 10^{12} such processors. And they must fit into a reasonably small volume. If the failure rate is only 0.00001% processors per year, a system composed of 10^{12} processors will have lost one million processors after 12 months of operation. It will also have had dynamically to reconfigure its connectivity and load allocations as individual processors failed. Biological systems, therefore, appear much more energy-efficient and robust than even the most advanced engineered systems.

An important element in obtaining a deeper understanding of how to engineer such systems is the observation that their behaviour is not just a function of the behaviour of their individual component elements. It is also a function of the interrelationship – the network or system architecture – between them. Consequently, a deeper understanding of how network architecture affects dynamics issues, both in computer systems and in biological systems, requires more than just an understanding of the component elements. It also requires an understanding of the dynamics, stability and adaptability of the large coupled nonlinear systems produced by these interacting elements.

2.2.2 Networked and Distributed Computing

If we move from circuit level to system level, and to networked or distributing computing, we see a similar picture. Here system engineering endeavours to build large distributed computing systems at an ever-increasing scale. The Grid is one example.

However, there are other systems of significant economic importance, such as large telecommunication networks. Within the defence domain there are what are known as 'systems of systems' – that is, the integration of numerous systems and platforms, such as command and control, radar and unmanned autonomous vehicles. Here again, lack of understanding of non-linear and dynamic interactions between the components of a networked system often limits our understanding of the behaviour of the overall system. This is compounded by the fact that we increasingly require systems to cope with dynamic changes in the number of components and their connectivity. We also require systems to adapt to unforeseen events. Here again biological metaphors have been put forward as an alternative source of inspiration in the design of such systems. Often the intention is, rather than having a centrally controlled system, to allow it to self-organize in a distributed or decentralized fashion, that is, to engineer coarse-grained computer systems as complex adaptive systems.

2.2.3 *Artificial Autonomous Agents*

Over the past two decades, researchers have put significant efforts into exploring the design and construction of artificial autonomous agents. Autonomous agents are entities that are capable of coordinating perception and action, for extended periods of time, and without human intervention, in the pursuit of some set of goals.

Biological autonomous agents are better known as animals. Their 'goals' are generally to stay alive long enough to mate. Artificial autonomous agents are animal-like artefacts. They may be physical mobile robots, or purely virtual entities. Example applications include:

- *Autonomous mobile robots* These have many potential applications, from remote operation in hazardous environments, through to more mundane applications such as industrial automation, office cleaning or robotic 'pets'.
- *Virtual agents for simulation* Computer simulations of real agents within real environments can be used for applications in science and engineering. Examples include predicting the behaviours of animals or people in certain circumstances, perhaps to scientifically evaluate some hypothesis. Or they may be the product of an artist's imagination, finding uses in computer games and animations, or as a plausible synthetic actor in interactive education applications.
- *The construction of 'cyberspace' virtual agents* Artificial agents that are not direct correlates of real-world agents have a number of uses; for instance, in the management and manipulation of electronic data, performing useful roles in business and industry, such as stock market trading, or controlling manufacturing processes.

These and other applications may involve a single agent or groups of agents interacting either competitively or cooperatively. In all application areas, there are indications that the traditional von Neumann computer architecture that has served society so well for so long may well not be the most applicable for studying or creating biologically inspired complex adaptive systems. Some new alternative approaches appear to offer greater promise.

2.3 Opportunities for BICAS

The ultimate aim of research in BICAS is to understand and build artificial complex adaptive systems with the attractive properties of adaptation and resilience and self-organization that we find in naturally occurring complex adaptive systems. Over the past 15 years, dramatic falls in the real cost of processor power and memory and disk storage have led to a number of computing tools and techniques from BICAS. These have moved from being academic curios to powerful methods for critical industrial applications.

Many BICAS techniques, therefore, are now directly applicable to the bottom-up engineering of artificial cognitive systems. Furthermore, the same falls in the real costs of computing have opened up the use of these tools and techniques for advanced computer modelling and simulation studies in the scientific understanding of natural cognitive systems. Thus, in contrast to the top-down approaches that have historically determined the engineering of complex distributed systems, the greater emphasis on biological inspiration gives more hope for the genuinely productive interplay between the scientific and engineering approaches to the production of artificial complex adaptive systems. We can anticipate the influence of BICAS techniques on the development of artificial cognitive systems and on large-scale networked computing systems.

Biological metaphors in computer systems engineering therefore become relevant to cognitive systems for a number of reasons. They could help to meet the need for new computer architectures and could support the engineering of artificial cognitive systems. These techniques could also play a part in conducting computationally intensive scientific simulation studies of biological cognitive systems.

2.4 Structure of this Chapter

It is beyond our scope in this chapter to review relevant background literature. (See Cliff, 2003 for a historical overview of the influence of biological ideas and metaphors in computer science approaches to cognitive systems since the mid-1980s.) Each of the subsequent three sections reviews, from a computer science perspective, the current capability and outstanding problems in a particular area. Section 3 explores BICAS in the context of creating new massively parallel computing devices. Section 4 discusses the application of BICAS ideas in the development of new complex distributed networks of heterogeneous computing devices. Section 4 then examines the state of the art in research in biologically inspired artificial autonomous agents, dealing with both robotic agents and 'virtual' ones. We provide some conclusions in Section 6. The Appendix, Section 8, presents recommendations for future research.

3 CIRCUIT-BASED SYSTEMS

Jim Austin[2]

3.1 Introduction

In this section we consider how cognitive science may influence the development of massively parallel computers and how such development will aid the understanding of the nervous system. To date, massively parallel processing (MPP) systems contain, at most, tens of thousands of processing elements. With the continued development of VLSI, we can now build systems with millions of processing elements. We can think of this as ultra-massive parallel processing (UMPP).

UMPP has widespread applications – in parallel simulations for weather forecasting, simulation of biological systems, modelling of protein interactions and other areas of bioinformatics, and in modelling biological systems such as hearing and vision. UMPP could also bring benefits to engineering through improved systems for computational fluid dynamics and complex system modelling, for example.

It is a fact that biology has, through evolution, created UMPP in neural systems, thus, the study of such systems may tell us how we can build such complex computer systems. Such a study is likely to concentrate on the following issues:

- *Dynamics* This concerns how simple elements provide complex behaviour over time. Does the brain exploit dynamics more widely than conventional computer systems? How can we exploit such properties in computer systems?
- *Self-organization* This concerns how UMPP systems organize themselves for a

[2] Flaviu Adrian Marginean, York University, assisted in the early drafts and background work on this section.

particular task without intervention from outside. How does the brain self-organize its processing? Can this be used in UMPP to deal with component failure and more efficient operation?

- *Synchronization* This concerns how many processors operate without a single timing signal (clock). How does the brain synchronize its processing? Can we use such methods to help to design artificial systems?
- *Processing speed* The brain works fast, yet uses slow processors. How does the brain achieve this? Can such methods be exploited in UMPP where the conventional wisdom has always been to use very fast computing elements.
- *Power use* The brain can effectively manage power to very large numbers of processors. How does it regulate this? Can we use such methods when building UMPP systems?
- *Timing* This concerns the brain's ability to ensure everything happens on time and in the correct sequence, despite very variable components. How does the brain achieve this? Computing has a good understanding of this area; can these ideas be used to understand the brain?
- *Robustness* The brain can tolerate failures in components and continue to operate. How does the human brain achieve this? In UMPP systems component failure is inevitable; can we learn from how the brain manages?
- *Information representation in transmission* The brain uses different methods to communicate between its components than computers. How does it do this? Can we use such methods in computer systems to give better operating characteristics?
- *Construction and micro circuitry* The brain is organized in a complex hierarchy. Why is it organized like this, what benefits does it give? Can we use such methods in computing systems?

In section 3.2 we describe work in each area, while in section 3.3 we consider the key questions in more depth.

3.2 Key Questions

3.2.1 Dynamics

Does the brain exploit dynamics more widely than conventional computer systems? There is evidence that the brain uses a more dynamic computation framework than today's computers. Memory systems could exploit dynamic storage methods. In addition, the behaviour of groups of computers operating in a loosely coupled way has major similarities with how groups of organisms operate. The analysis of dynamics thus operates both at the large system level as well as at the circuit scale.

3.2.2 Self-organization

Does the brain have a level of self-organization? If so how is this achieved in a controllable and advantageous way? An understanding of this issue may allow us to build better computer systems that can react to change or damage in a predictable way (IBM, 2003; HP, 2003). It would also allow a system to adapt to and explore new and unforeseen situations.

At the micro-architecture level, this relates to how the interconnection circuits can rewire themselves to operate more effectively. Clearly this has a wide range of applications. For large systems the ability to adapt in a robust and predictable way so that a system maintains, or improves, its operation is a major question. Recently, large computer companies such as IBM and Hewlett-Packard have proposed the concepts of 'autonomous computing' and 'adaptive infrastructure'. These concepts involve self-optimization and self-repair in large-scale IT systems.

3.2.3 Synchronization

How does the brain synchronize its processing? The brain does not appear to operate with a central clock. Thus it must use some form of asynchronous operation. How is this achieved? We know little about how to build large asynchronous computer systems.

Most computer systems use a central clock. This consumes power every time a circuit is 'clocked', even if it is doing nothing, wasting energy. There is growing interest in asynchronous systems. In practical terms, this results in low-power processors such as the Crusoe processor. Research continues to develop tools that allow designers to develop and analyse such systems (Furber, 2003; Moore, 2003).

This knowledge could be of great value to neurobiologists in the analysis of biological systems. Computer scientists could benefit by studying the brain's use of asynchronous methods.

3.2.4 *Processing Speed*

The brain uses processing elements that operate in millisecond time intervals while computer systems operate in nanoseconds. How does the brain achieve rapid and real-time performance with relatively slow computing elements?

Researchers have done little to investigate the advantages, if any, of using many millions of slow processing elements rather than relatively few fast processing elements. There is little work on computer systems made up of a very large number of slow computing elements.

3.2.5 *Timing*

How does the brain deal with real time? Biological systems must be able to meet timing deadlines, such as moving an arm to catch a ball. If it is too slow to compute the motion, we miss! How is this reflected in the construction of biological networks?

One avenue in computer systems is the use of 'any-time' methods, systems that can report a result at any time during processing rather than waiting for the computation to complete. Does the brain use such a method? How well do natural systems achieve this? Can we learn from any methods they may use? These issues apply both at the large and small scale.

3.2.6 *Robustness*

How does human memory continue to operate despite 'component failures'? What are its properties? Today's computers use fast but brittle components while brains are slow but robust. Can we design systems made up of unreliable elements? UMPP will require the ability to survive component failures.

3.2.7 *Information Representation in Transmission*

We know very little about how the brain communicates data. We do know that it uses frequency encoding. However, it is unclear how the brain encodes messages. We do not know clearly how neurons and spike trains complement each other to create effective systems.

The application of the use of inter-neuron communication methods (i.e. spike-train representations) in computer systems may give a lead to cognitive science.

3.2.8 *Construction and Microcircuitry*

We have a good understanding of how to build artificial neural networks for real-world tasks. However, we do not know how to bring these together to build systems. Studying the brain may give us ideas on how to solve these problems. A major question is how the brain is organized hierarchically.

3.3 Five-year View

In five years' time we should have a clear understanding of these issues within cognitive and computer science. To achieve this we need to build bridges between the two disciplines supported by focused funding in these areas. Ideally, there would be a few 'core' teams, which provide outreach and support to other teams working in individual areas (listed in the key questions section). The engagement of computer hardware and software suppliers should be in place. Support from these major companies will enable pull-through of the technologies. SMEs would also be involved to provide specific underpinning technologies for UMPP.

3.4 Twenty-year View

Underpinned by the rapid change in computer hardware and software, the industry will have taken up the key technologies and concepts that have been shown to be of value in applications. By this time UMPP systems based on the issues we have described should be in use in many areas, particularly in engineering, medicine and weather modelling. Such systems will have overcome many current problems with computer systems. We should have systems which self repair, use very little power, utilize tens of millions of processors and can deal with the timing and synchronization problems that such systems raise. The development of concepts from this area will be used in modelling cognitive systems, and the value of the interaction at the circuits and systems level will have become accepted.

4 DISTRIBUTED, NETWORKED COMPUTING SYSTEMS

Andy Wright

4.1 Introduction

Computing is making ever-greater inroads into everyday life, both at home and at work. Corporate computer systems are rapidly expanding, linking together computers or networks that in the past have stood separate from each other. Increasingly, the problem of having insufficient data is eclipsed by the issue of dealing with huge quantities of data derived from different sources, and making sense of this via data-mining and visualization techniques. The ability to link systems, via communication and data networks with ever-increasing bandwidth, is a key enabler of this progress.

As such systems grow, it becomes increasingly difficult to control and maintain them. Apart from the issue of engineering reliable software, problems arise in the interaction of new system components with existing legacy subsystems.

As systems become larger, and the ability to network new and existing systems increases, the behaviour of these super systems can become less predictable. One reason for this is that the complex spatial and temporal coupling within the network produces non-linear feedback. This can compound across the system to produce highly non-linear global system dynamics. When these behaviours are unforeseen, they are said to 'emerge' from the system, and are little understood.

These unpredicted, or emergent, behaviours have had significant negative consequences on a number of occasions. One example is the collapse in 1996 of the power generation grid on the eastern seaboard of the USA. This started with a small fault on one transmission line.

It is an article of faith that the same process of compounded non-linear interactions within distributed networks of *biological* components frequently gives rise to highly desirable emergent behaviours. Thus, a central challenge for the design and construction of distributed network computer systems, acting as the massively parallel computing infrastructure for cognitive systems research, is to develop an engineering methodology that can design desirable emergent behaviours into systems and that can predict and minimize, or even eliminate, undesirable emergent behaviours.

If this is possible, it raises the possibility of an alternative bottom-up approach to designing distributed networked computer systems. Here the large-scale functional structure of the system is allowed to emerge from the interactions between the system's components at the small scale.

There is a significant interest in these ideas within industrial research labs in the UK. BAE Systems has a Biofutures Strategic Option. BT Exact has a research programme on Nature-Inspired Computing (BT, 2003). Hewlett-Packard Labs Europe is home to the Biologically Inspired Complex Adaptive Systems research group.

In 2001, IBM launched its 'autonomic computing' initiative (IBM, 2003). This ambitious programme draws inspiration from the autonomic nervous system, rather than the central nervous system, in the creation and maintenance of enterprise computing systems. IBM's claim is that future networked computing systems should be:

- s*elf-configuring*, i.e., adapting to dynamically changing environments
- *self-healing*, i.e., able to discover, diagnose and act to prevent disruptions
- *self-optimizing*, i.e., able to tune resources and balance workloads to maximize use of IT resources
- and *self-protecting*, i.e., able to anticipate, detect, identify, and protect against attacks.

The Adaptive Infrastructure programme of Hewlett-Packard has similar ambitions (HP, 2003).

4.2 State of the Art

4.2.1 Biologically Inspired Models of Collective Organization

In biology, it has long been recognized that simple collective animal systems can exhibit useful self-organizing or emergent behaviours (Gueron, 1996). Seminal work by Reynolds (1987) demonstrated that it was possible to reproduce some of these biological behaviours in simulations of groups of simple autonomous agents called 'Boids'. These replicated the collective behaviours of flocks of birds or shoals of fish. The Boids produced coherent 'flocking' behaviour, with each Boid following three simple rules to determine its speed and direction of movement. The global flocking behaviour of the system is not embodied directly in the Boids themselves but rather emerges from the nonlinear interactions of the three rules that dictated the interaction between them.

Notions of collective self-organization have parallels in statistical physics, particularly the physics of many-bodied systems. Indeed, the work by Toner and Tu (1998) showed that a set of equations of motion could describe the rule base for Boids and that tools from dynamical systems theory could generate a simple measure of the stability of the Boids system. From this it was possible to show that the system had transitions to chaos and so was unstable in certain regions.

From a systems point of view one disadvantage of the Boids model was that not all the rules were local in nature. One rule required the Boids to move toward the centre of the flock and so required global knowledge concerning the positions of all the other flock members. However, other work by Wright *et al.* (2000), using these physical models, has demonstrated that it is possible to produce fully decentralized models, based on Hamiltonian mathematics that exhibit similar behaviour. Here each agent tries to minimize some energy function. The emergent property of the system can be seen as an equilibrium state where the process of energy minimization causes each agent to balance its interaction with neighbouring agents. This research also demonstrated that it was possible to derive more descriptive entropic measures of the global system's behaviour that could be derived from the local properties of the individual elements.

The use of the flocking analogy is only one of a number that has provided the metaphor for the generation of novel algorithms within computer science. The computer science literature has examples of ideas derived from coherent groups or swarms of creatures, with particular inspiration coming from observations of the *hymenoptera* order of social insects, which includes bees, ants, wasps and termites (Yovits *et al.*, 1962). Means by which coherent collective 'swarms' in nature have influenced new approaches in computer science are reviewed by Bonabeau *et al.* (1997). A popular computer modelling environment for exploring issues in this approach is SWARM, originally developed at the Santa Fe Institute (SWARM, 2003).

4.2.2 Multi-agent Machine Learning

The significance of Reynolds's idea of Boids is that it demonstrated the possibility of reproducing desirable behaviours that can emerge from a collective system of multiple agents. However, methods for engineering the rules that produce these behaviours are frequently developed on an ad hoc basis. The design of controllable behaviours for more involved systems remains a significant challenge. Methods for automatic learning or evolutionary design-optimization are therefore a topic of intense research. Ideas from machine learning have been explored in this area (Hu and Wellman, 1998; Boutiler, 1999; Leslie and Collins, 2003).

One attraction of attempting to adapt established single-agent learning techniques to multi-agent learning is that certain single-agent reinforcement learning algorithms have been proven theoretically to converge to optimal solutions. Reinforcement learning techniques operate in situations where the agent receives payoff, reward or punishment, as a consequence of certain actions or sequences of actions in its environment. The aim of reinforcement learning algorithms is to develop a strategy that optimizes some function of total payoff received. A good review of these methods is given by Sutton and Barto (1998).

Algorithms for multi-agent reinforcement learning are complicated by the fact that in a multi-agent system each agent has its own strategy, but the reward to an individual is now not only a function of the environment but also of the strategies of the other agents. In consequence, the learning problem is non-stationary. This presents a significant obstacle to obtaining a provably convergent learning method. Nevertheless, such learning problems exist in any self-organizing or self-regulating distributed computer network where future actions of one node need to take account of possible future actions or counter-actions by other nodes.

Current approaches have other significant limitations. Typically they assume either unlimited communication between individuals or no direct communication. Clearly from a perspective of computer science or engineering this is unrealistic. Biological systems on the other hand appear to be able to cope with, and make use of, partial and incomplete information that may be communicated within a flock or colony. Many of the methods derived so far have yet to take this into consideration.

A separate issue is the notion of equilibrium. A Nash strategy, to be the best strategy to an individual, is predicated on other players acting rationally. That strategy may not be the most appropriate if an individual is playing with, or against, an irrational partner. What constitutes the most appropriate equilibrium and how this should be found is still an open question to which biological systems may offer some insight, and is of relevance in dealing with node failures or malfunction in distributed computer networks.

4.3 Key Questions

From the perspective of a network of systems, a number of key questions need to be addressed. A central requirement is obtaining an underlying theoretical structure for the analysis and design of large-scale complex biological and computing systems. In a similar way to the theoretical justification derived for artificial neural networks, such a theoretical framework will allow us to make assurances about the robustness and predictability of these collective systems. Such a theory is essential if such methods are to be of general use in computing and engineering where reliability is paramount.

Can we derive a theoretical basis, possibly based on complexity theory and game theory, that will:

- allow an understanding of complex biological systems such that their properties and mechanisms may be used in computing and engineering?
- provide a framework to build robust, predictable and practical collective multi-agent systems?
- provide a framework under which agent strategies can be learnt and adjusted as events occur and circumstances change?

Different components of the network will function with a degree of autonomy but at the same time a degree of coordination with related subsystems. Machine learning methods will be required both to learn appropriate control strategies for these subcomponents and to allow the system to update itself continually.

In the shorter term, ideas such as multi-agent reinforcement learning, co-evolution methods and their linkage to theoretical constructs such as evolutionary game theory will be important. A central difficulty with this approach is that it is not clear what equilibrium the system needs to achieve to function appropriately. If all components are rational, and have complete knowledge of their environment and other individuals in the system, then this should be the Nash equilibrium of the system. However, if any component is not fully rational, is uncertain about its state or the state of the system, then it is not clear what the individual's strategy should be, nor is it clear what the equilibrium points of the system are.

The development of robust multi-agent learning methodologies is a key requirement for the propagation of these systems. In the shorter term this raises two specific questions:

- What properties of a system is it desirable to learn and how can these be measured?
- How do we learn multi-agent strategies that are robust in an uncertain environment with limited communication?

4.4 Five-year View

The growth of networking and the greater uptake of personal and commercial computing systems over the next five years will increase the need for techniques that can cope with, and take advantage of, behaviours that emerge when these systems are integrated. The key to the exploitation of these system properties is the use of machine learning methods to learn appropriate behaviours for individual components at the small scale such that at the large scale the system has the required behaviours and attributes. In the short term, the initial progress will probably be through multi-agent learning methods derived from the ideas of reinforcement learning, game theoretic methods such as evolutionary game theory and co-evolution.

The past two years have seen increasing publication activity in this area. This will probably increase, providing the basis for a theoretical framework for these ideas and algorithms to implement them. Currently, application of these methods is limited to small toy problems. This is in part a reflection of the level of computational effort required for reinforcement learning.

Research has addressed more practical problems (e.g. Stone *et al.*, 2001), however, these overcome the computational burden by restricting the state space of the individuals. It is hoped that a greater understanding will lead to more practical, and less expensive, algorithms based on these ideas.

The immaturity of machine learning makes it difficult to predict the likely progress in this area of research over the next five years. However, a theoretical structure for the learning of multi-agent strategies is starting to emerge. This will mature over the next four or five years. Other research, in particular ideas from theoretical biology on animal behaviour (MacNamara *et al.*, 1997), and ideas relating to learning in games (Kearns *et al.*, 2000), will also have an impact, as will the development of other biological inspired methods.

4.5 Twenty-year View

Rapid and robust learning methods will help in the production and integration of large-scale collective multi-agent networked computer systems. The integration will take time and depends on the development of both theory and practical algorithms over the next five years or more.

System design is currently very different from this bottom-up approach, so we will need a good understanding of the relative advantages and disadvantages before people will adopt this new approach. This understanding will have to consider the robustness and practicality of adopting such methods. The techniques will not be adopted until the user community trusts them.

The integration of these ideas into new computing paradigms, such as software agents and collective robotics, will assist this process and provide the enabling technology that will allow these methods to be in future computing environments: environments such as the Grid and personal computing systems, where adaptability and system-to-system co-operation are important.

5 AUTONOMOUS AGENTS

Dave Cliff and Robert Ghanea-Hercock

5.1 Introduction

The creation of artificial autonomous agents with animal-like capabilities has attracted increased attention as a research topic over the past two decades. One of the main influences for this renewed interest was the desire to build truly autonomous mobile robots. Nevertheless, autonomous agents that exist purely as software entities, with no physical realization, also have a number of important commercial and scientific applications. This section reviews the state of the art, followed by a discussion of the role that biological metaphors might play in future developments. We then describe some key research questions, and close with speculations on potential applications in five years and 20 years time.

5.2 State of the Art

The word 'agent' has in the past decade come to mean many things to many people. There is a tension between some of those meanings. Some researchers consider the use of an agent metaphor to be a natural next step for use in computer programming in general.

Such a transition to 'agent-oriented programming' would be similar to the spread and adoption of object-oriented programming techniques – as embodied in programming languages such as C++ or Java – that occurred during the 1990s. Under such a view, any procedure, method or function performed by a computer program can in principle be referred to as an agent if the programmer chooses to do so. Indeed, non-agent legacy software systems can also be 'wrapped' in an agent interface, thereby hiding their non-agent origins.

This approach has some appeal when viewed within the context of the history of programming language design. However, it also widens the definition of the word 'agent' almost to the point of vacuity. The interested reader is referred to the *International Journal of Autonomous Agents and Multi-Agent Systems* (published by Kluwer since 1997) and the associated international conference series that has been running since 1998 (http://www.aamas-conference.org). Both are noted for their high editorial standards and for their very broad interpretation of what counts as an 'agent'.

Here we focus on those strands of autonomous agent research with a strong focus on complex adaptive systems, and where ideas or metaphors from biological systems have influenced the development of new techniques or technologies. That is, we focus here on the treatments of agents as biologically inspired complex adaptive systems (BICAS).

One obvious distinction within agent research is between real physical artificial autonomous agents, i.e. robots, and agents with no physical embodiment, i.e. software agents that exist purely in virtual environments. We discuss the state of the art in these two areas separately, in sections 5.2.1 and 5.2.2 respectively. However, the strong links and shared roots between these two areas mean that there are many common current research issues, as highlighted in sections 5.3 and 5.4.

5.2.1 Autonomous Robots

Although there are historic precedents, we can trace most current research in biologically inspired autonomous robotics to seminal papers published by Brooks in the mid-1980s (Brooks, 1985, 1986). In these papers, Brooks argued forcefully for a behaviour based or bottom-up approach to cognition. This approach presumes that displays of intelligence are the product of complex interactions between an agent's behavioural repertoire and its environment, where that agent's behavioural repertoire is itself the product of the non-linear system formed from multiple interacting behaviour generating modules within the agent.

The principal proposed advantages of this approach are adaptability and robust operation. Brooks and his students at MIT demonstrated these advantages in a series of 'insect-like' autonomous mobile robots. Yet it has proved difficult to scale up behaviour-based engineering techniques to deal with more cognitively challenging behaviours. Hybrid, top-down and bottom-up approaches met with limited success (Connell, 1992; Gat, 1992).

In the early 1990s, Brooks's group shifted its attention to the construction of a humanoid robot, called Cog (http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/cog.html), using behaviour-based control techniques. While the mechanical engineering and low-level sensory-motor coordination and control aspects of Cog were novel innovations (Williamson, 1999), successful demonstration of higher level cognitive functions again proved to be elusive.

One insight from the Cog project was the significance of social interactions between humans and humanoid robots. This was subsequently studied in more depth using the Kismet 'socially expressive' robot head developed by one of Brooks's PhD students (Breazeal, 2002). The lack of social ability in Cog led to research on behaviour based humanoid robots as models for the diagnosis and quantification of social development disorders such as autism (Scassellati, 2000).

Physical robots with animal-like capabilities for autonomous action and survival have many obvious applications in areas such as hazardous environments, including battlefields, industrial automation, domestic cleaning and security, and in the entertainment and leisure industries.

Robot models may also act as physical simulations of real creatures, to test scientific hypotheses concerning the organization of a real animal's sensory-motor control system, i.e. its nervous system (Franceschini *et al.*, 1992; Srinivisan *et al.*, 1997, 1998; Webb, 2000, 2002, 2003). Autonomous, biologically inspired, non-humanoid robots for the remedial therapeutic treatment of autistic children are also under development (Dautenhahn *et al.*, 2002). Biologically inspired control systems have recently been developed for 'intelligent' prosthetic limbs and other assistive robotic technologies; most notably at the MIT Leg Lab (http://www.ai.mit.edu/projects/leglab).

We can consider an individual robot with a behaviour-based control system to be a complex system. The small-scale interactions of its behaviour-generating modules give rise to its overall large-scale observable behavioural repertoire. Furthermore, the individual behaviour-generating modules may themselves be delivering behaviours as the large-scale consequences of small-scale interactions if, for example, a module involves the use of an artificial neural network. Moreover, at higher level of analysis, an individual autonomous robot can be looked at as a small-scale component in a large-scale system if it is one of a number of robots working together as some form of team.

Fruitful research in so-called 'collective robotics' has been under way for a little over a decade. Before that, material costs and high failure rates of the requisite technologies made serious research prohibitively expensive.

While studies of teams of humans collaborating and cooperating on the solution of tasks is a potentially valuable source of inspiration, much early work in collective robotics draws inspiration from a more lowly biological inspiration: the collective behaviour of social animals, and in particular the social insect order *hymenoptera*, which includes ants, bees, termites and wasps. More recently, a surprisingly large amount of research in collective robotics has been directed at an even lower form of life: soccer players, in the various leagues maintained by the international RoboCup (http://www.robocup.org) organization.

5.2.2 Software Agents

Research exploring BICAS approaches to the creation of autonomous agents that exist purely in software has the advantage that it eliminates the cost of constructing and maintaining real robots. While this is an advantage if we can simulate accurately in software the interactions of a real robot and its environment, the dangers of this approach are often seriously under-appreciated. If the simulation has been poorly verified, or not verified at all, then the simulation may not faithfully model the real-world system that it is intended to represent. Hence, the results from the simulation study may not be replicable in the real world. This danger is heightened when the agent is adaptive, as the adaptation mechanisms – for example, learning in a neural network, or the use of a genetic algorithm[3] to tune the design of the agent – may exploit flaws in the simulation, and this may go undetected. Among the biggest computational costs in creating accurate simulations of real-world robots, and their real-world environments, are those associated with simulating the mechanics, kinematics and dynamics in sensing and in acting. It can take considerable computer power to simulate the physical processes in the sampling of the ambient optic array by a video camera, or the results of torque applied by a motor.

Independent third-party 'middleware' software suppliers have developed general-purpose 'physics engine' libraries that can save time and money in the development of accurate simulations. Leading suppliers in this field are MathEngine (http://www.mathengine.com), Havok (http://www.havok.com), and Criterion Software (http://www.csl.com). Despite the potential heavy computational cost of simulating phenomena that 'come for free' when working with real robots, studies of simulated agents can collect rich streams of data, data that it may be impracticable or impossible to gather from a physical robot. Robot simulations also allow studies of failure modes that could be prohibitively expensive when working with real robots. For example, when developing 'flying robot' unmanned air vehicles (UAVs), many real-world failure modes would involve the loss, or destruction, of the UAV. A simulated crash is much less costly.

Various research teams have worked with well-validated simulations of real robots, where the lessons learnt in simulation have been demonstrated to be transferable to the real system. One notable body of work in this area was Jakobi's development

[3] Genetic algorithms (GAs) are computational search and optimization techniques inspired by Darwinian notions of evolution through random variation and directed selection. In a simple GA, a population of candidate solutions is maintained, and an iterative process evaluates the performance of each solution in the population against some operational measure of goodness or fitness. On each iteration, or generation, solutions with better fitness scores are 'selected' for 'breeding', a process that generations new candidate solutions via operations inspired by inheritance with recombination and mutation in sexual reproduction. Typically, the GA operates on encodings of the solutions, referred to as 'genotypes', which are expanded into actual testable solutions 'phenotypes', in the evaluation process. Because of this, GA theory is largely application independent. For further details see Goldberg (1989) or Mitchell (1998).

of a principled methodology for radically simplifying the computational cost of simulating agent–environment interactions, albeit one that is primarily applicable where those interactions are themselves simple (Jakobi, 1997). For realistically complex or dynamically varying interactions, it will take significantly more work to establish how best to make computational savings in the simulation.

However, not all software agents are accurate models of physical robots. With many software agents there is no need to model accurately real-world robots, or even real-world physics. We can talk of non-physically accurate software agents in two broad classes; abstract scientific agent-based models, and commercial engineering applications.

Agent-based models intended for scientific purposes are no less rigorous than simulated robot models. They eliminate major computational costs by working at levels of analysis where detailed and accurate models of physical interaction are not relevant. This is often valid where collective behaviour is the primary object of study. For example, in collective robotics, it is an item of faith that some form of inter-agent communication is useful for coordination among the group of agents.

Many interesting but different forms of communication, or different constraints on the space of communicative behaviours, can in principle be explored in simulations where software agents inhabit a world with minimally simple 'laws of physics'. For example, a limited vocabulary of communicative utterances (grunts) could be modelled as simply emitting one of a small number of types of grunt, which are heard instantly by all nearby agents, without modelling any details of sound production, or sound-wave propagation in air, or auditory sensing of sound waves. Exactly this approach has proved very successful in the simulation-based scientific study of the development or evolution of a number of communication systems, including the evolution of human language use (MacLennan, 1992; MacLennan and Burghardt, 1994; Noble, 2000; Noble *et al.*, 2001; Kirby, 2001). Two notable UK research clusters in this area are the BioSystems group at the Informatics Research Institute at the University of Leeds (http://www.scs.leeds.ac.uk/research/inf) and the Language Evolution and Computation group at Edinburgh (http://www.ling.ed.ac.uk/lec).

One very constrained form of inter-agent communication occurs in microeconomics, where traders interact within auction markets – i.e., buyers and sellers communicate by signalling prices of bids and offers. These and other abstract artificial economic systems have also been studied scientifically with some success using minimal simulation techniques (e.g., Epstein and Axtell, 1996). In the UK, Gilbert's team at Surrey University has pioneered agent-based simulations in economics and the social sciences (http://www.soc.surrey.ac.uk/research/cress).

Returning to the issue of abstracting away from accurate simulation of real-world physics, there are sound scientific models of agents moving over some area of space that pay little or no attention to modelling the physics of movement. For instance, a country-scale model of traffic flows across a highway network gains nothing by accurately simulating the physics of each car's individual movement, provided that the abstractions in the model preserve a representation of phenomena important at a higher level, such as the fact that if one car hits another, both are likely to stop and, at least partially, block the road.

Agent-based simulations of human activity have found increasing use over the past decade in health informatics, in epidemiology and the associated prediction of health-care demand for planning purposes, for example. They are also used in geographic information systems (GIS) applications, such as those used to predict the growth and spread of a city, and the effect of that growth on natural resources.

Business and Management

The use of complex systems thinking in academic schools of business and management has also grown steadily over the past decade. In the US, both the Santa Fe Institute (http://www.santafe.edu) and the New England Complex Systems Institute (http://www.necsi.org) appear to generate sizeable revenue from their offerings of business seminars and consultancies. In the UK, complex adaptive systems research applied to the sphere of business and management is well represented by the Complex Adaptive Systems Group at Oxford University's Said Business School (http://www.sbs.ox.ac.uk/html/faculty_seminars_complex_systems.asp), by the Complexity Research Programme at the London School of Economics (http://is.lse.ac.uk/complexity) and by NEXSUS, the Complex Systems Management Centre at Cranfield University School of Management (http://www.nexsus.org).

There are many potential commercial or applied-engineering uses of autonomous software agents. However, those actually deployed and making money are more rare. One of the more lucrative markets is computer-based entertainment. BICAS-type software agents have been used in computer games (Maes, 1995; Cliff and Grand, 1999) and in the animation of computer-generated characters for Hollywood movies. The Boids algorithm was used to animate stampeding herds of animals in Disney's cartoon feature *The Lion King* (Reynolds, 1987). In recent years, global revenues from computer games have consistently exceeded revenues from Hollywood movies. The production costs of main title computer games now routinely match those of medium-budget movies. A recent huge commercial success involving application of simulated human agents in a computer-game is the Sims series of games produced by Maxis (http://thesims.ea.com/). In these games, users create human-like agents, design their home and then guide their relationships and careers. We can also class as entertainment applications of BICAS more abstract computer games, such as the perennially popular SimCity series, also produced by Maxis, in which the player takes the role of town planner and mayor for an abstract simulation of a city.

In a style similar to these models developed for entertainment, commercial scientific modelling of real-world systems has recently turned to the techniques of autonomous software agents for the predictive simulation of real-world events or scenarios. Examples include training police strategists in the prevention, containment and control of crowds of rioters; or using demographic data and spatial geographic information models to predict the effects on revenue stream of relocating a factory or choosing a specific site for a new superstore. Prominent commercial companies in this space include GMAP (http://www.gmap.co.uk) in the UK, and the Bios Group (http://www.biosgroup.com) in the United States. Both offer, at various levels of abstraction, agent-based models of humans interacting in some space that represents a real-world geography or corporate organization. Their models can be used in management planning and training applications. Again, in such simulations, the small-scale interactions compound to give large-scale overall activity that is not readily predictable in advance.

However, not all software autonomous agents are designed to interact with a simulated environment that is intended to represent some real-world situation, or a realistically plausible but imaginary world, as is more often used in entertainment applications. Many researchers have studied the development of autonomous software agents intended to coordinate their perception and action in environments that are abstract 'cyber-spaces,' typically formed from a number of dynamic data-streams.

Auction Agents

One potential application is for individual 'personalizable' software agents that a user instructs to do their business on e-commerce sites such as online exchanges or auctions, simultaneously monitoring the bids and offers in multiple auctions so as to get the best deal; or possibly also so as to arbitrage across those auctions. Such an agent could be simultaneously active in tens or hundreds of different auctions, where those auctions do not necessarily all operate according to the same rules and protocols. Byde, Priest and Jennings (2002) recently explored this application area, albeit not using BICAS techniques.

For several years, academic researchers have developed software agents for autonomous automated trading on the international financial markets. However, the take-up of such technology for live applications by investment banks and financial exchanges appears to be very poor.

Solid data on successfully fielded applications in financial trading are notoriously sparse. The developer of any consistently profitable automated-trading method, agent-based or otherwise, has a manifest vested interest in keeping very quiet about that success, at least until they have banked enough money to retire comfortably. For reviews of BICAS-oriented approaches to the engineering design of trading agents, see Cliff and Bruten (1999) and Tesfatsion (2002).

Prominent international research groups with a strong interest in artificial autonomous agents for business and e-commerce include academic groups at the MIT Media Lab and at Michigan University in the United States, at Liverpool and Southampton Universities in the UK, at major industrial research labs such as IBM in New York State, and Hewlett-Packard Labs in Bristol, UK, along with smaller commercial enterprises such as Frictionless Commerce (http://www.frictionless.com) in the United States and LostWax (http://www.lostwax.com) in the UK. It is worth noting that in many of these groups the desire for raw profit typically takes much higher precedence than considerations of biological verisimilitude.

It has long been known from studies in experimental economics (e.g. Smith, 1962) that when groups of human traders come together in an appropriate free-market environment, the transaction prices in the market can rapidly and reliably converge on the market's theoretical equilibrium price. This is the price at which the quantity supplied by the population of sellers best matches the quantity demanded by the population of buyers, and so represents an optimal allocation of those scarce resources that are supplied by the sellers and that are demanded by the buyers. This view of real-world free-market economies as resource allocation mechanisms is appealing because they are typically asynchronous and decentralized. In particular, they do not require a centralized auctioneer to orchestrate proceedings. Hence, they offer another metaphor from the natural world that can influence the engineering design of distributed and decentralized systems where some population of consumers demands scarce resources. For instance, in a networked computer facility, the scarce resources demanded by users are likely to include processor time, disk space and network bandwidth.

If autonomous software agents are attached to each network resource, acting as sellers of the resource, and if autonomous software agents are also associated with each user's request for a job to be processed, then the agents can negotiate prices by, for instance, engaging in an auction. The intent is that at times of high demand the price of some of the facility's resources will rise, making them less attractive to some users. They then hold off from consuming those resources until demand falls and the price comes down.

This dynamic and decentralized market-based approach to computer load-balancing is one instance of a new approach to robustly solving dynamic resource-allocation problems, an approach known as market-based control (MBC). Much of the groundwork for MBC was laid in the collection of papers edited by Huberman (1988), who pioneered MBC approaches while a researcher at Xerox PARC. More recently, Clearwater's (1996) collection includes accounts of a number of successful MBC systems, including distributed computer system load-balancing, industrial job-shop scheduling and management of office-block air-conditioning.

Research groups with significant activities or investment in MBC include Southampton University and Hewlett-Packard Labs Bristol in the UK, and groups at the University of Michigan, University of Southern California, IBM T.J. Watson Research Labs, and Hewlett-Packard Labs Palo Alto in the United States.

5.3 Biological Metaphors in Research

Experience so far indicates that for any artificial autonomous agent, robotic or virtual, parallel distributed processing architectures, such as artificial neural networks, offer many advantages over centralized sequential control programs. Experience also demonstrates that purely manual design of such processing architectures is extremely difficult.

Traditional engineering design methodologies are not well-suited to creating asynchronous distributed networks of processors intended to operate without central control. Thus, automated adaptation techniques, both within the lifetime of an agent and also over successive evolving generations of agents, remain the most promising approach to creating processing architectures. For this reason, biological metaphors such as adaptive artificial neural networks, and evolutionary computation techniques such as genetic algorithms, are likely to remain strong influences in future BICAS research.

As the number of individual processing units – for example, artificial neurons, or behaviour-generating modules – in an agent increases, it becomes more difficult to specify an appropriate connectivity between the components in advance, and also to reconfigure the connectivity to account for component failure or malfunction. For this reason, ideas from developmental biology could become more influential, as artificial autonomous agents undergo some kind of embryological morphogenesis, with the processing architectures initialized by 'seeding' and then growing and self-organizing. The Amorphous Computing team at MIT produced a review of their pioneering work in this area (Abelson *et al.*, 2001).

5.4 Key Research Questions

A number of technical challenges currently limit the full exploitation of the BICAS approach in autonomous agent research. One pressing question concerns how an autonomous agent can generate novel behaviours through learning, and integrate these with its functioning set of behaviours and strategic objectives. Specifically, this involves enabling the development of architectures, control functions, interface approaches and, for robots, physical mechanisms that allow modules to dynamically and automatically reconfigure themselves.

Early claims that it would be easy to extend layered behavioural architectures have proven difficult to substantiate in practice. A further issue within collective robotics and multi-agent systems is the challenge of striking the right balance between individual robot, or agent, capability and inter-agent cooperation.

The use of evolutionary optimization techniques such as genetic algorithms (GAs) in the semi-automated design of autonomous agent architectures requires a space of possible designs to be defined. This definition is often made implicitly, via the specification of how the 'agent genotype' genetic encodings operated on by the GA are interpreted as agent phenotypes in the evaluation of the genotype's 'fitness' value.

Research has demonstrated many successful applications of GAs in the design of autonomous agents both in robotics and software. However, the design of appropriate genetic encodings, and their associated mappings onto agent phenotypes via a morphogenesis process, and also of productive fitness evaluation functions, remains an ad hoc art, rather than an operationalized engineering discipline. This has long been recognized by practitioners of GA agents, but no clear solutions are in sight.

Recent developments in neuroscience have identified the presence of *gaseous* neurotransmitters, nitric oxide in particular, that operate in addition to the long-known direct neurotransmitter and electrical synaptic signalling mechanisms. This discovery indicates that neurons may be capable of signalling in a 'diffuse' manner, by release of gases to nearby neurons (Elphick *et al.*, 1995, 1996).

One of the first artificial neural network models that interacted with an associated computational biochemistry was described by Grand, Cliff and Malhotra (1997) and Grand and Cliff (1998). There, though, the interaction was closer in spirit to the way in which hormones may grossly modulate neural activity.

Recent research in artificial neural networks that incorporate models of gaseous neuro-modulation, in addition to direct connections between the neurons, includes work that offers some intriguing insights (Phillippides *et al.*, 2000; Husbands *et al.*, 2001). Nevertheless, the full computational and engineering implications of diffuse neurotransmission, possibly also with interactions from an artificial endocrine system, remain to be further explored in artificial autonomous agents.

One final issue that is starting to cause concern among practitioners and sympathetic observers of research in BICAS autonomous agents is the relatively slow rate of increase in the desired or intended cognitive complexity of the autonomous agents studied, whether GA-evolved or hand-designed. It is now over 15 years since Brooks's papers established the field of biologically inspired behaviour-based systems. It is becoming harder to use the excuse that this is a relatively new approach to explain why it has failed to tackle problems that are more cognitively challenging than navigating an environment while avoiding collisions.

The fear is that the BICAS approach is reaching an impasse similar to that which occurred in traditional logic-based top-down AI around the time that Brooks wrote his seminal papers. One response to this is that it is a fear based on impatience and ignorance, symptomatic of failing to appreciate the inherent difficulty of creating artificial systems that can attain the cognitive complexity needed for even simple, restricted, task domains and environments. The counter to this response is that it is exactly the excuse made by practitioners of logic-based top-down AI.

The problem is perhaps most acute in academic research in behaviour-based robotics in the UK. Severely limited budgets, in comparison to labs in the United States, often force researchers to do as much as possible with little expenditure on robot hardware. As a result, the replacement cycle for robot hardware is much longer in the UK than in better-funded labs elsewhere. This may force researchers to employ robot platforms that are ill-suited to, or simply incapable of, studying behaviours that are more cognitively complex. Support for this hypothesis comes from the observation that the teams most actively trying to accelerate the cognitive complexity of tasks studied within BICAS-agent research use abstract idealized simulation studies (Beer, 1996; Slocum *et al.*, 2000), despite being in the United States and having previously worked with advanced robotic hardware.

Finally, it is worth noting that two UK initiatives could potentially complement the Foresight Cognitive Systems Project, particularly with respect to the BICAS topics addressed here.

First, in November 2002 the UK Computing Research Committee (UKCRC), a body funded by the IEE and the British Computer Society (BCS), held a two-day workshop to explore the possibility of formulating 'grand challenge' research agendas for computer science. The challenges are intended to be similar in spirit to the Hilbert problems in mathematics or to President Kennedy's succinct mission-statement for the Apollo Moon-shot programme. From 109 outline submissions, the UKCRC has formulated seven grand-challenge proposals. In many cases these proposals subsume a large number of the outline submissions. Two of these seven are manifestly relevant to BICAS approaches. The first, 'In Vivo – In Silico,' intends to explore high-fidelity reactive computational modelling of development and behaviour in plants and animals. The second, 'The Architecture of Brain and Mind,' aims to develop integrated computational accounts of low-level neuronal brain processes and high-level cognitive behaviours. The UKCRC Grand Challenge website has further detail of these proposals: http://umbriel.dcs.gla.ac.uk/NeSC/general/esi/events/Grand_Challenges.

Second, at the time of writing, a proposed UK National Centre for the study of complex IT systems is in the early stages of planning. Significant funding has been agreed in principle from the Engineering and Physical Sciences Research Council (EPSRC), the Department of Trade and Industry (DTI) and the Information Age Partnership (IAP), a consortium of UK-based IT industries. Research on incorporating BICAS approaches within future IT systems will probably be on the agenda for this centre.

5.5 Five-year View

5.5.1 *Autonomous Robots*

The next generation of Mars rovers will incorporate highly flexible planning and reactive control architectures, thereby increasing their autonomy. This will require integrating the full range of recent advances in behavioural vision systems, behaviour management and reasoning.

Compact and high power parallel computing systems, such as those discussed in sections 3 and 4, will enable onboard strategic task reasoning and long-term planning and coordination. Parallel work on high-resolution sensing systems, for example, high accuracy GPS, and ultra-wide-band radio sensors, enable precision navigation and the co-ordination of many robots. Novel drive mechanisms, such as PZT actuators, enable very small mobile robots, (http://www.darpa.mil/mto/drobotics). Military requirements for covert remote surveillance produces autonomous units smaller than 10 centimetres that can operate in coordinated teams. Detailed biological neuro-behavioural patterns are mapped into basic robotic systems.

Next generation AIBO-style home entertainment robots will become commodity items as costs fall and functionality increases to include home monitoring, child care and networked integration with personal computers and media systems.

5.5.2 *Autonomous Software Agents*

Entertainment applications:

- 'Live' but entirely computer-generated versions of some sports become available over Internet broadband and/or broadcast TV and/or on mobile phones, where the participants in the sports – e.g. soccer players, race-horses, car drivers, or robot warriors – are synthetic agents, possibly with BICAS architectures. These agents are trained and/or bred, evolved, by individual users/players/viewers, or networked syndicates of users/players/viewers, on desktop PCs. The funding model is based on income from: online gambling; provision of cheap 'filler' content to TV broadcasters; and advertising hoardings on the virtual trackside. A clear precursor to this development is the iRace virtual horse-racing system planned as a joint venture by Telewest and VIS Entertainment, to be broadcast on Sky Digital (http://www.irace.com).

Engineering applications:

- Stable market-based control systems, populated entirely by artificial agents, used for resource allocation in clustered computer facilities and perhaps also in national Grid computer networks.
- Small-scale live trials of online international financial markets populated, at point of execution, entirely by artificial autonomous trader-agents, operated by smaller 'boutique' exchanges in major

financial centres such as London and New York.

- Federated networks of warehoused central computing facilities housing tens of thousands of server machines, all connected on an ultra-high-bandwidth network, providing what IBM refers to as 'computing on demand' and Hewlett-Packard as 'utility data centres' come on-stream. They use BICAS techniques to provide 'autonomic' or 'adaptive infrastructure' self-healing resilience to load fluctuations, component failures, and attack by computer viruses and worms; possibly with market-based control for load-balancing and thermal resource management.

Scientific applications:

- First genuinely predictive computer-simulation model of a simple invertebrate, perhaps *C. elegans*, less possibly *D. melanogaste* (but see Hamahashi and Kitano, 1998), allowing accurate studies *in silico* of morphogenesis and development processes, and lifetime adaptation/ habituation.

5.6 Twenty-year View

5.6.1 *Autonomous Robots*

The goal:

- Cost reductions bring advanced socially aware robots into commercial and domestic environments.
- Human–robot interaction is smoothly integrated at a verbal and social level, possibly using implants for direct control from a human user's nervous system in assistive and prosthetic applications.
- Robots operate in large numbers in hazardous environments and where there are labour shortages.
- Robotic healthcare nurses are a premier application, as they can lift and move patients around a hospital or home, and advise medical staff of the patient's condition from pervasive bio-sensors and provide access to online sources of medical knowledge. Some routine surgery – e.g. appendectomies, or eye-cataract operations – is also performed by autonomous robots.
- Low cost versions provide domestic health care to the elderly and disabled.
- Robotic classroom assistants become available for interactive teaching and remedial therapies.
- Cooperative teams of autonomous robots provide real-time surveillance and intervention in battlefields and emergency situations; and are used in the manufacture and assembly of complex machines such as combat aircraft.

5.6.2 *Autonomous Software Agents*

Entertainment:

- Real-time semi-improvised interactive movies, or, more likely, soap operas, with emotionally and cognitively plausible synthetic actors, 'synthespians', become common in domestic computer entertainment, blurring the distinction between a movie and a computer game.

Engineering:

- Silicon is possibly no longer the dominant substrate for engineered computing devices, as methods for reliable computation in genetically engineered organic substrates mature. However, quantum computation technique threatens the future of this bio-computation revolution.
- Nano-scale computing devices linked to microelectromechanical systems (MEMS) can be mass-produced at a unit cost so close to zero that MIT's late-twentieth-century vision of amorphous computing becomes common.
- Notion is widespread of computer systems being composed of many thousands of connected processors, collectively having their own 'immune system'.
- International financial markets are populated entirely by artificial trader agents.

Science:

- First full predictive computer simulation, or robotic, model of a simple vertebrate, e.g. an anuran (Arbib, 1987, 2003).

6 CONCLUSIONS

In this chapter we have considered three domains that relate to:

- How an understanding of the global behaviour of a system is related not only to it constituent elements but also the inherent interactions between these elements.
- How, with this understanding, we can derive methods and frameworks that will allow the design, development and control of large-scale distributed systems and groups of agents where the control exists at the local level – that is, it is distributed – while the behaviour of the overall system is predictable and controllable.

The domains we have considered relate to three application areas:

- Circuit-level computation, the production of tightly coupled ultra-massive parallel processors (UMPP).
- Distributed and networked computing and the design and control of the dynamics of loosely coupled networked computer systems such as may be produced by the Grid.
- Artificial autonomous agents, the design and control of both software and robotic agents and their environments.

These areas depend on the ability of the agents and system elements to adapt to stimuli and environmental change, to produce the desired overall behaviour of the agent collective or system. This ability to adapt offers many properties that are desirable in current computational systems such as the ability to: self-organize and to cooperate, to make the best of limited resources or to overcome some adversary, such as a software virus; to self-repair and maintain to overcome damage, component failure; and to improve functionality. These are similar to the published requirements of IBM's Autonomic Computing programme, and to aspects of Hewlett-Packard's vision of an Adaptive Infrastructure for IT systems.

In all the domains we described, there is progress on a broad front. Specifically in:

- derivation of a theoretical understanding of these systems
- experimental testing of ideas mostly inspired from biology but also from other areas of science and social science
- the practical implementation and demonstration of these ideas.

Although we have identified a number of key questions for each domain, three common issues emerge. These are:

- To build upon existing theoretical work and produce a coherent theory that will underpin complex adaptive biological systems. The advent of a theoretical understanding revolutionized the area of artificial neural networks.
- To build on existing methods to produce robust learning or evolution methods that are able to adapt these systems to their environments.
- To extend the practical application of these ideas. This will demonstrate the practicality of this bottom-up approach in new domains, breaking down barriers to their use and facilitating integration into existing systems.

Acknowledgements

The authors thank Flaviu Marginean for his help and background work in the construction of this chapter.

Bibliography

Abelson, H., Allen, D., Coore, D., Hanson, C., Rauch, E., Sussman, G.J. and Weiss, R. (2001) Amorphous computing. *Communications of the ACM*, 43 (5): 74–82.

Aberystwyth (2003) University of Aberystwyth website, http://www.aber.ac.uk/~dcswww/Research/robots/.

ActiveRobotics (2003) ActiveRobotics website, http://www.activrobots.com/.

Arbib, M.A. (1987) Levels of modelling of mechanisms of visually guided behaviour. *Behav. Brain Sci.*, 10: 407–465.

Arbib, M.A. (ed.) (2003) *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press Bradford Books.

Arkin, R.C. (1990) Integrating behavioural, perceptual, and world knowledge in reactive navigation, *Robotics Autonomous Syst.* 6: 105–122.

Ashby, W.R. (1952) *Design for a Brain.* London: Chapman & Hall.

Asynchronous Embryonics (2003) http://www.elec.york.ac.uk/bio/bioIns/asynchronous/asynchronous.html.

AURA (2003) The AURA research on 'high performance pattern recognition with neural networks', http://www.cs.york.ac.uk/arch/nn/.

Beer, R.D. (1996) Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M.J. Matariae, J.-A. Meyer, J. Pollack and S.W. Wilson (eds), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press Bradford Books, pp. 421–429.

Bonabeau, E.G., Theraulaz, D.J.L., Aron, S. and Camazine, S. (1997) Self-organization in social insects. *Trends Ecol. Evol.*, pp. 188–193.

Boutiler, C. (1999) Sequential optimality and coordination in multiagent systems, *IJCAI-99*, 5: 393–400.

Breazeal, C. (2002) *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Brooks, R.A. (1985) A robust layered control system for a mobile robot. AI Memo 864, MIT AI Lab. Subsequently published in *IEEE J. Robotics Automation*, 2 (1):14–23.

Brooks, R.A. (1986) Achieving artificial intelligence through building robots. AI Memo 899, MIT AI Lab.

BT (2003) BT website on 'Nature-Inspired Computing', http://more.btexact.com/projects/ftg.htm.

Byde, A., Priest, C. and Jennings, N.R. (2002) Decision procedures for multiple auctions. *Proc. 1st Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, Bologna, Italy, pp. 613–620.

C4ISTAR (2003) Network Enabled C4ISTAR Swarming Conference, http://aan.xservices.com/swarm/information.htm.

Claus, C. and Boutiler, C. (1998) The dynamics of reinforcement learning in cooperative multiagent systems. AAAI-98, 1.

Clearwater, S.H. (1996) *Market-Based Control: A Paradigm for Distributed Resource Allocation.* World Scientific Publishing Co., Singapore, London and New Jersey.

Cliff, D. and Noble, J. (1997) Knowledge-based vision and simple visual machines. *Phil. Trans. R. Soc. Lond. B*, 352: 1165–1175.

Cliff, D. and Grand, S. (1999) The creatures global digital ecosystem. *Artif. Life*, 5 (1): 77–93.

Cliff, D. and Bruten, J. (1999) Animat Market-Trading Interactions as Collective Social Adaptive Behavior. *Adapt. Behav.*, **7** (3&4): 385–414.

Cliff, D. (2003) Biologically-Inspired Computing Approaches to Cognitive Systems: A Partial Tour of the Literature. HP Labs Technical Report HPL-2003-011.

COGS (2003) Sussex University COGS group website, http://www.cogs.susx.ac.uk/ccnr/.

CMU (2003) Carnegie Mellon Robotics Laboratory website, http://www.ri.cmu.edu/general/research.html/.

Connell, J.H. (1992) SSS: a hybrid architecture applied to robot navigation, *Proc. IEEE Robotics and Automation Conf.*, pp. 2719–2724.

Czirok, A., Ben-Jacob, E., Cohen, I., Shochet, O. and Vicsek, T. (1996) Formation of complex bacterial colonies via self-generated vortices. *Phys. Rev. E*, 54: 1791–1996.

DARPA (Defense Advanced Research Projects Agency) (2003) DARPA, http://www.arpa.mil/grandchallenge/overview.htm.

Dautenhahn, K., Werry, I., Rae, J., Dickerson, P., Stribling, P. and Ogden, B. (2002) Robotic playmates: analysing interactive competencies of children with autism playing with a mobile robot', in K. Dautenhahn, A. Bond, L. Canamero, B. Edmonds (eds), *Socially Intelligent Agents – Creating Relationships with Computers and Robots.* New York: Kluwer Academic.

Edinburgh (2003) Edinburgh University, http://www.dai.ed.ac.uk/groups/mrg/MRG.html.

Edwards, P. and Murray, A. (1998) Fault-tolerance via weight-noise in analog VLSI implementations – a case study with EPSILON. *IEEE Transactions on Circuits and Systems II : Analog and Digital Signal Processing*, 45 (9): 1255–1262.

Elphick, M.R., Kemenes, G., Staras, K. and O'Shea, M. (1995) Behavioural role for nitric oxide in chemosensory activation of feeding in a mollusc. *J. Neurosci.*, 15 (11): 7653–7664.

Elphick, M.R., Williams, L. and O'Shea, M. (1996) New features of the locust optic lobe: evidence of a role for nitric oxide in insect vision. *J. Exp. Biol.*, 199: 2395–2407.

EPFL (2003) EPFL website, http://www.epfl.ch/pages/research/contents.html.

Epstein, J.M. and Axtell, R.L. (1996) *Growing Artificial Societies: Social Science from the Bottom Up.* Cambridge, MA: MIT Press.

Exeter (2003) Exeter 'Neural Computing Group' http://www.dcs.ex.ac.uk/research/neural/prj.htm#Goal.

Evolutionary Electronics (2003) Sussex 'Evolutionary Electronics' website, http://www.cogs.susx.ac.uk/users/adrianth/

Franceschini, N., Pichon, J.-M. and Blanes, C. (1992) From insect vision to robot vision. *Phil. Trans. R. Soc. Lond. Ser. B*, 337 (1281): 283–294.

Fudenberg, D. and Levine, D.K. (1998) *The Theory of Learning in Games.* Cambridge, MA: MIT Press.

Furber, S. (2003) Amulet group research pages, http://www.cs.man.ac.uk/amulet/.

Garis, H. (2003) Group pages, http://www.cs.usu.edu/~degaris/.

Gat, E. (1992) Integrating reaction and planning in a heterogenous asynchronous architecture for controlling real world mobile robots. *Proc. Tenth Natl Conf. Artif. Intell. (AAAI).*

Gatsby (2003) UCL Gatsby Computational Neuroscience Unit, website, http://www.gatsby.ucl.ac.uk/.

Goddard, N. (2003) website, http://anc.ed.ac.uk/~ngoddard/.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison–Wesley.

Grand, S. and Cliff, D. (1998) *Creatures*: Entertainment Software Agents with Artificial Life. *Autonomous Agents Multi-Agent Syst.* 1 (1): 39–57.

Grand, S., Cliff, D. and Malhotra, A. (1997) Creatures: artificial life autonomous software agents for home entertainment. In Johnson, W.L. (ed.), *Proceedings of the First International Conference on Autonomous Agents*. New York. ACM Press, pp. 22–29.

Grossberg, S. (2003) ART networks website, http://cns-web.bu.edu/Profiles/Grossberg/.

Gueron, S.A., Levin, S. and Rubenstein, D.I. (1996) The dynamics of herds: from individuals to aggregations. *J. Theoret. Biol.*, 182: 85–98.

Guestrin, C., Koller, D., Parr, R. (2001) *Multiagent Planning with Factored MDPs.* NIPS.

Hamahashi, S. and Kitano, H. (1998) Simulation of *Drosophila* embryogenesis. In *Proc. Sixth International Conference on Artificial Life (AlifeVI)*. Cambridge, MA: MIT Press. pp. 151–160.

Hogg, T. and Huberman, B. (1991) Controlling chaos in distributed systems, *IEEE Trans. Systems, Man, and Cybernetics (Special Section on DAI)*, 21 (6): 1325–1332.

HP (2003) HP's website on Biologically-Inspired Complex Adaptive Systems, http://www.hpl.hp.com/research/bicas/.

Hu, J. and Wellman, M.P. (1998) Multiagent reinforcement learning: theoretical framework and algorithm. In *Proc. Fifteenth International Conference on Machine Learning*, pp. 242–250.

Huberman, B.A. (ed.) (1988) *The Ecology of Computation*. Amsterdam: Elsevier/North-Holland.

Huberman, B.A. and Adamic, L.A. (1999) Evolutionary dynamics of the World Wide Web. Technical Report, Xerox Palo Alto Research Centre.

Husbands, P., Philippides, A., Smith, T.M.C. and O'Shea, M. (2001) Volume signalling in real and robot nervous systems. *Theory Biosci.*, 120: 253–269.

IBM (2003) IBM Autonomic Computing website, http://www.research.ibm.com/autonomic//

Jakobi, N. (1997) Evolutionary robotics and the radical envelope-of-noise hypothesis. *Adapt. Behav.*, 6 (2): 325–368.

Jakobi, N., Husbands, P. and Harvey, I. (1995) Noise and the reality gap: the use of simulation in evolutionary robotics. In F. Moran, A. Moreno, J. Merelo and P. Chacon (eds), *Advances in Artificial Life: Proc. 3rd European Conference on Artificial Life*. Springer-Verlag, Lecture Notes in Artificial Intelligence 929, pp. 704–720.

JPL (2003) NASA JPL website, http://robotics.jpl.nasa.gov/.

Kearns, M., Mansour, Y. and Singh, S. (2000) Fast planning in stochastic games. *Proc. 16th Conf. Uncertainty* in *AI*. San Francisco, CA: Morgan Kaufmann. pp. 309–316.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. Evolutionary Computation*, 5 (2):102–110.

Kirkpatrick, S., Gelatt, C. and Vecchi, M. (1983) Optimization by simulated annealing. *Science*, 220: 671–680.

Kohonen, T. (2003) http://www.cis.hut.fi/research/som-research/teuvo.html.

Kuhn, D.R. (1997) Sources of failure in the public switched telephone network. *IEEE Computer*, 3 (4): 31–36.

Langton, C.G. (1989) Artificial life. In C.G. Langton (ed.), *Artificial Life: Proceedings of* *the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*. Los Alamos, NM, September 1987, volume 6 of Santa Fe Institute Studies in the Sciences of Complexity, pp. 1–47. Redwood City, CA: Addison–Wesley.

Leslie, D.S. and Collins, E.J. (2003) Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Ann. Appl. Probab.*, 13(4): 1231–1251.

Lund, H., Webb, B. and Hallam, J. (1998) Physical and temporal scaling considerations in a robot model of cricket calling song preference, *Artif. Life*, 4 (1): 95–107.

Maass, W., Natschläger, T. and Markram, H. (2002) Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.*, 14 (11): 2531–2560.

Macaque (2003) Collations of connectivity data on the Macaque brain. http://www.cocomac.org/.

MacLennan, B.J. (1992) Synthetic ethology: an approach to the study of communication. In Langton, C.G., Taylor, C., Farmer, J.D. and Rasmussen, S. (eds), *Artificial Life II: Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*, volume X of *Santa Fe Institute Studies in the Sciences of Complexity*. Redwood City, CA. Addison–Wesley, pp. 631–658.

MacLennan, B.J. and Burghardt, G.M. (1994) Synthetic ethology and the evolution of cooperative communication. *Adapt. Behav.*, 2 (2): 161–188.

MacNamara, J.M., Webb, J.N., Collins, T., Slzekely, E.J. and Houstton, A. (1997) A general technique for computing evolutionary stable strategies based on errors in decision making. *Theoret. Biol.*, 189: 211–225.

Maes, P. (1995) Artificial life meets entertainment: lifelike autonomous agents. *Communications Assoc. Computing Machinery*, 38 (11): 108–114.

Mass, W. and Bishop, C.M. (1999) *Pulsed Neural Networks*. Cambridge, MA: MIT Press.

Mataric, M.J. (1998) Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior. *Trends Cogn. Sci.*, 2 (3): 82–87.

McCarthy, J., Minsky, M.L., Rochester, N. and Shannon, C.E. (1955) *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Available at: http://www.formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Mathemat. Biophys.*, 5: 115–133.

Mead, C.A. (2003) Carver Mead's 'Physics of Computation' Research Group, http://www.pcmp.caltech.edu/.

MIT (2003) MIT Humanoid Robots Group website, http://www.ai.mit.edu/projects/humanoid-robotics-group/.

Mitchell, M. (1998) *An Introduction to Genetic Algorithms.* Cambridge, MA: MIT Press Bradford Books.

Moore, S.W. (2003) Self timed Logic group pages, http://www.cl.cam.ac.uk/Research/Rainbow/projects/selftimed.html.

Murray, A.F. (2003) A Murray group web pages, http://www.see.ed.ac.uk/~neural/pcdf/NeuralPage/Neural.html.

Nash, J.F. (1951) Non-cooperative games. *Ann. Math.*, 54: 286–295.

Neumann, P.G. (1990) Cause of AT&T network failure. *The Risks Digest*, 9.

NIPS (2003) NIPS workshop on Multi-Agent Learning: Theory and Practice http://www.cs.rutgers.edu/~mlittman/topics/nips.html.

Noble, J. and Cliff, D. (1996) On simulating the evolution of communication. In P. Maes, M.J. Matariæ, J.A. Meyer, J. Pollack and S.W. Wilson (eds), *From Animals to Animats 4: Proc. Fourth Int. Conf. on Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press Bradford Books, pp. 608–617.

Noble, J. (2000) Co-operation, competition and the evolution of pre-linguistic communication. In C. Knight, J.R. Hurford and M. Studdert-Kennedy (eds), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge: Cambridge University Press.

Noble, J.J., Di Paolo, E.A. and Bullock, S. (2001) Adaptive factors in the evolution of signalling systems. In A. Cangelosi and D. Parisi (eds), *Simulating the Evolution of Language.* Berlin: Springer-Verlag, pp. 53–78.

Philippides, A.O., Husbands, P. and O'Shea, M. (2000) Four-dimensional neuronal signaling by nitric oxide: a computational analysis. *J. Neurosci.*, 20 (3): 1199–1207.

Plymouth (2003) Plymouth Centre for Neural and Adaptive Systems, website, http://www.tech.plymouth.ac.uk/soc/research/neural/.

Reynolds, C. (1987) Flocks, herds, and schools: a distributed behavioral model. *Computer Graphics*, 21 (4): 25–34.

Reynolds, C. (2003) 'Boids' website, http://www.red3d.com/cwr/.

Robocup (2003a) Robocup website. http://www.robocup.org/.

RoboCup (2003b) RoboCup Special Interest Group (SIG) on 'Multiagent Learning'. http://sserver.sourceforge.net/SIG-learn/.

Rosenblatt, F. (1962) *Principles of Neurodynamics*. New York: Spartan.

Rosin, C.D. and Belew, R.K. (1997) New methods in competitive coevolution. *Evolutionary Computation*, 5: 1–29.

Scassellati, B. (2000) Investigating models of social development using a humanoid robot. In B. Webb and T. Consi (eds), *Biorobotics*. Cambridge, MA: MIT Press.

Slocum, A.C., Downey, D.C. and Beer, R.D. (2000) Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In J. Meyer, A. Berthoz, D. Floreano, H. Roitblat and S. Wilson (eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*. MIT Press, pp. 430–439.

Smith, V.L. (1962) An experimental study of competitive market behavior. *J. Polit. Econ.*, 70: 111–37.

Sony (2003) Sony Laboratories website, http://www.csl.sony.fr/Research/Topics/DevelopmentalRobotics/index.html.

Srinivasan, M.V. (2003) Website, http://cvs.anu.edu.au/bioroboticvision/brv.html.

Srinivasan, M.V., Chahl, J.S., Nagle, M.G. and Zhang, S.W. (1997) Embodying natural vision into machines. In M.V. Srinivasan and S. Venkatesh (eds), *From Living Eyes to Seeing Machines.* Oxford: Oxford University Press, pp. 249–265.

Srinivasan, M.V., Chahl, J.S., Weber, K., Venkatesh, S., Nagle, M.G. and Zhang, S.W. (1998) Robot navigation inspired by principles of insect vision. In A. Zelinsky (ed.), *Field and Service Robotics*. Berlin: Springer Verlag, pp. 12–16.

Stanford (2003) Stanford University Robotics website, http://robotics.stanford.edu/.

Stone, P., Balch, T. and Kraetzschmar, G. (2001) RoboCup 2000: Robot Soccer World Cup IV. Berlin: Springer Verlag.

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Sun, R. and Bookman, L. (1994) Computational Architectures Integrating Neural and Symbolic Processes: A Perspective on the State of the Art. New York: Kluwer. http://www.wkap.nl/prod/b/0-7923-9517-4.

SWARM (2003) Website, http://www.swarm.org/index.html.

Task (2003) DARPA 'TASK (Taskable agent software kit) program' website.{?}

Tarassenko, L. (2003) Oxford Neural Network Group website, http://www.eng.ox.ac.uk/World/Research/Summary/B-Neural.html.

Tesauro, G. (1994) TD-Gammon: a self-teaching backgammon program achieves master-level play. *Neural Computation*, 6 (2): 215–219.

Tesfatsion, L. (2002) Agent-based computational economics: growing economies from the bottom up. *Artif. Life*, 8 (1): 55–82.

Toner, J. and Tu, Y. (1998) Flocks, herds, and schools: a quantitative theory of flocking. *Phys. Rev. E*, 58 (4): 1998.

Walter, W.G. (1950) An imitation of life. *Sci. Am.*, 182 (5): 42–45.

Walter, W.G. (1951) A machine that learns. *Sci. Am.*, 185 (2): 60–63.

Webb, B. (2000) What does robotics offer animal behaviour? *Anim. Behav.*, 60: 545–558.

Webb, B. (2002) Robots in invertebrate neuroscience. *Nature*, 417: 359–363.

Webb, B. (2003) Can robots make good models of biological behaviour? *Behav. Brain Sci.*, 24 (6).

Widrow, B. and Hoff, M.E. (1960) Adaptive switching circuits. IRE WESCON convention Record, pp. 96–104.

Williamson, M.M. (1999) Robot Arm Control Exploiting Natural Dynamics, PhD Thesis, MIT Department of Electrical Engineering and Computer Science.

Wright, W.A. (2001) Sequential strategy for learning multi-stage multi-agent collaborative games. *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, pp. 874–884.

Wright, W.A., Smith, R.E., Danek, M.A. and Greenway, P. (2000)A generalisable measure of self-organisation and emergence. *Proc. Int. Conf. Artif. Neural Networks (ICANN)*, pp. 857–864.

York (2003a) York 'Bio-Inspired Engineering Group' website, http://www.elec.york.ac.uk/bio/bioIns/main.html.

York (2003b) York 'Advanced Computer Architectures Group' website, http://www.cs.york.ac.uk/arch/NeuralNetworks/index.htm.

Yovits, M.C., Jacobi, G.T. and Goldstein, G.D. (1962) *Self-organizing Systems*. Washington, DC: Spartan Books.

8 APPENDIX: KEY QUESTIONS

Cognitive function	Issue	5-year horizon	20-year horizon
Circuits and systems	Dynamics	Understanding of the relationship between dynamics in the brain and in computer systems	A general acceptance of these issues into mainstream computer systems design.
	Self-organization	Understanding of how computer systems may self-organize to be robust to failure, adapt to other functions etc.	Systems in everyday use that exploit these properties
	Synchronization	Better theoretical understanding of how neural-like computer systems are able to operate without a clock	-
	Processing speed	Better understanding of how new computers can use slow processing elements yet compute faster than we do now.	
	Timing	Merging of ideas between cognitive science and computer science on how we are able to achieve real time behaviour	-
	Robustness	Theoretical understanding of how computer science may exploit the brain's methods for more robust systems	
	Information representation in transmission	New ideas on how computer systems may exploit biological communication methods to achieve more effective operation	
	Construction and microcircuitry	Acceptance that neuromorphic and neural inspired systems can benefit the design of new computer systems	
Distributed computing systems	Complexity	Derivation of a theoretical basis for real large-scale distributed systems	
	Learning	Derivation of reinforcement learning methods for the developing strategies for multi-agent systems	Extensive use of multi-agent machine learning methods for the development of desired behaviours in large-scale systems

(*Continued*)

(*Continued*)

Cognitive function	Issue	5-year horizon	20-year horizon
	Bottom-up system design	Construction of framework for the development of bottom-up derived large-scale systems	Integration of a robust bottom-up design philosophy for large-scale distributed systems
	Behaviour	Derivation of a theoretical basis to explain the emergent behaviours in large-scale systems. This must allow for the presence of non-rational agents within the system	Development of a framework for the integration of human and automated decision makers in large-scale distributed systems
	Use of emergence in systems	Development of useful structures that make use of system's emergent properties	Understanding of what classes of behaviour may be derived by exploiting the emergent properties of a system
	Coping with uncertainty	Derivation of systems that work with uncertainty and limited communications	
Behaviour-based systems	Learning	Development of learning methods for individual agents that allow the production of global properties for the whole system	The ability to learn complex social and situational responses over broad time scales
	Architectures	What architectures, control, data and communication facilitate the derivation of multi-agent systems	The development of an autonomic network approach to multi-agent architectures.
	Utility of multi-agent system	Understanding of when and how advantage can be gained from the use of a multi-agent system in comparison to a single-agent system. Such a study would consider not only the theoretical advantage but also the relative computational and implementation costs of the different approaches	Role-based autonomic adaptation of multi-agent system utility responses

SECTION

2

Cognitive Systems in Touch with the World

Natural and artificial cognitive systems have to sense and respond to their environment.

Section Contents

This page intentionally left blank

Cognitive Systems in Touch with the World: Introduction

Lionel Tarassenko and Richard Morris

The human brain continuously processes data from the five senses: sight, hearing, smell, touch and taste. In the chapter on sensory processing (Chapter 4), Lionel Tarassenko and Mike Denham review the progress made in the past 20 years in the design of artificial cognitive systems that can process data from the first three of these senses. Most of their chapter describes the achievements of the traditional engineering or 'IT-centric' approach. This relies on building mathematical models of artificial sensory systems that lay no claim to biological plausibility or relevance. This approach has reaped real dividends when applied to object recognition, visual tracking or speaker-independent continuous speech recognition, for example.

These examples illustrate a fundamental paradox frequently encountered during the Foresight Cognitive Systems Project: the capabilities of artificial cognitive systems designed using the engineering approach still fall well short of those of natural systems. Yet artificial systems designed using the principles of biology (biomimetic systems) are mostly inferior to those based purely on the engineering approach. This is in part due to the *lacunae* in our neurobiological account.

Although we have developed a reasonably good understanding of lower-level processes in the human brain, for example, the processing of sounds by the cochlea, it is much harder to interpret what is happening at a higher level. How does the brain understand the meaning of a word? In their chapter on representation (Chapter 3), Vincent Walsh and Simon Laughlin address what happens at the different levels. They note that at the lower level, 'early processing', different senses employ common mechanisms. Visual, auditory and somatosensory cortex all contain fine spatial resolution maps with repeated small units. At the 'higher' levels of cortical organization, the units of organization change in functional complexity but the basic organization is preserved.

The original thesis that sensory representation can be understood as a feedforward or constructive process in which the details of the external world build up from simple features to more complex representations now needs to be modified as a result of the realization that we sense in order to act. The feedforward view is inadequate as it does not take into account knowledge acquired in recent physiological and perceptual studies about feedback mechanisms.

One of the key issues discussed by Walsh and Laughlin is that of cross-modal integration. Recent evidence from functional imaging studies of normal adults indicates that the senses do not interact solely within brain sites such as the superior colliculus, which receives inputs from more than one sensory

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

modality, but that they can also produce multimodal effects upon areas traditionally considered as unimodal. For example, the response in the visual cortex to a light can be boosted by concurrent touch at the same location in external space.

The same issue of cross-modal integration features prominently in the chapter on sensory processing written from a physical scientists' perspective. Here it goes under the name of 'sensor fusion' and is defined as the 'merger of multiple measurements of sets of parameters to form a single representation'. The engineering approach to sensor fusion, based on probabilistic techniques such as Kalman filtering or sequential Monte Carlo methods has been very successful but there remain many open questions, especially when dealing with fully distributed, as opposed to centralized, systems.

Walsh and Laughlin make the further point, with examples from invertebrate animals ranging from the nematode worm to the fruit fly, that the study of non-mammalian species offers many advantages. With smaller numbers of identifiable neurons, it is much easier to map the circuits, and to observe the transformation of signals in neurons. We can then relate the resulting processing of information to the generation of behaviour.

In their chapter, Tarassenko and Denham review the state of the art in computer vision ('sight'), automatic speech recognition ('hearing') and electronic noses ('smell'). Even if biomimetic sensory processing systems do not yet attain the performance levels of their IT-centric counterparts, biology has often provided the inspiration, if nothing else, for the design of the latter. The development of edge detectors in image processing, for example, was largely based on similar pre-processors found in the visual systems of animals. A more fundamental inspiration, however, is the *existence proof;* without the evidence that the human brain *can* form a 3D representation of the world from its visual sensors, it is doubtful that researchers in artificial cognitive systems would have attempted to do the same.

Animals from ants to humans can home in on an unseen object, even if they have moved since they last observed the object. In order to home in on a remembered object, the brain must associate a representation of its 3D position with the object and update this representation as the head and body move. We do not have a complete model of how this is achieved, even in insects, and certainly not in humans. On the other hand, we have recently seen great advances in the design of artificial systems which can *emulate* this ability and the interaction between computer scientists and neuroscientists working in vision is becoming more productive all the time.

Natural dialogue is the *leitmotiv* for researchers working on speech and language. The chapter with this title (Chapter 5), written by William Marslen-Wilson, argues that the ability to communicate with natural (human) cognitive systems should be a desirable feature of all artificial cognitive systems. Such a goal is driving researchers in the field of automatic speech recognition to examine recent advances in our knowledge of neurobiology, as the best of today's automatic speech recognition systems still struggle to handle continuous spontaneous speech where language is used in its natural context.

There are strong echoes of the chapter on representation in Marslen-Wilson's review of the latest developments in the cognitive neuroscience account of speech and language. He notes that the field of linguistics has generated a wealth of hypotheses about representations and processes at all levels, ranging from views on the nature of phonological representation and how they are accessed from the speech stream, to speculation about the organization of the mental lexicon and about the nature and representation of linguistic and conceptual meaning.

Sometimes, the hypotheses are in conflict with each other. Nowhere is this better illustrated than in the argument about the nature of the functional system which relates the auditory information to the stored knowledge of words in the language. Here is one of the key battlegrounds between traditional symbolic artificial intelligence (AI) and connectionism (sometimes also known as parallel distributed processing). The proponents of the former see the form and meaning associated with a given word as being represented in terms of labelled symbolic nodes. Connectionists, on the other hand, consider that the mapping from form to content emerges as a function of the learned pattern of connections among the elementary units in a multi-layer neural network. Marslen-Wilson's view is that additional constraints are needed to converge on a unique scientific account.

Here again, recent advances in non-invasive imaging may help to provide some of these constraints, as the results from these techniques begin to allow us to build a picture of the neural and functional organization of the human auditory cortex. One fascinating clue provided by a very recent fMRI study is the evidence for multiple processing streams, emerging in parallel from primary auditory cortex and from surrounding areas specialized for the processing of complex auditory inputs such as speech.

Marslen-Wilson ends with a list of open questions facing researchers in the field of speech and language. In some areas, there will be significant progress only if neuroscientists, psychologists, physical scientists and computer scientists join forces. One aim of this book is to promote such multidisciplinary research. The benefits of such an approach in the context of speech and language might be:

- greater scientific understanding of human cognition and communication
- significant advances in noise-robust speech recognition, understanding and generation technology
- dialogue systems that can adapt to their users and learn on-line
- improved treatment and rehabilitation of disorders in language function, including novel language prostheses.

It is clear from this list, and again it is a generic theme in this book, that the benefits will be spread across both the life sciences and physical sciences.

CHAPTER

3

Representation

Vincent Walsh and Simon Laughlin

In this chapter we highlight developments at the centre of research in sensory and cognitive processing in the biological and cognitive sciences. Our aim is to alert those working in the computational and engineering sciences to aspects of sensory representation that could be most relevant to developing artificial systems. To this end, we emphasize research and theoretical approaches that have yet to reach the textbook canon in the cognitive sciences. Physical scientists interested in modelling the senses tend already to be *au fait* with research on, for example, depth perception, object recognition and colour segmentation. We perceive a need for greater integration in the domains of the development of sensory representations, the response of neural systems to damage, the integration of information from more than one sense and the transformation of sensory inputs to representations useful for action. We also emphasize developments in understanding so-called simple systems, such as the fly brain and the vertebrate retina. It is here that tractable proposals for applications may emerge first, followed by a convergence of computation and cognition. Further, considering 'simple systems'

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

introduces constraints on modelling – in particular, a detailed knowledge of neural circuitry – that are not present when considering pure cognitive factors.

1 INTRODUCTION: THEORIES AND APPROACHES

In the domain of sensory representation, the disciplines of sensory and cognitive neuroscience have amassed what appears to be a great deal of knowledge. Detailed maps of visual, auditory and somatosensory regions of cortex exist for several species. The similarities between these maps suggest that the different senses employ common mechanisms. For example, the primary visual cortex contains a detailed retinotopic spatial map of the visual scene. To some degree there is selection for visual attributes in subregions of the visual cortex and the maps are iterative – i.e., based on small units of organization such as orientation – or direction-selective columns. Other primary sensory maps also exhibit these principles of organization. Auditory and somatosensory cortex both contain fine spatial resolution maps with repeated small units.

At other levels of cortical organization, sometimes called 'higher' levels, a problem we address below, the basic units of organization change in apparent functional complexity but the basic organization is preserved. For example, in extrastriate regions of cortex, the receptive fields of cells are larger, but still preserve retinotopic mapping and columnar architectures. At these and later levels, strict retinotopy does break down, but there is still some degree of organization by eccentricity.

The responses of cells in these regions, however, tend to be driven by more complex stimuli – specific shapes, faces etc. – than the responses of primary sensory cells. This forms the basis of what characterizes the dominant approach to understanding sensory processes. We can characterize this, at only a small risk of caricature, as follows:

> Sensory representation and signal processing can be understood as a feed forward or constructive process in which the details of the external world build up from simple features to more complex representations.

The principles of these constructive processes are likely to be similar across different modalities and it is thus sufficient to study a single sense in detail.

It has to be said that this approach has been immensely successful, as any of the literatures on vision, audition, tactile processing etc. attest. Scientists in the UK have been at the forefront of many of the most important advances. Yet many of the fundamental questions that are the spur for this research remain unanswered.

We sense in order to act. The global question is, how do the sensory systems provide an output for action? This poses several challenges for traditional approaches.

First, the output of sensory systems is not just a description of the sensory scene that corresponds to our experience of the scene. It is a description of the scene that is useful for action or prediction. Adding purpose to the intended outcome of signal processing raises many questions: Are objects of graspable size given primacy? If action is based on more than one sensory cue, do the senses cooperate or compete? If either, under what conditions? How does experience change the representation of the environment? Are the same brain regions involved in parsing a scene when one is naïve to the scene or task also involved when one is familiar with the scene or trained on the task? How do the consequences of the action affect signal processing?

The feed forward view is now clearly inadequate. There are many examples of critical feedback influences both from physiological and perceptual studies and several studies under the umbrella of attention. Key advances in recent years make these questions tractable. Some of the advances have come from studying sensory processing and cognition in humans or other mammals, others have relied on the study of smaller systems, such as insects and electric fish.

1.1 What are Neural Representations and Why are they Relevant to New Technology?

1.1.1 The Nature of Representation

For every unique percept, action and memory there should be a unique state of the brain. This state represents the information required to specify and generate these situations and events. The representation exists at a number of levels, from the patterns of activation of signalling molecules observed by molecular neurobiologists, to the patterns of activation of brain regions observed in functional imaging. No matter what its level, the representation could in principle take any one of a very large number of forms. This is because to discharge its function it need fulfil only two basic requirements, namely:

- The representation is unique and specific to events.
- The representation performs the tasks associated with these events.

1.1.2 What Factors Determine Representations and Why are these Factors Technologically Relevant?

A nervous system is made up of a large number of richly connected components (nerve cells = neurons). This rich connectivity of neurons offers many ways of mapping inputs onto outputs that specify events and actions. However, evolution has moulded the representations used by nervous systems through necessity, opportunity and efficiency. These three requirements draw the structure of neural processing closer to technology.

1.1.3 Necessity is What Brains have to do

The brain has evolved to collect information, process and store this information, and use the information to produce an output that increases, and perhaps even maximizes, the probability of a favourable outcome for the parent organism. Causal relationships – especially physical laws and *a priori* knowledge of what is to be expected in the normal environment – determine the information that has to be used for a given task, and the way it must be used. In other words, the nature of the real-world task defines the minimum set of measurements and operations required to execute that task to given specifications of speed, accuracy, cost etc.

As we shall see, the study of nervous systems suggests a variety of technologically relevant solutions to the problems of gathering, processing and outputting the information required for sensing, evaluation, control, decision-making, navigation and coordination. As technology advances, the number and range of these shared problems will increase: nervous systems will make increasing contributions to the development of quicker, cheaper and more intelligent uses of information.

1.1.4 Opportunity

Any application of technology is determined by the range of available devices. New devices create opportunities to improve design and expand capabilities. The opportunities offered by nerve cells differ from those offered by current information processing technologies. Nerve cells use signalling molecules – usually large protein molecules anchored in membranes, or freely diffusing low molecular weight messengers – to transmit and process chemical and electrical signals. Compared with electronic devices, nerve cells are slow (usually <1 kHz cf. >3 GHz in the Pentium 4 processors), unreliable (signal-to-noise ratio of <50:1) and weak (dynamic response range 0.15 V). However, nerve cells are flexible and adaptable. They make their own connections and assemble specific groups of signalling molecules at specific places at specific times to process and transmit information. More importantly, nerve cells can adapt their connections and molecular configurations in response to specific patterns of input and in response to the previous signalling history.

The nerve cell is, therefore, a versatile and adaptable self-assembling component with a rich and highly elastic repertoire of computational primitives. We can be pretty certain that nervous systems have exploited the versatility of nerve cells to process large quantities of information reliably enough to solve the many problems of inference and control that confront an animal. Thus, as suggested by von Neumann, the nervous system may well lead us to new methods for computation that depend more upon correlations and relationships than upon numerical accuracy.

1.1.5 Efficiency

Natural selection favours individuals who make better use of finite resources. Thus evolution promotes efficiency. We can see efficiency at three levels:

- the design and organization of devices that makes best use of available space and energy
- the efficient coding of information within the constraints of space and energy
- the efficient use of this information to produce rapid, reliable and appropriate outputs.

Thus, design principles gained from studying nervous systems and the behaviour that they generate are likely to be both effective and efficient.

2 KEY FINDINGS AND CONCEPTUAL ADVANCES OF THE PAST DECADE

There have been many advances over the past decade, both technical and conceptual. The technical advances have, to some extent, generated the headlines, but it is the conceptual advances that are the key to further progress. They have provided several novel views of sensory processing to present a new profile for sensory research.

2.1 Protocortex and Rerouting of Development

The view that common principles underlie processing in different sensory domains receives backing from the theoretical and anatomical studies supporting the idea of a unified proto-cortical plan from which all the primary areas develop (Dennis and O'Leary, 1989). Recently, a series of experiments by Sur and colleagues, showed that, in the developing ferret, fibres from the retina can be rerouted to the medial geniculate nucleus (MGN), an auditory structure. This auditory structure was disconnected from its inputs and as a consequence was then innervated by visual fibres. Subsequent electrophysiological recordings from the MGN revealed that the cells in this now newly wired auditory structure responded to visual input from the eyes. The next stage in the auditory pathway, A1, the primary auditory cortex also responded to visual inputs.

The organization of the rewired MGN was a compromise between the normal fate of the inputs and the normal structure of the target site. Thus, eye-specific populations of neurons were segregated in the MGN, but the size and organization of these elements was similar to that expected by normal inputs from the inferior colliculus.

There were other visual-dominated features of the now visual MGN. The cells were organized retinotopically and were monocular and not orientation-selective. The newly imposed visual structure extended to the auditory cortex (A1). In normal ferrets, A1 re-maps the cochlea, along which frequencies are encoded in a single dimension, but in the rewired animals there was a two-dimensional retinotopic map of the visual world: 'Visual cells in rewired A1 have orientation-tuning, direction tuning and velocity tuning indices that are indistinguishable from V1 cells' (Sur and Leamy, 2001: 258).

This work invites us to consider the extent to which the behavioural functions of an area are determined by afferent inputs during development and/or to intrinsic features. In other words, what does a rewired ferret see/hear? Behavioural studies of the perceptual abilities of rewired animals have shown convincingly that a rewired ferret's auditory cortex performs visual functions.

Sur has concluded: 'Everything the cortex knows about the external world is contained in the spatiotemporal activity of its afferents' (Sur and Leamy, 2001: 260).

2.2 Remnants of Protocortex or Real Interactions?

Even allowing for the equal potential of the sensory areas to develop, the end result is commonly treated as modular. Visual cortex is visual, auditory/auditory etc., but there are several reasons to question this as too simplified.

Relatively neglected work in the 1980s had already shown that when a monkey was responding to the *tactile* orientation of lines, neurons in the visual cortex gave orientation-selective responses (Haenny and Schiller, 1988; Haenny *et al.*, 1988). One interpretation is that intermodal interactions between vision and touch convey stimulus-specific information. More recent work has shown that interactions between vision and touch or vision and audition occur as early as V1. An alternative, but less likely, view is that these are mere intermodal minglings, vestigial from earlier stages of development.

The position that the interactions are meaningful is supported by the work of Shamma (2001), who has argued that the algorithms used in vision are but one example of 'a unified computational framework [that] exists for central auditory, visual and other sensory processing' (see also Statistics of the environment, below).

There are, of course, many examples of two senses being better than one. We perceive speech better if we can see the speaker's lips (Risberg and Lubker, 1978). Our tactile sensitivity is greater if we can see our fingers, and improved even if one is about to saccade to the fingers making judgements (Rorden *et al.*, 2002). We can visually locate an object more accurately if it is also a source of sound (King, 2002).

There is also good evidence that several brain structures are specialized to receive inputs from more than one sensory modality – the superior colliculus, the intraparietal sulcus and the superior temporal sulcus being key sites. In the superior colliculus, for example, the superficial layers devoted to vision are co-registered with the underlying deeper layers which are concerned with auditory and multisensory inputs.

Recent evidence of functional imaging from normal adults indicates that the senses do not interact solely within 'multimodal' brain sites, such as those mentioned above. The senses can also produce multimodal effects upon areas traditionally considered as unimodal. For example, the response to a light in the visual cortex can be boosted by concurrent touch at the same location in external space.

The most recent convincing demonstration of integration from early stages of cortical processing comes from Falchier *et al.* (2002). They traced projections from auditory cortex and a polysensory area of the temporal lobe (STP) to the primary visual cortex (V1). The central few degrees of V1 contain few inputs from auditory cortex but a non-negligible number (approx 5%) from the STP. Between 10 and 20 degrees, however, the auditory cortex contributes projections which amount to around 10% of the projections from V5 to V1 and STP contributes approx 35% on the same scale.

The functional corollary of this is that multisensory integration that relies on interactions between visual processes in V1 and information from other modalities will be enhanced in the peripheral rather than the central regions of the visual field. Candidate behaviours are reduced time taken to orient, decreased sensory thresholds and visual imagery (Klein *et al.*, 2000).

2.3 Cross-Modal Integration

2.3.1 Competition, Cooperation and Compensation

The question of interaction between modalities and brain regions raises the related question of when these interactions are cooperative or competitive.

Competition Rather than Cooperation?

The apparent outcome of sensory processing is coherent experience and coordinated action. However, at the level of single receptive fields, single sensory modules and the cerebral hemispheres, the outcome does not reflect the underlying mechanisms. Competition between stimulus inputs is now established at every level of analysis.

Desimone and colleagues (see Desimone 1998 for a review) have established stimulus based competition in the analysis of receptive field profiles of neurons in extrastriate cortex. Walsh and colleagues have established competitive interactions between different visual areas at the same level of the processing hierarchy. Interactions between the hemispheres have long been examined.

Compensation

There is evidence that tactile stimuli can activate the visual cortex of blind people (Wanet-Defalque *et al.*, 1988; Uhl *et al.*, 1991; Rauscheker, 1995; Sadato *et al.*, 1996). The question is whether this activity has a function.

We know from physiological studies that visual cortical areas respond to tactile orientation (Haenny and Schiller, 1988; Haenny *et al.*, 1988), but until recently it was not clear whether visual cortex in human subjects could be shown to be necessary for tactile discrimination (Zangaladze *et al.*, 1999). We now know that under some circumstances visual cortex is an aid to tactile perception. One view of reorganization in blind subjects is as unmasking or strengthening previously extant connections and responses.

Cohen *et al.* (1997) established the relevance of visual cortex for Braille reading. They disrupted occipital cortex processing in blind subjects who were given the task of identifying Braille characters or embossed Roman letters. Interference, caused by brief magnetic pulses, called transcranial magnetic stimulation (TMS), disrupted tactile performance in the blind subjects but not in sighted controls. In this study, stimulation of the somatosensory cortex did not impair tactile discrimination performance in blind subjects. One explanation of this phenomenon is that the effects of mid-occipital TMS are related to 'interference with more complex discriminative operations performed by occipital cortex in the blind' (1997: 182). This leaves open the problem of how to explain the earlier demonstration of tactile disruption in blind subjects following stimulation of somatosensory or motor cortex (Pascual-Leone and Torres, 1993; Pascual-Leone *et al.*, 1995) and with the suggestion that the occipital activity may partly explain the superior tactile abilities of blind subjects. Supporting evidence comes from brain imaging studies in which the visual cortex of early-blind individuals was active during auditory localization and also during Braille reading.

2.4 Learning

Strategies in perception and action change with increasing practice or experience. We are familiar with the reports that deficiencies in one sense are associated with compensatory improvements in another. An examination of these reports showed that increases in the areal representations in the somatosensory cortex as a function of the frequency of using body parts could be established as functionally relevant or epiphenomenal.

Blind subjects who could read Braille and sighted subjects who could not were given a tactile detection task. They received TMS over sites in the somatosensory cortex where electrical stimulation of the index finger had evoked potentials. The subjects experienced single-pulse TMS, applied 50 ms after the electrical pulse was delivered to the finger. TMS over the somatosensory cortex impeded detection of tactile stimulation over a three-fold greater area of the scalp in the blind group. There was also a difference between the dominant and non-dominant hands of the Braille readers. TMS over Braille-dominant hands disrupted tactile thresholds over twice as many scalp locations as the non-dominant hand of the same subjects.

This experiment might allow one to conclude that, in the case of blind Braille readers, the change in the somatosensory representation was a consequence of the differential sensory input between the Braille readers' fingers and the sighted subjects', and between the two hands of the Braille readers. Another possibility is that the effects of somatosensory TMS were due to an expansion of the motor cortex due to the repeated finger movements made in reading Braille.

The plasticity observed in the Braille subjects does not mark the end of the reorganization. Plasticity is not a special case of perception: rather, it is the normal state of the nervous system. Any reorganization due to amputation or blindness would be pointless, if not maladaptive, as in cases of phantom pain, if the new map could not constantly change with the demands of behaviour.

Evidence of the plasticity of expanded representations of motor areas was seen in a group of blind subjects, all of whom became blind before the age of 10 and learned to read Braille before the age of 13 (Pascual-Leone *et al.*, 1995). Motor evoked potentials (MEPs) were recorded from the first dorsal interosseous (FDI) of both hands and the abductor digiti minimi (ADM), not used for Braille, of the Braille-dominant hand. The subjects read Braille for up to six hours a day at work, but MEP amplitudes diminished markedly after 10 days of vacation without much Braille activity. Just one week back at work reinstated the increased amplitude and the number of scalp locations from which a TMS induced MEP could be elicited. Shorter-term changes in the motor maps were also observed. The scalp area from which an MEP could be elicited from the FDI increased in size and sensitivity during the working day, but there were no changes on rest days or in the ADM of the Braille-dominant hand.

2.5 Reverse Hierarchy Theory

The dominant approach to explaining the physiology of vision has been, and is, bottom-up. Small receptive fields pass on information for further analysis to the higher, secondary visual areas. This has been challenged on two levels, theoretical and anatomical. The weight of evidence for this view depends partly on the kinds of questions that have been asked of V1 neurons. When illusory contours were considered, and indeed termed 'cognitive contours', it was presumed that the perception was generated in a cognitive, higher area such as inferotemporal cortex. However, subsequent studies of V2 (Peterhans and von der Heydt, 1991) and later V1 neurons (Grosof *et al.*, 1993) showed that these lower areas contained the necessary architecture to retrieve the form of illusory contours. Similarly, because attention, in one of its many flexible guises, is considered to be dominated by inferotemporal and parietal cortices, few experiments set out to assess whether a V1 or an extrastriate neuron changed its responses to stimuli depending on behavioural relevance. Researchers who did probe these and other questions aimed at reassessing the role of these areas in visual processing consistently found that V1 neurons were indeed sensitive to the context of a visual scene (Zipser *et al.*, 1996; Nothdurft *et al.*, 1999), showed responses that can be described as attentional (Motter, 1993; Somers *et al.*, 1999), and respond at times later than the mean latencies for secondary visual areas (Bullier, 2001). Further, an influential new theory of vision provides principled reasons for reconsidering the role of V1 in vision (Ahissar and Hochstein, 2000).

In their reverse hierarchy theory of vision, Ahissar and Hochstein propose that visual processing follows a global-to-local trajectory in which extrastriate neurons with large RFs carry out an initial coarse grained analysis of the visual field followed by a more detailed analysis in earlier visual areas. The predictions of this theory include that easy tasks are learned at higher levels of cortex, while harder tasks demand the resolution of primary visual cortex. The anatomical version of this theory (Bullier, 2001) also emphasizes global-to-local operations and describes V1 and V2 as 'active blackboards' for other cortical areas.

3 STATISTICS OF NATURAL SIGNALS

Much of sensory science is based on the analysis of responses of neurons or subjects to simple visual stimuli such as gratings, bars or spots of light. This precludes understanding the way in which sensory systems use global information in the environment. To remedy this deficiency, physiologists and computational neuroscientists have resurrected the approach of understanding the processing of natural images. It is difficult to overestimate the impact of this approach. The environment presents many regularities to the sensory and action systems from luminance to language.

A major advance, promoted by cheaper instrumentation and computing power, is to measure these regularities and describe them mathematically, usually as statistical distributions of signal along relevant stimulus dimensions such as intensity, wavelength, spatial location, time and so on. Such natural distributions are used in two ways. The first is to describe how the nervous systems responds to natural signals. Physiologists commonly measure transfer functions that describe, mathematically, how signals change as they pass through neurons. By applying these transfer functions to a natural input distribution we derive the natural output distribution. This shows how the nervous system handles the signals it has evolved to process, but does not necessarily tell us why.

The purposes of neural transformations are explored by the second approach. This takes the mathematical description of the input and, by applying appropriate theory (e.g. Information Theory), derives an optimum coding procedure that maximizes desirable attributes of the output within constraints. If neural processing approaches this theoretical optimum, we have established a plausible function for it. The desirable attributes of the output we have considered so far are information content, completeness and the salience of important features. The constraints include the limited signal capacity of neurons, and the energy required to generate neural signals. The optimum codes are derived analytically, using appropriate theory and empirically, by training networks.

These applications of natural signal statistics have revealed two types of neural code. One maximizes the information coded per neuron, within the constraint of noise and bandwidth. It is implemented in invertebrate and vertebrate retinas by filters that adapt to reduce noise at low levels and remove redundancy at high light levels (Atick, 1992; van Hateren, 1992). The other, sparse coding, tends to maximize the amount of information coded per signalling event – e.g. per action potential or per excited cell – by activating a small proportion of a large population of otherwise inactive neurons. Information is coded by the combination of neurons that are activated. The sparseness of activity increases the salience of features.

Experiment and theory show that the receptive fields of neurons in primary visual cortex (V1) generate sparse representations (Simoncelli and Olshausen, 2001). Because the efficiency of representation depends upon correlations in the input, these receptive fields resemble filters derived from the statistical techniques of principal components analysis and independent components analysis.

The analysis of sensory signal statistics and coding efficiency is contributing important techniques and concepts to our understanding of coding and representation at all levels. The statistical analysis of signal quality provides powerful tools for describing and analysing the quality of neural representations. These methods will help us to understand what is gained and what is lost when neurons transform inputs to generate new representations.

The use of constraints demonstrates that the limited ability of neurons to code and transform signals has a profound effect on neural function. To be effective, neural processing must play to its strengths (nonlinearity and plasticity) and avoid exposing its weaknesses (poor time resolution, low signal-to-noise ratio). The most obvious contribution will be of most lasting importance. The success of the natural signal statistics approach emphasizes the value of analysing biologically relevant tasks.

Signal statistics is rekindling interest in one of the most basic forms of plasticity in neural systems, adaptation. Simply defined as a change in response to a constant signal, adaptation turns out to be much more than a gain control or differentiator. In the visual system of the fly and visual and auditory cortex of mammals , adaptation optimally allocates the limited number of neural signalling states to inputs, by dynamically rescaling signals to take into account changes in signal distribution (Fairhall *et al.*, 2001). Adaptation also reduces redundancy by removing predictable components from neural signals (Schwartz and Simoncelli, 2001). Thus adaptation not only neatly packages signal in channels of limited capacity, but transforms signals according to their predictability.

Perhaps the most important contribution made by natural stimulus statistics will be to consolidate the primacy of the statistical view of neural function. Predictability aligns our understanding of sensory codes (i.e. representations) with the ultimate purpose of the nervous system, namely the identification of regularities and the association of these regularities with favourable outcomes via appropriate action. An increasing number of investigators are using natural stimulus statistics to implement a Bayesian approach to sensory coding. This approach is starting to bridge the gap between lower level operations – adaptation, filtering and coding – to the higher level functions of inference and learning.

To emphasize this point, we note that the consequences and generality of the statistical approach extend to functions that would not, at first sight, be thought to depend upon the structure of the environment. Saffran and colleagues (1996), for example, have shown that infants as young as 8 months use patterns in auditory inputs as an aid to learning language.

Kirkham and colleagues have extended this work to 2-month olds. Infants were habituated to sequences of stimuli that followed a statistically predictable pattern. When presented with novel sequences, these were preferred by even the youngest subjects. The authors suggest that this approach to learning 'is consistent with the thesis that early development is highly attuned to the multifaceted structure of the infant's environment, and suggest that learning the statistical regularities of the environment may be a critical part of the cognitive apparatus with which infants make sense of the world' (2002: (B35)).

4 SIMPLE SYSTEMS

The advances we discussed above represent work carried out in mammals, including humans. The study of non-mammalian species offers many advantages, and has led to recent advances, when trying to describe how the features of the environment and the mode of representation are successfully translated into action.

4.1 Background

It is difficult to define and analyse the neural circuits that ultimately process information in large and complicated brains. It is more practical to analyse this fundamental structure in simpler nervous systems. With smaller numbers of individually identifiable neurons, we can map out the circuits, observe the transformation of signals in neurons, and relate the resulting processing of information to the generation of behaviour. For this straightforward technical reason, the analysis of simple systems continues to establish fundamental principles of operation and design of neural circuits. Because many of the current generation of simple systems are from invertebrates, we should emphasize that the defining characteristics of simplicity are the tractability and completeness of the system, not its position in the animal kingdom.

As techniques for unravelling neural circuits and describing behaviour improve, our knowledge will increase so that more vertebrate systems will fulfil the criteria established for simple systems (Marder, 2002). Productive systems were developed over 20 years ago for studies of sensorimotor integration, motor pattern generation, neuromodulation and cellular mechanisms of learning. These include:

- vertebrate retinas (early visual coding)
- insect (especially fly) visual systems (early visual coding, pattern recognition, motion detection, colour discrimination and learning, visual navigation)
- orthopteran (e.g. cricket) auditory systems (auditory coding, pattern recognition, song production)
- the central nervous systems (CNS) of arthropods, molluscs, leeches, tadpoles and lampreys (sensorimotor integration, motor pattern generation, learning).

Because the analysis produces definitive unambiguous descriptions of what neurons do, work on these systems continues to expose new problems.

Researchers are also developing promising new systems. These include: the nematode worm *Caenorhabditis* and the fruit fly *Drosophila*, which are favourable for molecular, genetic and developmental studies; insect olfactory systems; and a directional hearing system in a parasitic fly that localizes targets with an accuracy approaching 1 degree using a pair of sensors separated by <0.5 mm (Mason *et al.*, 2001; Robert and Gopfert, 2002). A major advantage of studying simple systems is that, with limited resources, they have evolved simplified 'quick and cheerful' solutions to apparently complicated problems of pattern recognition and control.

Several topics with widespread applications in neuroscience and technology are emerging.

4.2 Sensorimotor Integration

As we observed for cortex, the nervous system represents information for a purpose, action. This must influence the representation of sensory inputs. However, beyond some obvious adaptations to specialized signals and behaviours – for example, classic work on reflexes and eye movements in bats have reduced vision – this is eye reflexes and eye movements in general – we know remarkably little about the level at which motor factors start to operate in the nervous system.

When it passes from sensors to effectors, information is transformed from sensory coordinates to motor coordinates. Work on insects is starting to identify and analyse these transformations. This is possible because the sensory inputs, the intermediate neurons and the motor outputs are exceptionally well described. Several groups combine studies of circuitry, signalling and computation to describe representations by ensembles of neurons, to define coordinate systems, to determine why these representations are valuable for guiding behaviour, and to see if they are modified by motor context. Promising systems are as follows:

- In the CNS of the locust, mechanosensory inputs produce reflex motor responses and modulate motor output to control walking, swimming and flight. For the scratch reflex, positional information mapped by an array of mechanoreceptors on the wing, is converted into a motor pattern that guides the locust's hind foot to the point of stimulation (Matheson, 1998).
- In cricket auditory communication systems motor context (the generation of song) regulates sensory processing, providing an exceptional opportunity to study an important organizational principle, efference copy and the interplay between sound reception and production at the circuit level (Poulet and Hedwig, 2002).
- The analysis of the multisensory control of insect flight is developing rapidly in the UK (Krapp, Rind, Simmons), Germany (Egelhaaf, Borst), the United States (Dickinson) and Australia (Srinivasan) because of its potential for neuroscience, robotics and flight control (Frye and Dickinson, 2001).

The aim of this work is to establish the engineering principles underlying stabilization and the manner in which they are executed by sensors and circuits. These studies look at systems in their totality by combining the analysis of sensors, sensory signal processing, sensory integration, motor control and motor response within the context of animals moving freely in natural settings.

Opportunities for such complete syntheses are rare in neuroscience. They are likely, therefore, to yield new principles. Because these discoveries will show how physical components are organised to function, they relate directly to the construction of practical devices.

4.3 Coding

Simple systems are small and neurons are slow, unreliable and produce weak signals. So by studying them we can learn how systems have evolved to optimize representations within severe constraints of physical size, bandwidth, signal-to-noise ratio and, a related constraint, energy consumption. This constraint-based approach provides compelling accounts of retinal function that are being incorporated in neuromorphic VLSI 'silicon retinas'.

Laurent's work on olfactory coding by neural circuits in insects (bees and locusts) and vertebrates (zebra fish and rats) is an especially exciting development (Laurent, 2002). Experiment and theory suggest a new general coding scheme for handling signals that are distributed in a space of high dimensionality – those receptors among 100s or 1000s of different types that are activated by an odour. Electrical oscillations indicate interactions that could group cells to provide a compact representation. This representation is established more quickly when the odour is smelt again. Laurent's findings are drawing into the realms of pattern recognition and learning the analysis of low-level sensory coding by tractable networks of identifiable neurons.

4.4 Pattern Recognition and Learning

The visual circuitry for determining platform rotation and translation from optic flow – the trajectory of looming objects and reactions to surfaces (e.g. landing, obstacle avoidance) – is being analysed in insects and crabs. Retinotopic interactions extract looming edges (Rind and Simmons, 1999), matched filters are used for pattern retrieval from optic flow (Krapp *et al.*, 1998), and work is starting on adapting filters that optimize retrieval by taking into account the prevailing signal statistics (Harris *et al.*, 2000; Fairhall *et al.*, 2001). These studies are closely tied to the development of VLSI electronic devices for vehicle control (Rind). A classic analysis of the motion cues that flies and bees use for flight stabilization, landing and obstacle avoidance forms the basis for the development of small autonomous flying machines (Chahl *et al.*, 2003). Insect inspired prototypes are flying.

Improvements in the definition and analysis of signals and circuits in crickets promises a number of simple algorithms for auditory localization and pattern recognition Work on olfaction., together with studies of adaptation in pattern recognition networks, provides a bridge to the establishment of patterns of causality by learning. In addition to the classic work on the cellular substrate of learning in the mollusc *Aplysia*, recent advances in optical imaging and electrical recording in the brain of worker bees (Faber *et al.*, 1999) indicate how cells represent complicated patterns so that they can be associated with specific events.

4.5 Navigation

Autonomous vehicles, be they animals or robots, need to establish their location and move to objects that they have previously encountered. Behavioural and modelling studies of Hymenoptera (ants, bees and wasps) reveal how these animals use landmarks to establish and memorize locations, and how they combine visual cues (e.g. movements across terrain, sky compass) and motor cues to compute their track and establish and maintain headings (Collett and Collett, 2002). These studies benefit from simulations with robots (Webb, 2002) and animats (Dale and Collett, 2001). We are on the threshold of identifying neural mechanisms. Work on locusts and *Drosophila* suggests that circuits in the central complex of the insect brain are involved (Strauss, 2002).

5 TECHNICAL DEVELOPMENTS

5.1 Experimental

As it becomes increasingly apparent that representation depends critically upon the distribution of signals among cells, research is giving high priority to simultaneous recordings from identified neurons. Optical imaging techniques, using new Ca dyes and voltage sensitive dyes, easily resolve slow signals (>100 ms) but currently struggle to resolve individual action potentials. We need smaller, steerable and more tenacious recording probes. Developments in nanofabrication could, over the next 10 years, offer solutions.

Powerful and relatively inexpensive methods for recording, analysing and generating large data sets are revolutionizing research on simple systems. Complete descriptions of sensory inputs (natural scene statistics) and motor responses (flight trajectories, limb movements) define the full range of signals and, from statistical relationships, their structure. These measurements are complemented by statistical techniques for establishing the information content of neural signals.

When examining systems that have evolved for action, the action must be as real as possible. Techniques are being developed to monitor an animal's movements accurately through naturalistic surroundings, to reconstruct the sensory signals so generated, and to replay these to restrained animals while recording neural activity.

Given that neural systems process sensory information in a motor context we must strive to record signals from identified neurons in animals that move freely. Micro chips have recently been developed to record from flying moths. These devices, which weigh 50 mg, contain a recording probe, amplifier, transmitter and power supply. Smaller, lighter devices with steerable probes are required and developments in nanotechnology are eagerly anticipated.

5.2 Theoretical

Powerful and accessible simulation and modelling packages are invaluable for establishing that measured interactions can truly account for measured behaviour. The simulation of the purposeful behaviour of simple systems with robots (Webb) is particularly promising because these physical implementations expose practical requirements that are not always apparent in either the theory or the neurobiological data. As we discussed above, simple systems have contributed significantly to our theoretical and practical understanding of neural codes. This is, in turn, feeding back into the formulation of more appropriate hypotheses for coding and information processing and better empirical descriptions of the involvement of the underlying neural mechanisms.

6 FUTURE DIRECTIONS

The nervous system presents a wide variety of technologically relevant solutions to the problems of gathering, processing and outputting the information required for sensing, evaluation, control, decision-making, navigation and coordination. As technology advances, the number and range of the problems shared by engineering and the life sciences will increase. As a consequence, nervous systems will make increasing contributions to the development of quicker, cheaper and more intelligent uses of information.

The goal of artificial systems may not be to mimic biological mechanisms: in some cases the outcome may be more important than mechanism. However, as the behaviours to be modelled increase in complexity, the probability will increase that biological solutions will offer the most useful interface between technology and human users.

We can be reasonably confident that nervous systems have exploited the versatility of nerve cells in order to process large quantities of information reliably enough to solve the many problems of inference and control that confront a behaving animal. Thus, as von Neumann suggested, the nervous system may well lead us to new computation methods that depend more upon correlations and relationships than upon numerical accuracy.

This chapter has described some of the key considerations in developing converging goals between cognitive and physical scientists.

- Although humans are sight-dominated animals, it is clear that our senses act in concert rather than in turn. We need to go beyond the limited descriptions of perception and action that we can obtain by single modality studies.
- A key goal is to understand the dynamic nature of sensory representations that can change during development – as a response to injury or deprivation and as a response to the changing requirements of the world (learning). Much of what we know describes the behaviour of a subject required to make a simple response in an impoverished and unfamiliar environment. Indeed, readiness to meet change, rather than building a representation, might be considered the normal state of the sensory systems.
- Anatomical, physiological and theoretical studies now support the breakdown of hierarchical views of perception.

New initiatives await developments in bridging these three themes. For example, a reverse hierarchy approach may be as artificial as the original bottom-up approach – which cells dominate may be determined anew with changes in the tasks demanded and the information available.

6.1 Cross-Fertilization with Computer Scientists

There are rich pickings here. The study of developing neural systems, rewiring learning, co-registration of maps, are all problems not really begun to be exploited in interactions between biologists and computational scientists. The aims and even the language of the cognitive neuroscience and computational communities have converged greatly in recent years. Both groups ask such key questions as: How does a behaving system combine sensory information? What are the constraints of development? How are *intra*-modal features combined?

Bibliography

Ahissar, M. and Hochstein, S. (2000) The spread of attention and learning in feature search: effects of target distribution and task difficulty. *Vision Res.*, 40: 1349–1364.

Atick, J.J. (1992) Could information-theory provide an ecological theory of sensory processing? *Network*, 3: 213–251.

Baddeley, R., Abbott, L.F., Booth, M.C.A., Sengpiel, F., Freeman, T., Wakeman, E.A. and Rolls, E.T. (1997) Responses or neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lon. B*, 264: 1775–1783.

Baddeley, R.J. and Hancock, P.J.B. (1991) A statistical analysis of natural images matches psychophysically derived orientation tuning curves. *Proc. R. Soc. Lon. B*, 246: 219–223.

Bullier, J. (2001) Integrated model of visual processing. *Brain Res. Brain Res. Rev.*, 36: 96–107.

Chahl, J., Thakoor, S., Le Bouffant, N., Stange, G., Srinivasan, M.V., Hine, B. and Zornetzer, S. (2003) Bioinspired engineering of exploration systems: a horizon sensor/attitude reference system based on the dragonfly ocelli for mars exploration applications. *J. Robotic Syst.*, 20: 35–42.

Cohen, L.G., Celnik, P., Pascual-Leone, A., Corwell, B., Falz, L., Dambrosia, J., Honda, M., Sadato, N., Gerloff, C., Catala, M.D. and

Hallett, M. (1997) Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389: 180–183.

Collett, T.S. and Collett, M. (2002) Memory use in insect visual navigation. *Nature Rev. Neurosci.*, 3: 542–552.

Dale, K. and Collett, T.S. (2001) Using artificial evolution and selection to model insect navigation. *Curr. Biol.*, 11: 1305–1316.

Dennis, D. and O'Leary, M. (1989) Do cortical areas emerge from a protocortex? *Trends Neurosci.*, 12: 400–406.

Desimone, R. (1998) Visual attention mediated by biased competition in extrastriate visual cortex. *Phil. Trans. R. Soc. Lond. B*, 353: 1245–1255.

Faber, T., Joerges, J. and Menzel, R. (1999) Associative learning modifies neural representations of odors in the insect brain. *Nature Neurosci.*, 2: 74–78.

Fairhall, A.L., Lewen, G.D., Bialek, W. and de Ruyter Van Steveninck, R.R. (2001) Efficiency and ambiguity in an adaptive neural code. *Nature*, 412: 787–792.

Falchier, A., Clavagnier, S., Barone, P. and Kennedy, H. (2002) Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci.*, 22 (13): 5748–5759.

Frye, M.A. and Dickinson, M.H. (2001) Fly flight: a model for the neural control of complex behavior. *Neuron*, 32: 385–388.

Grosof, D.H., Shapley, R.M. and Hawken, M.J. (1993) Macaque V1 neurons can signal illusory contours. *Nature*, 365: 550–552.

Haenny, P.E., Maunsell, J.H.R. and Schiller, P.H. (1988) State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exper. Brain Res.*, 69: 245–259.

Haenny, P.E. and Schiller, P.H. (1988) State dependent activity in monkey visual cortex. 1. Single cell activity in V1 and V4 on visual tasks. *Exper. Brain Res.*, 69: 225–244.

Harris, L.R., Blakemore, C. and Donaghy, M. (1980) Integration of visual and auditory space in the mammalian superior colliculus. *Nature*, 288: 56–59.

Harris, R.A., O'Carroll, D.C. and Laughlin, S.B. (2000) Contrast gain reduction in fly motion adaptation. *Neuron*, 28: 595–606.

King, A.J. (2002) Neural plasticity: how the eye tells the brain about sound. *Current Biol.*, 12: R393–R395.

Kirkham, N.Z., Slemmer, J.A. and Johnson, S.P. (2002) Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83: B35–B42.

Klein, I., Paradis, A.L., Poline, J.B., Kosslyn, S.M. and Le Bihan, D. (2000) Transient activity on the human calcarine cortex during visual-mental imagery: and event related fMRI study. *J. Cogn. Neurosci.*, 12: 15–23.

Kosslyn, S.M., Pascual Leone, A., Felician, O., Camposano, S., Keenan, J.P., Thompson, W.I., Ganis, G., Sukel, K.E. and Alpert, N.M. (1999) The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science*, 284: 167–170.

Krapp, H.G., Hengstenberg, B. and Hengstenberg, R. (1998) Dendritic structure and receptive-field organization of optic flow processing interneurons in the fly. *J. Neurophysiol.*, 79: 1902–1917.

Kujala, T., Alho, K. and Näätänen, R. (2000) Cross-modal reorganization of human cortical functions. *Trends Neurosci.*, 23 (3): 115–119.

Lamme, V.A. and Spekreijse, H. (2000) Modulations of primary visual cortex activity representing attentive and conscious scene perception. *Front Biosci.*, 5: D232–D243.

Laurent, G. (2002) Olfactory network dynamics and the coding of multidimensional signals. *Nature Rev. Neurosci.*, 3: 884–895.

Lee, T.S., Mumford, D., Romero, R. and Lamme, V.A. (1998) The role of the primary visual cortex in higher level vision. *Vision Res.*, 38: 2429–2454.

Motter, B.C. (1993) Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.*, 70: 909–919.

Marder, E. (2002) Non-mammalian models for studying neural development and function. *Nature*, 417: 318–321.

Mason, A.C., Oshinsky, M.L. and Hoy, R.R. (2001) Hyperacute directional hearing in a microscale auditory system. *Nature*, 410: 686–690.

Matheson, T. (1998) Contralateral coordination and retargeting of limb movements during scratching in the locust. *J. Exp. Biol.*, 201: 2021–2032.

Nothdurft, H.C., Gallant, J.L. and Van Essen, D.C. (1999) Response modulation by texture surround in primate area V1: correlates of 'popout' under anesthesia. *Vis. Neurosci.*, 16: 15–34.

Olshausen, B. and Field, D. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381: 607–609.

Pascual-Leone, A. and Torres, F. (1993) Plasticity of the sensorimotor cortex representation of the reading finger in Braille readers. *Brain*, 116, Pt 1: 39–52.

Pascual-Leone, A., Wassermann, E.M., Sadato, N. and Hallett, M. (1995) The role of reading activity on the modulation of motor cortical outputs to the reading hand in Braille readers. *Ann. Neurol.*, 38 (6): 910–915.

Peterhans, E. and von der Heydt, R. (1991) Subjective contours – bridging the gap between psychophysics and physiology. *Trends Neurosci.*, 14: 112–119.

Pons, T.P., Garraghty, P.E., Ommaya, A.K., Kaas, J.H., Taub, E. and Mishkin, M. (1991) Massive cortical reorganisation after sensory deafferentation in adult macaques. *Science*, 252: 1857–1860.

Poulet, J.F. and Hedwig, B. (2002) A corollary discharge maintains auditory sensitivity during sound production. *Nature*, 418: 872–876.

Rauscheker, J. (1995) Compensatory plasticity and sensory substitution in the sensory cortex *Trends Neurosci.*, 18: 36–43.

Rind, F.C. and Simmons, P.J. (1999) Seeing what is coming: building collision-sensitive neurones. *Trends Neurosci.*, 22: 215–220.

Risberg, A. and Lubker, J. (1978) ‘Prosody and speechreading’, Speech Transmission Lab. Q. Prog. Stat. Rep. 4, 1–16.

Robert, D. and Gopfert, M.C. (2002) Novel schemes for hearing and orientation in insects. *Curr. Opin. Neurobiol.*, 12: 715–720.

Rorden, C., Greene, K., Sasine, G.M. and Baylis, G.C. (2002) Enhanced tactile performance at the destination of an upcoming saccade. *Curr. Biol.*, 12: 1429–1434.

Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M.P., Dold, G. and Hallett, M. (1996) Activation of primary visual cortex by Braille reading in blind subjects. *Nature*, 380: 526–528.

Saffran, J.R., Aslin, R.N. and Newport, E.L. (1996) Statistical learning by 8-month old infants. *Science*, 274: 1926–1928.

Schwartz, O. and Simoncelli, E.P. (2001) Natural signal statistics and sensory gain control. *Nature Neurosci.*, 4: 819–825.

Shamma, S. (2001) On the role of space and time in auditory processing. *Trends Cognitive Sci.*, 5 (8): 340–348.

Simoncelli, E.P. and Olshausen, B.A. (2001) Natural image statistics and neural representation. *Ann. Rev. Neurosci.*, 24: 1193–1216.

Somers, D.C., Dale, A.M., Seiffert, A.E. and Tootell, R.B. (1999) Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proc. Natl Acad. Sci. USA*, 96: 1663–1668.

Strauss, R. (2002) The central complex and the genetic dissection of locomotor behaviour. *Curr. Opin. Neurobiol.*, 12: 633–638.

Sur, M. and Leamey, C.A. (2001) Development and plasticity of cortical areas and networks. *Nature Rev.*, 2: 251–262.

Uhl, F., Franzen, P., Lindinger, G., Lang, W. and Deeke, L. (1991) On the functionality of the visually deprived occipital cortex in early blind persons. *Neurosci. Lett.*, 124: 256–259.

van Hateren, J.H. (1992) Theoretical predictions of spatiotemporal receptive-fields of fly LMCs, and experimental validation. *J. Comp. Physiol. A*, 171: 157–170.

Walsh, V., Ellison, A., Battelli, L. and Cowey, A. (1998) Task-specific impairments and improvements following magnetic stimulation of human visual area V5. *Proceedings of The Royal Society B*, 265: 537–543.

Wanet-Defalque, M.-C., Veraart, C., Volder, A.D., Metz, R., Michel, C., Dooms, G. and Goffinet, A. (1988) High metabolic activity in the visual cortex of early blind human subjects. *Brain Research*, 446: 369–373.

Webb, B. (2002) Robots in invertebrate neuroscience. *Nature*, 417: 359–363.

Zangaladze, A., Epstein, C.M., Grafton, S.T. and Sathian K. (1999) Involvement of visual cortex in tactile discrimination of orientation. *Nature*, 401: 587–590.

Zipser, K., Lamme, V.A. and Schiller, P.H. (1996) Contextual modulation in primary visual cortex. *J. Neurosci.*, 16: 7376–7389.

CHAPTER

4

Sensory Processing

Lionel Tarassenko and Mike Denham

1 INTRODUCTION

The sensory systems of humans and animals achieve levels of performance that far exceed those of any artificial sensory system. The almost instantaneous processing by the brain of data from the five senses – sight, hearing, smell, touch and taste – remains well beyond the capability of artificial cognitive systems.

What we can describe as 'conventional' or 'IT-centric' systems pay little attention to the fundamental principles or mechanisms of information processing in biological sensory systems. Their implementation is within conventional computing architectures, using conventional algorithmic computation methods. While this approach has not been able to match natural sensory systems, it has

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

achieved some impressive results, as this report will describe.

In contrast to these IT centric approaches, it is highly likely, and to some extent already apparent, that the brain uses an entirely different 'computational paradigm'. The brain's processing involves flexible deployment of highly parallel, asynchronous, nonlinear and adaptive dynamical systems.

This chapter reviews progress in building artificial systems that can process data from the first three senses; sight (computer vision), hearing (speech recognition) and smell (olfaction). We then go on to review the issue of 'data fusion', the ability to combine disparate sensory data, a task biological systems perform with great success. We consider systems built on the principles of deriving a probabilistic model of sensory data and applying essentially Bayesian methods of information processing (the IT-centric approach), and contrast these with the unique characteristics of biological sensory information processing.

This review is tailored to promote the objective of the Foresight Cognitive Systems project and hence is necessarily incomplete and intentionally selective. The aim of the project is to investigate ways in which the life sciences and physical sciences can learn from each other, either by working together or by greater understanding of recent progress in each others' areas of research. For this reason, each section ends with a brief overview of 'open questions', which relates the state of the art in artificial cognitive systems to our current knowledge of the human brain, where appropriate. We also highlight fruitful areas for collaborative research.

2 SIGHT: COMPUTER VISION

2.1 Introduction

The human eye focuses light on the retina. This has two sets of photosensitive cells: rods, detecting black and white light; and cones, detecting colour. The brain interprets incoming images and associates meaning with them. In this section we begin with a description of the IT-centric approach to artificial vision before going on to describe recent progress in learning from natural systems.

The goal of computer vision is to mimic natural vision by making 'useful decisions about real physical objects and scenes based on sensed images' (Shapiro and Stockman, 2001). We present here a summary of the state of the art for some key problems in computer vision – object recognition, 3D reconstruction and visual tracking.

2.2 Object Recognition

Object recognition (OR) is one of the hardest problems in computer vision (Shapiro and Stockman, 2001; Jahne and HauBeker, 2000). There are few general algorithms for the automatic recognition and localization of arbitrary 3D objects in complex scenes. The most significant sources of difficulty in the OR problem are the large changes in appearance of a single object under change of viewpoint, lighting and occlusion by other objects in the scene. An object recognition system must be invariant to such changes, while being able to discriminate between different objects with similar appearance. Probabilistic models allow for the incorporation of within-class variation such as varying illumination conditions, sensory imperfections such as noise and segmentation errors, as well as the inclusion of prior knowledge and hence empirical data.

In computer vision, as in neuroscience, object recognition is frequently divided into two schools of thought, which might be labelled object-based and view-based. In the object-based paradigm, the computational model of an object is inherently three-dimensional, and recognition is a matter of deciding which object is seen, in which 3D orientation. In the view-based version, the many different appearances of an object are each modelled independently in 2D, and no explicit 3D computations are performed. Although the object-based model appeared initially to be supported by psychophysical evidence from 'mental rotation' experiments (Shepard and Metzler, 1971), this position is no longer as clear (Gauthier *et al.*, 2002).

Coincidentally, recent progress in computer vision has been almost entirely in the view-based paradigm, with significant advances over the past few years in generic face recognition and the building of object recognition systems from poorly labelled training data. Machine learning techniques, combined with the availability of large corpora of training data have brought some once-hard problems within reach.

We can divide object recognition into two stages, pre-processing and classification. In the pre-processing stage, images are converted into 'geometric primitives', i.e. vectors of parameters that uniquely define geometric or other features of the 2D image. Classification provides an estimate of the probability that the object belongs to one of the known classes, or to none at all.

The choice of the appropriate representation of objects for the comparison of models and observations is a complex issue. Successful techniques combine astute problem-specific pre-processing with powerful classification algorithms. In the area of generic face detection, where the algorithm must find all faces in an image, without reference to the identity of any individual, recent algorithms have achieved excellent performance. This is via a combination of a pre-processing step which allows fast computation of Haar wavelet responses, and classification using a carefully trained boosting algorithm.

In object-based recognition, the position and orientation of objects relative to a reference co-ordinate frame is also important. The degrees of freedom representing the position and orientation are found by pose estimation, which corresponds to a regression problem. The regression function maps the observation to pose parameters: the major problem is the representation of this function. Many applications restrict it to a parametric set of functions, in which case it becomes a problem of parameter estimation.

Recognition and pose estimation are thus a combination of two well-known problems in pattern recognition: classification and regression. The following information must be provided:

- prior distributions, including prior knowledge about the distribution of objects in images
- model density, i.e. the probabilities of observed images given the class and pose parameters
- learning algorithms to estimate probability density functions from empirical data
- inference algorithms to interpret the processed data.

A fundamental problem is to construct adequate statistical models with an appropriate trade-off between independency assumptions, the dimension of the parameter space, the size of the available sample data and the required discriminatory power to differentiate between different classes of objects. A general probabilistic model makes use of independencies and marginalization to switch between different levels of model densities. Mixture density models, Hidden Markov Models, Markov random fields and Probabilistic Graphical Models (see section 7) all attempt to build generic models of sensor data.

2.3 Three-dimensional Perception

In considering the state of the art in computer vision, it is interesting to look at one research direction that has recently seen significant progress – the automatic perception of three-dimensional (3D) information from a set of two-dimensional (2D) images. These images might be a stereoscopic pair emulating the human visual system, or a sequence of images captured by a moving hand-held camera. In both cases, humans can readily extract the 3D information from the 2D video stream. The past few years have seen the emergence of the first artificial systems that can reliably do the same. Such systems are now robust enough to be applied commercially in adding computer-generated special effects to movies and computer games. Two factors have contributed to this success: new mathematical models of the geometry of the problem, and advances in parameter estimation under non-Gaussian noise models.

In general, the input to an artificial 3D reconstruction system is a set of 2D images of an unknown scene, photographed from unknown locations. The desired output is a 3D description of the scene, which includes a description of the positions from which the photographs were taken (see Fig. 4.1). The key to a successful solution lies in solving the 'correspondence problem': identifying objects or patterns in each image which are also visible in some of the others. Each such correspondence between 2D images of the same 3D object constrains the 3D interpretation of the set of images.

In some sense, then, the 3D reconstruction problem depends on the object recognition problem – a program that solves it must be able to identify the same object in many images where, by definition, there has been a change in viewpoint, and hence a change in the object's appearance. Indeed, there is much overlap between current object recognition techniques and those of 3D reconstruction from widely spaced views.

However, evidence that a bottom-up approach – not requiring object recognition – was possible comes from psychophysics. Julesz's introduction of the random-dotstereogram showed that stereo reconstruction could precede any high-level interpretation of the scene (Julesz, 1960). The stimuli in his experiment consist entirely of random dots, which can resolve into an image only if the correspondence problem is solved. Humans can resolve such images, suggesting that the assumption that all images are of the same 3D scene constrains the problem so much that even a very poor solution to the correspondence problem can yield the correct interpretation.

2.4 Visual Tracking

Visual tracking is about making computers 'see' moving objects. The key advance over the past 15 years has been in getting computers to anticipate movement. For example, if a computer is to follow the

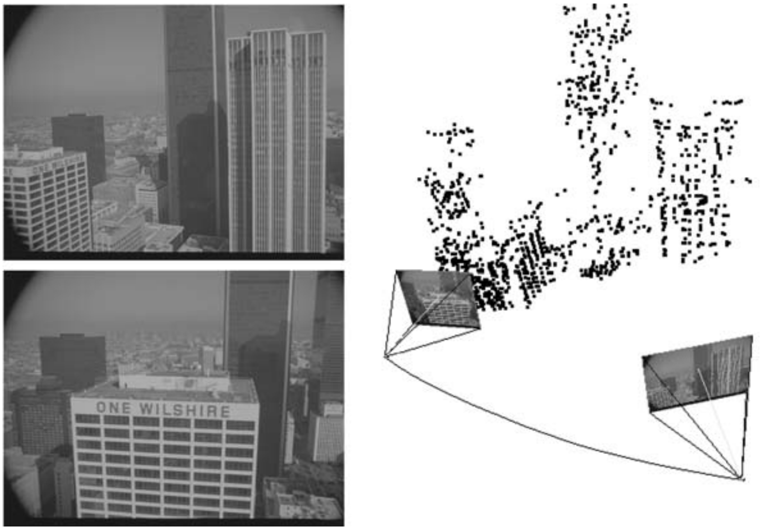


FIGURE 4.1 Reconstructing a scene in three dimensions (*right*) builds on 2D images (*left*). (This figure appears in the colour plate section)

trajectory of an intruder (see Fig. 4.2) captured on a security video, it helps enormously if the computer is programmed to expect a range of shapes ('roundish') for the intruder's head. With that information the computer can distinguish the true head from visual 'flotsam' in the room – books on a shelf or pictures on the wall, for instance.

A computer can anticipate not only shape but also movement (see Fig. 4.2). It 'expects' an intruder's motion to be smooth and largely horizontal, parallel to the floor. With this 'prior' information, the computer can concentrate, fixing on the moving head, without being continually distracted by the bright and attractive pieces of flotsam nearby.

Two main elements are required for the computer to achieve this ability to anticipate. One is the probabilistic geometric/ dynamical model and the other is the inference engine.

2.4.1 Geometric/Dynamical Model

The geometric/dynamical model has three elements. The first is state space – a space of possible shapes, geometry and appearances. We can build this state space with a combination of tools from graphics (for example, splines), projective geometry (affine spaces) and statistics (eigencurves and eigenimages) (Turk and Pentland, 1991).

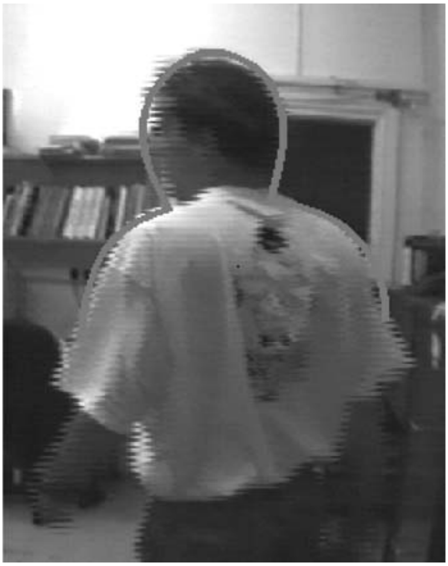


FIGURE 4.2 Computers can track intruders as they move around a room. (This figure appears in the colour plate section)

The second element is a prior probabilistic model over that state space. This may be static, treating each image in a sequence independently, or dynamic.

The third element of the geometric/ dynamical model is the likelihood model, a classical element in any probabilistic pattern recognition scheme. In the context of image processing, this is a measure of the degree of agreement between a hypothesized state and the pixels in an image or sequence of images.

The pioneering work that established this paradigm broke new ground by combining the techniques of interactive computer graphics with image processing in the 'snake' (Kass *et al.*, 1987). A simplified form of snake gained great popularity amongst practitioners (Cootes *et al.*, 1996). Learned forms of dynamical models using autoregressive processes, standard tools in time series analysis, proved powerful in focusing computational resource (Blake and Isard, 1998).

2.4.2 Inference Engine

Inference engines come in various guises for tracking movement in artificial vision. The snake employed local optimization of combined intrinsic and extrinsic energy, effectively foreshadowing full probabilistic treatment in terms of prior and likelihood. In a Gaussian, probabilistic setting, this becomes the classical Kalman filter, used widely, and subsequently related to the snake (Terzopoulos and Szeliski, 1992). Gaussian models are limited however to relatively clutter-free data. In the past five years there has been an explosion of interest in non-Gaussian techniques, especially sequential Monte Carlo methods (Blake and Isard, 1998).

2.5 Natural Systems

2.5.1 Attention

A fundamental aspect of the biological visual system is its ability to attend to the most salient regions and component parts of the visual scene in respect of the animal's goals. This manifests itself normally in humans in our ability to direct our gaze rapidly towards objects of interest in our visual environment. This has great evolutionary significance as a survival mechanism.

Laurent Itti, of the iLab at the University of Southern California, and Christof Koch, of Caltech, recently reviewed the important trends that have emerged from recent work on neurocomputational models of focal visual attention (Itti and Koch, 2001). They reach a number of conclusions:

- that perceptual saliency of a stimulus critically depends on the surrounding context. They also believe that a 'saliency map' that topographically encodes for stimulus conspicuity over the visual scene is an efficient and plausible bottom-up control strategy for attention
- that inhibition of return, in which the brain excludes the current attended location from future attention, is a crucial component of the process
- that attention and eye movements are tightly interactive, imposing specific computational challenges in attention control
- that scene understanding and object recognition strongly influence and constrain the selection of attended locations.

In terms of computational efficiency, the attentional mechanism avoids the need to process fully the massive sensory input ($\sim 10^7$–10^8 bits per second at the optic nerve) in parallel, by breaking down the problem into a rapid sequence of computationally less demanding problems of localized visual analysis.

There is strong potential in applying computational architectures and processing mechanisms for visual attention derived from neurocomputational models of the biological system. This is particularly promising for areas such as surveillance, automatic target and image recognition, navigational aids and robot control. There are already several examples of such detailed neurocomputational models of visual attention (Koch and Ullman, 1985; Lee *et al.*, 1999; Itti *et al.*, 2000; Deco and Zihl, 2001; Grossberg and Raizada 2000; Itti *et al.*, 1998 and 2000; Tsotsos *et al.*, 1995; Itti and Koch, 2000; Rybak *et al.*, 1998; Deco and Schurmann, 2000). One of these has led to a patented device for the computation of intrinsic perceptual saliency in visual environments and applications (Koch and Itti, 2001). One application described for this device is the automatic evaluation and optimization of sales or advertisement displays.

2.5.2 Invariant Object Recognition

Another fundamental problem solved by the biological sensory system in the neocortex is the recognition of objects relatively independently of size, contrast, spatial frequency, position on the retina or angle of view. VisNet is one attempt to capture this property in a neurocomputational model (Rolls and Milward, 2000).

VisNet2, a development of VisNet, is based on the organization of visual processing for object recognition as a set of hierarchically connected neocortical areas (V1, V2, V4, TEO, IT). Circuits in one cortical area receive information from circuits in the preceding area. Neurons in layers 4, 2 and 3 of one microcircuit mainly connect to neurons in layers 2 and 3 of a neocortical microcircuit of the previous area.

The model thus takes the form of a four-layer feed-forward network with convergence to each part of a layer from a small region of the preceding layer, with competition between the neurons within a layer and with a trace learning rule to help it to learn transform invariance. The trace rule is a modified Hebbian rule, which modifies synaptic weights according to both the current firing rates and the firing rates to recently seen stimuli. This enables a neuron to learn to respond similarly to the gradually transforming inputs it receives, which over the short term are likely to be about the same object, given the statistics of normal visual inputs.

The short response latencies of face-selective neurons in the inferotemporal (IT) area of the neocortex impose major constraints on models of visual-object recognition. It appears that visual information must propagate in a feed-forward fashion, with most neurons having time to fire only one spike.

Arnaud Delorme, of the Salk Institute for Biological Studies, and Simon J. Thorpe, of the Centre de Recherche Cerveau et Cognition in Toulouse, hypothesize that the order of firing of ganglion cells in the retina can encode flashed stimuli (Delorme and Thorpe, 2001). The researchers propose a neuronal mechanism that could be related to fast shunting inhibition to decode such information. Based on these assumptions, they have built a three-layered neurocomputational model of retinotopically organized neuronal maps. By using a learning rule involving spike timing-dependent plasticity, they showed that neuronal maps in the output layer can learn to recognize photographs of faces. The model was not only able to generalize to novel views of the same faces, it was also remarkably resistant to image noise and reductions in contrast. SpikeNet, a commercial system for face recognition, incorporates the sensory information processing principles embodied in this model.

2.6 Silicon Retinas

The most striking feature that emerges from observation of the structure of the retina is that it is made up of well-delineated sheets. One sheet contains the two types of photosensitive neurons, the rods and the cones. The rods are very sensitive detectors that are responsible for monochrome vision at night. In brighter scenes there is a larger input signal: cone cells become active at light levels corresponding to twilight. The human eye has three different types of cone cells with different sensitivities to various wavelengths of light. Using only these three different types of cells, the visual system can distinguish subtle changes in colour and can extract a large amount of useful information from a scene.

From the structure of the retina, it is clear that the output signal from each cone cell must flow through a bipolar cell before it reaches a ganglion cell that forms part of the output from the retina, the optic nerve. The structure alone suggests that the bipolar cells form a critical part of the visual system. Furthermore, comparisons of different vertebrate species show that the response of bipolar cells is the same in all species. This suggests a common image processing strategy within this first critical stage of the visual system. An understanding of this strategy could lead to the design of better cameras, for example.

Within the vertebrate's retina, each bipolar cell responds to the cone and horizontal cells that form the outer plexiform layer of the retina (Mahowald, 1994). Experiments have shown that the analogue output signal from a cone cell is proportional to the logarithm of the incident light intensity. The network of laterally connected horizontal cells then appears to respond to the local average output of the cone cells, with the influence of each cone decreasing with lateral distance along the network. The output of each bipolar cell is then the amplified difference between the response of the local cone and horizontal cells.

The result is that each bipolar cell has a centre-surround receptive field in which it is stimulated by inputs in a small central spot and inhibited by any stimulus in a larger surrounding annulus. This receptive field acts on the image as a high-pass spatial filter. Retinas therefore combine an array of logarithmic detectors and high-pass spatial filters.

The first sensors designed to mimic both the function and the form of the retina, often referred to as neuromorphic systems, were fabricated in the late 1980s (Mead, 1989). By designing a pixel circuit containing field-effect transistors operating in a regime known as sub-threshold or weak inversion, they demonstrated a 'silicon retina' that combined logarithmic detectors and spatial filtering based upon analogue voltages.

Joachim Buhmann and his colleagues of the University of Bonn provided the first quantified demonstration of the advantages of a silicon retina (Buhmann *et al.*, 1994). This group compared the performance of two systems for face recognition, one containing a silicon retina and the other a conventional CCD camera. They found that using well-controlled uniform illumination the CCD outperformed the silicon retina, a result they attributed to the limited resolution of the silicon retina. However, when the face was illuminated from one side, the performance of the CCD system fell from 75.2% to 62.4%. In contrast, the performance of the silicon retina system increased from 71.6% to 93.6%. By far the best performance was therefore with the silicon retina using extra information available from shadows created by non-uniform illumination.

The development of silicon retinas has highlighted two important results. The first is the clear demonstration that within an array of sensors collective, parallel analogue computation can extract useful information from a vast amount of input data. More specifically these systems clearly demonstrate the advantages of using a combination of logarithmic detectors and a high-pass spatial filter when imaging naturally illuminated scenes.

2.7 Open Questions

Whether the aim is object recognition or tracking, computer vision employs a representation of 'key features' which then need to be combined into collectives for recognition as objects. This involves the binding problem. We do not know how the brain accomplishes this gestalt process.

We have a better understanding of how the brain passes information and organizes its storage for the visual cortex than for any other area of the brain. However, we know relatively little about how the brain combines features, especially through time.

At a higher level, the analysis of the perception of a scene is also a long way off. To account for perception requires a better understanding of the binding problem and of how the brain encodes time, or dynamics. An information theory view of the use of spike trains in the neural code has been discussed but this is limited to discussions of rate coding (Rieke *et al.*, 1997).

Current object recognition systems depend on large volumes of manually labelled training data. Research efforts are directed now at learning object recognizers from training data which are less informative or even partially erroneous. For example, it is easier for a human to say whether or not an image contains a face than to indicate the location of a face if present.

The natural research question is then, 'given a large set of images known to contain faces, but without knowing where the faces are, can an automatic face detector be trained?' Even weaker are training sets where each image is associated with multiple labels (e.g. 'tree, grass, rabbit' or 'George, Paul, Ringo'), again without positional information. In each of these cases, classification becomes more difficult, but the potential for building more general systems is clear.

There are intriguing implications for neuroscience research arising out of the latest studies of visual tracking. One possibility is to look at neural spatiotemporal processing in the light of Kalman filtering. Kalman filter models predict the time-course of spatial and temporal frequency properties of visual channels under different conditions, for example, tracking coherent motion, seeking 'lost' objects. It would be interesting to see if these predictions had value in understanding neurophysiological findings.

A second possibility is the random element inherent in sequential Monte Carlo techniques. Computationally this enables tractable inference, albeit approximate, over spaces with dimensions greater than two but less than about 30. There may be potential here for modelling neural mechanisms that can deal with high dimensional state spaces.

Existing silicon retinas suffer from three major problems. The circuits needed to perform parallel analogue computation increase the area of each pixel, limiting the spatial resolution of the 'retina'. They are monochrome sensors and the extension of the existing design to support three different colours is not trivial. Most importantly, variations between the responses of individual logarithmic detectors severely degrades the quality of the output images.

The solution to an efficient, biologically inspired vision sensor may be to mimic the function of the retina, but not its structure. This would incorporate logarithmic detection followed by high-pass spatio-temporal filtering and would take the filter outside the pixel circuitry to make high-resolution cameras, considerably reducing pixel size.

3 HEARING: SPEECH RECOGNITION

3.1 Introduction

Sound, produced by vibrating objects, has three characteristics; pitch, volume and quality. When sound waves reach the ears, they cause the eardrums to vibrate and induce movement in the three smallest bones in the human body; the hammer, anvil and stirrup. These three bones press fluid in the inner ear against membranes, which brush tiny hairs, triggering nearby nerve cells, sending messages to the brain, which interprets them. The goal of speech recognition is to mimic this process.

As in the previous section, our starting point is IT-focused, followed by a review of natural systems.

Although the technology is relatively new, over the past 20 years there has been considerable progress in continuous speech recognition (CSR) using probabilistic models. Much of what follows is based on a recent overview of speech recognition (Young, 2001).

3.2 Continuous Speech Recognition Using Hidden Markov Models

The now standard approach to continuous speech recognition uses a probabilistic model to find the optimal word sequence from a sequence of acoustic vectors describing the speech signal in the frequency domain. This requires maximizing the probability of the word sequence given a sequence of acoustic vectors. This is equivalent to finding the maximum product of the acoustic model, the acoustic sequence probability given the word sequence, with the language model – the word sequence probability.

The acoustic model comprises a series of some 45 basic phones, from which any word can be constructed. This approach reduces an enormous vocabulary of words to a manageable sequence, although each phone is affected by its context (co-articulation effect). Each phone is represented by a Hidden Markov Model (HMM), the parameters of which are trained using a maximum-likelihood algorithm. Speech is thus modelled by a sequence of phone models.

The language model is found by calculating the word sequence probability. To simplify the model, each word is assumed to depend upon the last n-1 words, forming an n-gram language model, where n is typically 2 or 3. The probabilities are then estimated from training sets, although obtaining sufficient data is a significant problem.

The decoding method attempts to find the optimal phone, and hence word, sequence. This is a significant computational task because of the size of the set of possible word histories, determined by the value of n. The search problem is made tractable by limiting the number of possible solutions considered, or by performing multiple passes over the data, continuously refining the solution as the search space is reduced.

The context of a phone also affects its pronunciation. So recognition must also consider the preceding and following phones. The simplest way to overcome this is to use 'logical phones', giving a different phone model for every possible context of each base phone. However, obtaining sufficient data is again a problem. This is overcome by clustering the logical phones onto a reduced set of shared physical models, using decision trees. A method known as soft-tying can prevent the partitioning from becoming too coarse. This entails a post-processing stage that groups each state with its nearest neighbours.

Although a decision-tree tied-state context-dependent modelling scheme can handle carefully articulated speech, it cannot recognize speech that is less well-articulated. Incorporating context-dependent pronunciation in the decision trees increases complexity with only a marginal improvement in recognition rates. Linguists argue that this is a fundamental limitation of modelling speech using a sequence of basic phones.

The speech feature vectors often incorporate a mixture of Gaussians for the outputs from the Hidden Markov Model to allow for a more general distribution and to cope with variability in one speaker and between different speakers. The variances are constrained to be diagonal to reduce the number of unknown parameters as there are typically 10 to 32 Gaussians per state and 5000–10 000 states per system. Since the training set may be a poor representation of the test data, the model parameters can be adapted to improve the fit to the given data. For example, this can be done by treating them as random variables and estimating them with a standard Bayesian maximum a posteriori approach.

Three main assumptions underpin current work on continuous speech recognition using statistical approaches; frame independence, the 'sequence of phones' speech model and the *n*-gram language modelling approach. Researchers are developing segment models to weaken the assumption of frame independence. They use the concept of a segment model whereby speech features are based on segments, rather than frames, and introducing a distribution of segment durations. However, improving the accuracy of segment modelling tends to increase errors caused by assumptions on the sequence of phones.

Asynchronous parallel models have also arisen, as many phonological processes are more naturally based on a hierarchy of parallel feature streams than on a strictly sequential approach. The amount of parallelism introduced varies from independent streams that are only coupled at major boundaries to multiple observation distributions sharing the same underlying Markov chain.

More recently, research has led to a factorial HMM approach, whereby Markov chains evolve in parallel. However, the state spaces have to be constrained. This is currently done using either a mixed-memory approximation or parameter tying. This approach, although computationally very expensive, is still relatively new and may improve with experience.

Attempts to improve the fixed-length limitation of the *n*-gram model have been based on the use of longer range forecasters. This is done mainly with trigger models, whereby predictor words are counted if they lie anywhere within the current word's history. The trigger approach is combined with conventional *n*-grams using the Maximum Entropy (ME) method. Head words can also work with the ME method and *n*-grams to attempt to model long-range dependencies in a more principled way.

Speech recognition systems have attained very high performance levels. Speaker-independent, vocabulary-independent, near real-time systems provide acceptable accuracy for many non-critical applications. These systems, however, are complex and the size of the word vocabulary requires a large amount of training data. Speaker-independent systems are good enough for limited volume speaker identification and/or verification systems – controlled access to buildings for example – but fall short of being useful as a human-computer interface for applications such as telephone banking.

3.3 Natural Systems

Progress in building efficient and plausible neurocomputational models of biological auditory processing has been largely restricted to the peripheral processing of auditory stimuli (sounds). We now know that the brain computes many of the principal characteristics of an auditory stimulus before information about the stimulus reaches the neocortex. These 'preprocessed' characteristics include pitch and direction.

The UK has had a major research activity in this area for some years, resulting in detailed and comprehensive models of peripheral auditory processing in the brain. One recent outcome of this activity is a detailed computational model of the inner hair cell and auditory-nerve complex (Sumner *et al.*, 2002).

Building on previous research, this model is intended as a component of more comprehensive models of the auditory periphery. It combines smaller components that aim to be faithful to physiology in so far as is practicable and known. The model reproduces a wide range of observations of animal inner hair cell receptor potential and auditory nerve. When the input comes from a suitably non-linear simulation of the motion of the cochlear partition, the model can simulate the rate-intensity functions of low-, medium- and high-spontaneous rate auditory nerve fibres in response to stimulation both at rest frequency and at other frequencies. The model also reproduces quantitatively phase-locking characteristics, relative refractory effects, mean-to-variance ratio, and first and second-order discharge history effects.

André van Schaik, of the MANTRA Centre for Neuromimetic Systems, Swiss Federal Institute of Technology, and Ray Meddis, of Essex University, recently described an analogue VLSI implementation of a model of signal processing in the auditory brainstem (van Schaik *et al.*, 1999). The implementation is based on a model of amplitude-modulation sensitivity in the central nucleus of the inferior colliculus (CNIC). A single chip implements the three processing stages of the model, the inner hair cell, cochlear nucleus sustained-chopper, and CNIC coincidence-detection stages.

The chip contains 142 neurons and incorporates two circuits: an inner hair cell circuit and a neuron circuit. The input to the chip comes from a 'silicon cochlea' consisting of a cascade of filters that simulate the mechanical frequency selectivity of the basilar membrane. The chip was evaluated using amplitude-modulated pure tones. Individual cells in the CNIC stage demonstrate bandpass rate-modulation responses using these stimuli. An array of these cells represents the frequency of modulation as the location of the cell generating the highest rate of action potentials. The chip processes acoustic signals in real time and demonstrates the feasibility of using analogue VLSI to build and test auditory models with many component neurons.

As in visual attention and figure-ground separation, an essential feature of auditory sensory processing is the ability to focus on the part of the sound signal of interest against a background of distracting signals, and to be able to direct this focus at will. This is particularly important in understanding speech in the presence of background noise. The biological auditory system is very good at carrying out this process, the so-called 'cocktail party effect'. The ability to analyse speech against background noise has applications in both speech recognition systems and hearing aids.

Several researchers have developed neurocomputational models of the process (McCabe and Denham, 1997). The model they described shows that streaming results from interactions between the tonotopic patterns of activity of incoming signals and traces of previous activity that act as a feedback signal and influence processing of subsequent signals. The model's behaviour agrees with a number of well-known psychophysical results, including the relationship between presentation rate, frequency separation and streaming, the temporal development of streaming, and the effect of background organization on streaming.

The fundamental principle of sensory information processing that underlies the neurocomputational models of the process formed the basis of a highly efficient noise reduction system in NeuVoice, a commercial speech recognizer for such applications as mobile telephones and 'pocket' computers.

3.4 Open Questions

The technology of high-quality speech recognition remains based on systems composed of two separate parts; a front-end which uses acoustic waveform representations, followed by an inference mechanism built around a Hidden Markov Model framework. There has been little effort in neurobiologically inspired processing to replace the HMM framework. Similarly, there have been few attempts to consider structured modifications of the feature coding of the acoustic waveform using higher-level feedback from the inference part of speech recognition, an issue of data fusion.

4 SMELL: OLFACTION

4.1 Introduction

Smell, also termed olfaction, is the detection of floating molecules by receptor cells in the nasal cavity behind the bridge of the nose. Some one hundred million of these cells, each less than 1 micrometre in diameter, interact with odour molecules. The olfactory bulb pre-processes the information from these receptors before it proceeds to the pattern recognition processes in the olfactory cortex. The number of primary odours is thought to be in the region of seven. Humans can typically distinguish somewhat fewer than 100 odours.

The olfactory system has four main parts; the receptor array, the olfactory bulb, the nucleus, and the cortex. The olfactory bulb performs dynamic fusion of data from the receptors. We know that the olfactory bulb constructs a distributed code for odours, i.e. cells are not tuned for specific smells. There are feedback paths from the nucleus and the cortex to the olfactory bulb. Every neuron in the bulb seems to be involved with every stimulus.

Although each sniff of the same odorant creates a very different spatial response in the receptor array, the pattern of electrical activity of the macroscopic olfactory bulb is very similar for the same odorant and different for different odorants. The olfactory bulb appears to use phase-locked dynamic oscillators to encode odorant information in a robust and distributed manner.

4.2 Electronic Noses

The attempt to mimic the sense of smell is termed an electronic nose. This is defined as 'an instrument which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern recognition system, capable of recognizing simple or complex odours' (Gardner and Bartlett, 1999). A number of research groups are working on devices that mimic the biological olfactory system. This is one area where there is closer connection between the IT-centric model and biological approaches.

Many different crude artificial nose systems have demonstrated that representing an odour as a high-dimensional vector allows broad discrimination.

With most electronic noses, an array of chemical sensors separates out different odours. Each sensor responds to a range of odours with varying specificity and selectivity. Most commercial systems use either metal-oxide sensors or conducting polymer resistive sensors, although alternatives are being investigated. The polymer film is attractive because it is much faster than other methods.

A team at the California Institute of Technology has developed an array-based electronic nose technology (Freund and Lewis, 1995). Their technology uses polymer sensors mixed with carbon black to make them conductive. An odorant causes the polymers to swell and their resistance to change. An array of different polymers swells to different degrees, giving a signature of the odorant. The company Cyrano Sciences has commercialized this technology with the launch of a handheld electronic nose (Cyrano Sciences). Further development of the technology has produced chips with arrays of thousands of sensors on CMOS VLSI substrates. The chips are standard CMOS with a post-processing electrodeless gold deposition step that forms the sensor contacts.

Our knowledge of the complex relationship between the structural features of molecules and the response of a sensor is limited. Hence sensor design is largely an ad-hoc process.

Sensors should operate at ambient pressure with a high-frequency response and their operation should be independent of temperature or humidity. Sensor design is also limited by the requirements of size, detection range and reproducibility over time and over different sensors of the same type. A severe limitation on the use of electronic noses is the compensation needed to accommodate temperature and humidity effects and, in some cases, the flow rate.

The characterization of odours with an electronic nose is divided into pre-processing and pattern recognition. Pre-processing typically involves the calculation of relative or fractional changes in each parameter, to compensate for noise and drift, and some form of normalization. A Bayesian or neural network classifier then exploits the distributed representation of the normalized sensor responses as high-dimensional vectors.

We may need more sophisticated processing techniques to detect more subtle changes in odours. In particular, adaptive techniques, using dynamical or time-varying models, should be able to exploit the temporal information in transient signals. The use of transient responses and the frequency spectrum may also give additional information to aid in odour discrimination, possibly reducing the number of sensors required.

4.3 Natural Systems

For all the progress we have outlined here, research in developing an electronic nose is hindered by a lack of realistic biological models, and of practical chemical sensors whose function is similar to biological receptors.

Walter J. Freeman and his colleagues at the University of California at Berkeley have made extensive study of the olfactory sensory system over the past 30 years. This has resulted in a neurocomputational model based on chaotic attractor neural networks (Eisenberg *et al.*, 1989).

John J. Hopfield of Princeton University has also developed a neurocomputational model of olfactory processing (Hopfield, 1999). He points out that highly olfactory animals must solve several basic olfactory tasks, including background suppression, multiple object separation, mixture separation and source identification. The large number, *N*, of classes of olfactory receptor cells – hundreds or thousands – permits computational strategies and algorithms that would not be effective in a low-dimension stimulus space.

Hopfield constructs a model of the patterns of olfactory receptor responses based on the broad distribution of olfactory thresholds. Representing one odour from the viewpoint of another then allows a common description of the most important basic problems and shows how to solve them when *N* is large. One possible biological implementation of these algorithms uses action potential timing and adaptation as the ‘hardware’ features that are responsible for effective neural computation.

Zhaoping Li, of the Gatsby Computational Neuroscience Unit at UCL, and John Hertz, of Nordita, Copenhagen, have described a model of an olfactory system that performs odour segmentation (Li and Hertz, 2000). Based on the anatomy and physiology of the biological olfactory system, their model consists of coupled bulb and cortex modules. The bulb encodes the odour inputs as oscillating patterns.

The cortex functions as an associative memory. When the input from the bulb matches a pattern stored in the connections between its units, the cortical units resonate in an oscillatory pattern characteristic of that odour. Further circuitry transforms this oscillatory signal to a slowly varying feedback to the bulb. This feedback implements olfactory segmentation by suppressing the bulb's response to the pre-existing odour, thereby allowing the brain to single out and recognize subsequent odours.

4.4 Open Questions

Olfaction is interesting when compared with the other senses, because there is no obvious spatial topography of odorant space as there is for vision and hearing. This is one reason why olfaction remains less well understood from the perspective of information processing. Very few artificial olfactory systems have attempted to exploit the dynamical aspects of processing in the bulb, nucleus or cortex. There is clearly a lot of work to do before we can fully understand and reproduce the physics of pattern processing in the human olfactory system.

5 SENSOR FUSION

5.1 Introduction

The performance of the individual senses is in itself remarkable, but that is only the beginning. The living brain manages extraordinary feats of 'data fusion', combining the input from the senses. It may be that artificial systems could benefit from a greater understanding of how natural systems handle data fusion.

Sensor fusion is the merger of multiple measurements of sets of parameters to form a single representation. Sensor fusion can be divided into:

- *Complementary* Sensors do not depend upon one another but merge information to form a more complete picture.
- *Competitive* Each sensor provides equivalent information about the underlying parameters.
- *Cooperative* Sensors combine to provide information that each sensor alone could not provide.

The simplest approach is to average the signals, or 'pixel intensities,' from different sensors at each point. However, this is clearly unsuitable in many situations and is very sensitive to noise. The most rigorous way to perform sensor fusion is to use a Bayesian approach. A Bayesian framework has advantages over other methods. It allows for the development of a probabilistic scheme to combine multiple information sources, particularly since these sources may be multimodal.

If a linear, or linearized, model is used, with uncertainty in sensor and motion models modelled by Gaussian densities, it is possible to employ classical Bayesian methods based on linearized Kalman filters. The sequential Monte Carlo method, or particle filter, provides an approximate solution to the problem of non-linear and non-Gaussian estimation.

Since the particle filter approximates the optimal solution, it can outperform the Kalman filter in many cases, given sufficient computational resources. However, the performance of particle filters degrades rapidly as the dimension of the state vector increases, leading to various proposals for ways of avoiding this problem.

A choice of state coordinates making the state equation linear reduces computation time. This opens up the possibility for Rao-Blackwell techniques, i.e. marginalizing the full conditional posterior density with respect to almost linear parts. Calculating the likelihood one step ahead of re-sampling is crucial, together with adding extra noise, in avoiding divergence.

Applications of Data Fusion

The combination of computer vision with speech recognition could bring significant benefits. Computer vision can aid in word recognition by the use of visual phones or 'visemes', since very similar acoustic phones frequently have very different visual phones, giving a degree of orthogonality in phone selection. Combination of the two techniques could overcome some problems of speech recognition, for example its sensitivity to noise and its speaker dependence, giving dramatic improvements in word recognition in noisy environments. However, there are significant problems involved with fusing such disparate data, for example, in the difference in the numbers of acoustic and visual phones, time delays and differences in sampling rates.

Researchers have adopted two main approaches, feature fusion or decision fusion. The first extracts features using the available techniques and combines them before the recognition process. The second combines the outputs from separate recognition processes. The second approach seems to be preferred as combination occurs at a higher level.

The fusion of different approaches to imaging – mainly computer tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) and single photon emission computer tomography (SPECT) – has also made significant advances. It has enormous clinical importance as each technique provides different information, anatomical for CT and MRI, and functional for PET and SPECT.

The relevant images have to be matched, or 'registered,' before fusion so that the target image is modified to match the reference image. All matching algorithms have four stages; extraction of key structures, matching key structures, evaluation of similarities between structures, and transformation of the target images. Fusion then occurs by combining the equivalent images. It is also useful to incorporate prior knowledge about the underlying physiology, for example the anatomical shapes of different regions.

Open Questions

In many real-world problems, Bayesian methods are inferior to simpler approaches. The problem is that fusion should be considered from a 'top-down' or systems-level approach, rather than from a 'bottom-up' approach.

The correct way to combine sensors, or information generally, requires knowledge of the correct 'utility function' for decision-making. The two key research problems in data fusion are the approximation of what is a full many-body problem and the incorporation of different types of uncertainty into the analysis.

Most fusion systems are hierarchical and hence require modelling of some form of locality. Hierarchical systems are less robust than fully distributed systems, but fully distributed systems require complicated non-linear, stochastic and temporal mathematics which are currently lacking. Approximating the many-body problem has links to graphical models (see below), and has to face the issue of how to construct tractable but principled models.

As far as incorporating uncertainty is concerned, unless local distributions are known exactly, which is never the case in practice, simple combination models outperform the Bayesian approach to model integration, even assuming independence between sensors. This is because errors propagate through to the decision-making part and hence the errors can dominate the decisions (Kittler, 1998).

Although fusion does occur at some levels in the human brain, for example, at the individual neuron level for signal enhancement, at other levels the brain prefers not to fuse but leaves data sources fragmented and works with multiple distributed representations of information (Abbot and Sejnowski, 1999), leaving the fusion to the high-level binding problem.

6 THE NEOCORTICAL MICROCIRCUIT

The neocortex of the brain serves perception, attention, memory and a spectrum of other higher cognitive functions. These combine to provide the outstanding powers of biological sensory systems.

Markram has identified a number of characteristics of the neocortex that suggest that the biological system solves the computational problem of sensory processing in a fundamentally different way from conventional computers (Markram, 2002). First, an apparently stereotypical microcircuit of neurons carries out all of the tasks of the neocortex. And, whilst individual circuits might adapt to reflect the species-specific or modality-specific nature of the stimulus, the underlying computational process remains the same. Secondly, it appears that the same microcircuit simultaneously performs multiple processing functions.

This form of unlimited parallel computation allows simultaneous integration of multiple sensory features in a manner that intrinsically solves the so-called 'binding' problem. In addition, the boundaries between microcircuits are not fixed. The circuits form dynamically from a continuous sheet of interconnected cells, according to the processing requirements and the topographical mapping of sensory data into, and motor data from, the neocortex. Each functional microcircuit consists of an extremely dense set of neurons – of the order of 10^4 neurons within a volume of less than 1 mm^3. There is a great diversity of cell types within the microcircuits – at least nine major anatomical classes of cells, 15 major electrophysiological classes and 20 major molecular classes. In addition, these cells are precisely connected to each other, through synapses that act as unique adaptive filters for the transmission of data between any pair of cells. Thus while a single neuron may connect to many hundreds of other neurons, each target neuron will interpret a signal sent by a neuron in a unique way. Also, these connections are not static but change their transmission characteristics dynamically and asynchronously, on a millisecond timescale, and are organized in a highly precise way in relation to the function of the different neuron types that they connect.

Finally, the neocortical microcircuit appears to be able to maintain a virtual continuum of timescales of information processing, with time constants of activity ranging from a few milliseconds to years. Thus it seems to bear a close resemblance to a dynamical system where the state activity moves along continuous trajectories in a complex, extremely high-dimension space, these paths being uniquely defined over very long, if not infinite, periods of time. Moreover, the neocortical microcircuit has the remarkable ability to adapt continuously and to optimize itself to meet the requirements of new tasks and environments. This occurs both through unsupervised modification of its dynamic parameters, in particular those that define the response characteristics of individual synaptic connections, and through modulatory control of the circuit dynamics by specific neurotransmitters which inform and teach the microcircuit about the desired goals of, and objectives achieved by, the animal.

6.1 Open Questions

Arguably, the most important challenge in research at the interface between neuroscience and computer science is to understand the cortical microcircuit and its role and function in information processing, in particular in sensory systems. This ubiquitous neural circuit can then be a building block in novel computational architectures for artificial sensory systems. By integrating our understanding of the neocortical microcircuit with our knowledge of the global characteristics of sensory processing – such as attention, figure-ground separation or object recognition – we may be able to build systems that begin to match the performance of biological systems.

7 LEARNING: PROBABILISTIC GRAPHICAL MODELS

Probabilistic models form the foundation for much work in machine learning, computer vision, signal processing and data analysis. The formulation and solution of such models rests on the two simple equations of probability theory, the sum rule and the product rule. However, the simplicity of these equations is deceptive. In practice all but the simplest models require highly complex manipulations and can become analytically and/or computationally intractable.

Probabilistic graphical models can be seen as a marriage between probability theory and graph theory. We gain several benefits by augmenting the laws of probability with diagrams:

- We derive new insights into existing models, for example by providing a way to understand readily their conditional independence properties.
- It becomes much easier to formulate and motivate new, sophisticated probabilistic models, simply by expressing them diagrammatically.
- Complex computations, for example those required to perform inference, can be expressed in terms of purely graphical operations (such as moralization, triangulation, finding the maximal spanning tree, etc.).
- We can even go directly from graphical representation all the way to computational evaluation of predictions.

There are two principal kinds of graphical model, directed graphs and undirected graphs, corresponding to graphs with directed edges (i.e. endowed with arrows) or undirected edges. Each node of the graph represents a (group of) random variables. In the directed graph representation, the joint distribution of all the variables is defined by a product of conditional distributions, one for each node, conditioned on the states of the variables corresponding to the parents of the respective nodes in the directed graph. For an undirected graph, the joint distribution is given by the product of clique potentials (non-negative functions) defined over all cliques (maximal fully connected subgraphs) in the graph. Examples for well-known models corresponding to directed graphs include Kalman filters, Hidden Markov Models and belief networks while examples of undirected graphs include Markov random fields and Boltzmann machines.

A joint probability distribution over sets of random variables may exhibit one or more properties of conditional independence, and this corresponds to one of the key forms of prior knowledge which is built into a probabilistic model. These correspond to expressions of the form: $P(A \mid B, C) = P(A \mid B)$ which says that A is independent of C given B. Using the product rule of probability, we can easily rewrite this in the form $P(A, B \mid C) = P(A \mid C)\ P(B \mid C)$ which says that, given C, the joint distribution of A and B factorizes, so that A and B are independent. A very powerful feature of the graphical model representation is that such conditional independence properties can be read off directly from the graph without having to perform any mathematical manipulations of probability distributions.

Graphical models appear particularly useful in a Bayesian setting, where the uncertainty in unknown quantities is captured by expressing them as random variables endowed with prior probability distributions. Thus, in a mixture of Gaussians for example, the means, covariances and mixing proportions of the Gaussians (as well as the latent variables describing which components generated each data point) are all unknown and hence are described by stochastic variables. These additional stochastic variables correspond to additional nodes in an expanded graphical model having a hierarchical structure. Furthermore, we may wish to use flexible priors, governed by hyper-parameters which themselves are unknown and hence described by hyper-priors, thereby extending the hierarchy to higher levels.

From a Bayesian viewpoint, learning simply corresponds to inference on this expanded graph (Jordan, 1998), in which we condition on the observed variables (the 'data') and then infer the posterior distributions over quantities of interest (e.g. means of the Gaussian components) while marginalizing (integrating) out any remaining variables (such as the hyper-parameters).

Thus the application of graphical models to practical problems requires the solution of inference problems, and here graphical models are particularly powerful in allowing the general inference problem to be solved exactly through graphical manipulations. For tree-structured graphs the framework of belief propagation (Pearl, 1988) provides an exact solution in time linear in the size of the graph. For more complex graphs having loops, the graph is first transformed into a tree structure (a 'junction' tree) in which each composite node comprises multiple variables from the original graph, and then a local message-passing algorithm (a generalization of belief propagation) is performed. While the junction tree framework is exact, and indeed optimal, for complex models corresponding for instance to densely connected graphs, the algorithm can become computationally intractable. It scales exponentially with the number of variables in the largest composite node.

We must therefore seek approximation schemes in such cases. Three important classes of approximation are currently being explored: Markov chain Monte Carlo (MCMC), a numerical approach in which the approximation usually arises through the use of finite computer time; variational methods, which are very powerful deterministic approximation schemes and which have recently been shown to scale to very large information retrieval problems well beyond the scope of MCMC methods; and belief propagation, in which the tree algorithm is applied directly to a general graph. The last is a purely ad hoc approach which sometimes fails but which surprisingly often produces spectacular results and which is currently the focus of much theoretical investigation. In all cases, the graphical representation offers considerable assistance in the formulation and solution of the inference problem.

7.1 Open Questions

Learning is an integral part of human behaviour. State-of-the art algorithms in machine learning are highly complex and require massive computational power. Synaptic plasticity, which is thought to be the key mechanism for learning in animal brains, is reasonably well understood but there is still some debate about the precision with which it operates. As the approach in machine learning is so different, it is not clear whether a deeper understanding of human learning will help in the design of improved algorithms.

Acknowledgements

This report draws heavily on contributions from experts in the field. Professor Tarassenko would like to thank the following for their help: Professor Andrew Blake (section 2.4 on visual tracking); Dr Steve Collins (section 2.6 on artificial retinas); and Professor Chris Bishop (section 7 on probabilistic graphical models). Professor David Lowe helped to formulate the Open Questions and Dr Stephen Payne worked on the literature survey.

References

Abbot, L. and Sejnowski, T.J. (1999) *Neural Codes and Distributed Representations*. Cambridge, MA: MIT Press.

Blake, A. and Isard, M. (1998) *Active Contours*. Berlin: Springer.

Buhmann, J.M., Lades, M. and Eeckmann, F. (1994) Illumination-invariant face recognition with a contrast sensitive silicon retina. *Adv. Neural Inform. Processing Syst.*, 6: 769–776, Morgan Kaufmann, San Francisco, CA.

Cootes, T.F., Taylor, C.J., Cooper, D.H. and Graham, J. (1996) Active shape models – their training and application. *Comput. Vision Image Understanding*, 61 (1): 38–59.

Cyrano Sciences. http://cyranosciences.com.

Deco, G. and Schurmann, B. (2000) A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Res.*, 40, 2845–2859.

Deco, G. and Zihl, J. (2001) A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system. *J. Comp. Neurosci.*, 10: 231–253.

Delorme, A. and Thorpe, S.J. (2001) Face identification using one spike per neuron: resistance to image degradations. *Neural Netw.*, 14: 795–803.

Eisenberg, J., Freeman, W.J. and Burke, B. (1989) Hardware architecture of a neural network model simulating pattern recognition by the olfactory bulb. *Neural Netw.*, 2: 315–325.

Freund, M.S. and Lewis, N.S. (1995) A chemically diverse conducting polymer-based 'electronic nose'. *Proc. Natl Acad. Sci. USA*, 92: 2652–2656.

Gardner, J.W. and Bartlett, P.N. (1999) Electronic noses: principles and applications. Oxford: Oxford University Press.

Gauthier, I., Hayward, W.G., Tarr, M.J., Anderson, A.W., Skudlarski, P. and Gore, J.C. (2002) BOLD activity during mental rotation and viewpoint-dependent object recognition. *Neuron*, 34: 161–171.

Grossberg, S. and Raizada, R.D. (2000) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Res.*, 40: 1413–1432.

Hopfield, J.J. (1999) Odor space and olfactory processing: collective algorithms and neural implementation. *Proc. Natl Acad. Sci. USA*, 96: 12506–12511.

Hyder, A.K., Shabazian, E. and Waltz, E. (eds) *Multisensor Fusion*. (Nato Science Series). New York: Kluwer.

Itti, L. and Koch, C. (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.*, 40: 1489–1506.

Itti, L. and Koch, C. (2001) Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2: 194–203.

Itti, L., Koch, C. and Braun, J. (2000) Revisiting spatial vision: towards a unifying model. *J. Opt. Soc. Am. A*, 17: 1899–1917.

Itti, L., Koch, C. and Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 20: 1254–1259.

Jahne, B. and HauBeker, H. (eds) (2000) Computer Vision and Applications: A Guide for Students and Practitioners. New York: Academic Press.

Jordan, M.I. (1998) *Learning in Graphical Models.* Cambridge, MA: MIT Press.

Julesz, B. (1960) Binocular depth perception of computer-generated patterns. *Bell System Technical J.* 39: 1125–1162.

Kass, M., Witkin, A. and Terzopoulos, D. (1987) Snakes: active contour models. *Proc. ICCV*, 259–268.

Kittler, J. (1998) Combining classifiers. *Patt. Anal. Appl.* 1: 18–27.

Koch, C. and Itti, L. (2001) Patent pending. Filed July 23, 2001, following provisional applications No. 60/274,674 filed March 8, 2001 and 60/288,724 filed May 4, 2001 http://ilab.usc.edu/publications/patent.html.

Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4: 219–227.

Lee, D.K., Itti, L., Koch, C. and Braun, J. (1999) Attention activates winner-take-all competition among visual filters. *Nature Neurosci.* 2: 375–381.

Li, Z. and Hertz, J. (2000) Odour recognition and segmentation by a model olfactory bulb and cortex. *Network*, 11: 83–102.

Mahowald, M. (1994) *An Analog VLSI Stereoscopic Vision System*. New York: Kluwer Academic.

Markram, H. (2002) Structural and functional principles of neocortical microcircuits. Available at http://www.igi.tugraz.at/telematik/tele1-02_markram.pdf.

McCabe, S.L. and Denham, M.J. (1997) A model of auditory streaming. *J. Acoust. Soc. Am.*, 101: 1611–1621.

Mead, C.A. (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison Wesley.

NeuVoice Ltd (http://www.neuvoice.com/).

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems.* San Francisco, CA: Morgan Kaufmann.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. and Bialek, W. (1997) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Rolls, E.T. and Milward, T. (2000) A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.*, 12: 2547–2572.

Rybak, I.A., Gusakova, V.I., Golovan, A.V., Podladchikova, L.N. and Shevtsova, N.A. (1998) A model of attention-guided visual perception and recognition. *Vision Res.* 38: 2387–2400.

Shapiro, L.G. and Stockman, G.C. (2001) *Computer Vision.* Englewood Cliffs, NJ: Prentice-Hall.

Sharma, R.K., Leen, T.K. and Pavel, M. (2001) Bayesian sensor image fusion using local linear generative models. *Optical Eng.*, 40: 1364–1376.

Shepard, R.N. and Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science*, 171: 701–703.

SpikeNet. http://www.spikenet-technology.com/.

Sumner, C.J., Lopez-Poveda, E.A., O'Mard, L.P. and Meddis, R. (2002) A revised model of the inner hair cell and auditory-nerve complex. *J. Acoust. Soc. Am.*, 111: 2178–2188.

Terzopoulos, D. and Szeliski, R. (1992) *Tracking with Kalman Snakes, in 'Active Vision'*. Cambridge, MA: MIT Press.

Tsotsos, J.K. *et al.* Modeling visual-attention via selective tuning. *Artif. Intell.* 78: 507–545.

Turk, M. and Pentland, A. (1991) Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3: 1.

van Schaik, A. and Meddis, R. (1999) Analog very large-scale integrated (VLSI) implementation of a model of amplitude-modulation sensitivity in the auditory brainstem. *J. Acoust. Soc. Am.*, 105: 811–821.

Young, S.J. (2001) Statistical modelling in continuous speech recognition (CSR).*Proc. Intl Conf. Uncertainty in Artif. Intell.*, Seattle, WA.

CHAPTER

5

Speech and Language

William D. Marslen-Wilson

1 INTRODUCTION

This chapter focuses on speech in relation to language, and language in relation to speech. This focus is in the context of the Foresight Cognitive Systems Project, which set out to see where bridges can be built between research on artificial cognitive systems and research in neuroscience and related areas.

Experience suggests that such bridges can exist only where there is interest and motivation on both sides, and that they can be maintained only where it is possible to set out a structured program for future research with achievable medium-term goals. This requires a well-specified scientific characterization of the shared domain.

A desirable feature for artificial cognitive systems, assuming some autonomy as independent computational agents, will be the ability to communicate with natural (human) cognitive systems. Since the preferred human mode of communication is through speech, a potential shared interest is in understanding how speech conveys meaning. From the life sciences side, this is the problem of how the human brain is functionally and neurally organized to support this process.

This document argues that the study of speech and language in the context of modern cognitive neuroscience – incorporating recent advances in neurobiology and driven by new techniques in neuroimaging – should, over the next two decades, lead to major breakthroughs in our scientific understanding of these crucial human capacities. This offers the potential for a detailed neuroscientific analysis of a major natural cognitive system. This is a domain that can interact very profitably with the exercise of building explicit, implementable computational models of speech and language processing systems.

2 STATE OF THE ART

The scientific study of speech and language is entering a period of transition. From its origins as a field of study rooted in the cognitive sciences and in the humanities, with a strong reliance on the empirical and theoretical techniques of classical experimental psychology, it is now coming to grips with the challenge of an emerging cognitive neuroscience. This challenge is driven by the emergence of new techniques for imaging

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

the activities of the intact human brain, and by strong inputs from neurobiological sources – in particular neurophysiological and neuroanatomical research into auditory processing structures and pathways in non-human primates.

In this dynamic scientific context, any characterization of the state of the art will overlap considerably with the future direction of the field and emerging developments. The discussion on the following pages, therefore, should be seen in this light.

2.1 Language and Speech – the Cross-Disciplinary Challenge

The evolution of a cognitive neuroscience of language requires us to bring together several different strands. Collectively, these constitute both the 'state of the art' and the raw material out of which the future science of language will be constructed.

2.1.1 The Cognitive and Psycholinguistic Framework for the Study of Language Function

The historically most important and most extensive inputs to the study of language as a cognitive system are the data and theory generated by experimental psycholinguistics, incorporating influences from linguistics, computer science and computational modelling. In the 50 years since the emergence of modern linguistic theory, a standard view has come to predominate, mapping primary distinctions in types of linguistic knowledge onto a model of the psychological system assumed to represent and deploy this knowledge in actual human performance. It is from this source that we acquired the basic segmentation of linguistic knowledge into different categories, which has strongly constrained the kinds of scientific questions that have been and are being asked.

Every student in the field will be familiar with the typical four-way classification into:

- knowledge of linguistic form (sounds and letters)
- knowledge of words – the mental lexicon
- knowledge of grammar (or syntax) – how words fit together to form meaningful sentences
- knowledge of meaning (or semantics).

The scientific study of language – whether in language acquisition (how language is learned as a first or second language), language production or language comprehension – takes for granted that these are the categories of linguistic knowledge that are learned and mentally represented by the language user, and that must be deployed swiftly and effectively in the on-line processes of speaking, listening and understanding.

Research within this framework has led to detailed hypotheses about the content and structure of the mental representations of linguistic knowledge, and about the mechanisms that operate on this knowledge in real-time performance. Out of this has emerged a clear appreciation of the complexity and richness of language as a human cognitive system.

The field has developed a wealth of hypotheses and observations about representations and processes at all levels of the system, ranging from views on the nature of phonological representations and how they are accessed from the speech stream, to those about the organization of the mental lexicon, and about the nature and representation of linguistic and conceptual meaning. Accompanying this, there has been a tremendous development in the sophistication of experimental and statistical methodologies available for probing the functional properties of the language system, and a keen awareness of the many traps that language prepares for the unwary experimenter.

These achievements make the concepts and techniques of experimental psycholinguistics an essential component of the future science of language. Without them, we can neither grasp what it is we are trying to explain, nor put together adequate experiments to probe the detailed properties of the human language system. None the less, history also suggests that, on its own, this approach is not sufficient to pin down the underlying scientific facts of the matter.

Despite the accumulation of detailed observations and hypotheses about many aspects of the language system, there remain long-standing disagreements about its underlying properties. A case in point is the mental lexicon – the stored mental representation of what we know about the words in our language. This type of knowledge plays a crucial functional role in the language system. We can neither understand language nor produce it without reference to these lexical representations, which link together knowledge of the phonological form of words and knowledge of word meanings.

Lexical processing in this environment imposes complex functional demands on the system. These include the basic processes of access and selection – how, as the speech signal is heard, listeners relate this sensory information to their knowledge of words in the language and identify the word being uttered.

There is good agreement about the basic functional properties of these processes, centring around the view that lexical access is based on the simultaneous activation of multiple word candidates, and that the selection of a single correct candidate is based on a relational process of choice between these competitors. Selection of the correct candidate reflects not just the evidence favouring that specific candidate, but also the relevance of this evidence to other possible candidates. One source of evidence for this is the timing of word-recognition processes, showing that spoken words can be recognized as soon as they diverge from their cohort of competitors (e.g., Marslen-Wilson, 1987).

There is major disagreement, however, about the nature of the functional system supporting this rapid, efficient, on-line recognition process. The most fundamental issue concerns the underlying computational properties – whether they are based on distributed representations or on a network of interacting localist nodes. Both types of approach can capture the main cohort phenomena, but they do so in very different ways, leading to contrasting approaches both to the nature of psycholinguistic representation and to the mechanisms involved in spoken language processing.

Localist models approach the mental lexicon in terms of classical concepts of symbolic computation. The form and meaning associated with a given word is represented in terms of labelled symbolic nodes. Lexical access and selection depend on excitatory and inhibitory interactions between these nodes, at different levels of representation, as words are seen or heard.

In distributed models, the mapping from form to content emerges as a function of the learned pattern of connections among the elementary units in a multi-layer neural network. Nowhere in the system is either the phonological form or the semantic content of the word explicitly represented, except indirectly as a pattern of connections over units. As a word is heard, overlapping patterns of activation, corresponding to different lexical interpretations of the material heard up to that point, will be simultaneously activated and will interfere with each other to form a blend state. This will resolve over time into a unique representation corresponding to the word being heard, as a function of the properties of other words in its competitor neighbourhood.

These are two very different ways of explaining the functional properties of the same cognitive process. One finds this state of affairs repeated across the field. Although there is agreement about the functional properties of the human speech comprehension system, there can be quite outstanding levels of disagreement about how these phenomena are to be explained. Thus, despite the methodological and theoretical sophistication of current psycholinguistic research over the past two or three decades, there is little doubt that additional constraints are needed to converge on a unique scientific account.

2.1.2 Cortical Localization of Language Function

A second critical strand is the long-standing tradition of research into the brain

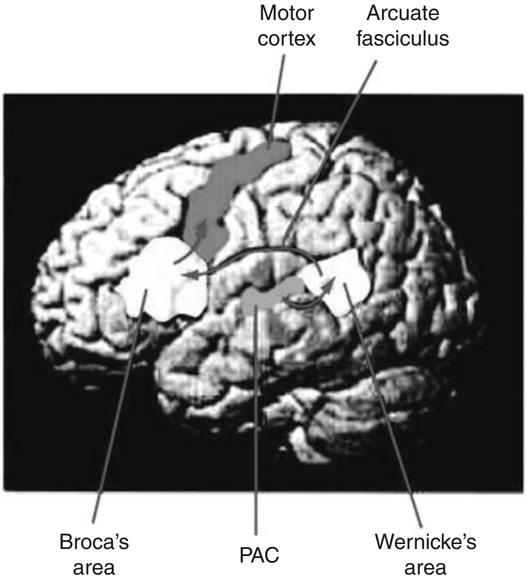


FIGURE 5.1 The classic Broca–Wernicke diagram linking brain regions and language functions (PAC is the primary auditory cortex). (This figure appears in the colour plate section)

bases of human language function, developing from the Broca–Wernicke–Lichtheim framework of the nineteenth century, linking to the more functionally oriented fields of neuropsychology (and later cognitive neuropsychology) in the second half of the twentieth century, and beginning to re-invent itself in the last decade with the advent of new techniques for neuroimaging. The nineteenth-century framework, evolving out of the needs of the neurologist for bedside diagnosis of aphasic patients, typically dealt in broad-brush functional and neural labels (see Figure 5.1).

Disorders of comprehension were associated with damage to the superior temporal lobe ('Wernicke's area'), for example, while problems in language production – so-called telegraphic speech, for example – were associated with damage to Broca's area in frontal cortex. More subtle aphasic deficits were analysed in terms of damage to connections between these areas and primary sensory and motor areas, but the grain of neural and functional analysis remained relatively coarse.

The emergence of neuropsychology brought with it more detailed functional analyses of the language system and its deficits – especially in the sub-field of cognitive neuropsychology – and introduced a more active hypothesis-testing approach, reflecting the hard-earned advances in experimental method of mid-twentieth-century psychology. However, despite the fact that the primary empirical input to the field was the effects of injury to the brain on sometimes very specific aspects of language function, linkage to the function of particular brain areas was never the most successful feature of this research programme.

If a patient exhibited, say, a specific deficit in reading words with irregular pronunciations but not in saying them, this might lead to a correspondingly specific claim about the functional organization of the language system. But it would typically not lead to claims about which brain systems were responsible for the postulated functions. This lack of specificity was at least in part a response to the awkward messiness of data from the damaged brain, where sometimes the most specific deficits seemed to emerge from the largest lesions, and where experimental, replicable science was rarely possible.

A lack of detailed engagement with the neural level is no longer tenable in the era of neuroimaging, where research with positron emission tomography (PET), function magnetic resonance imaging (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG) with both normal and brain-damaged populations could deliver direct information about what brain areas are involved in which cognitive tasks (see the Foresight Research Review, 'Advanced Neuroscience Technologies'). A number of landmark papers are emerging, combining neuropsychological and neuroimaging data to breathe new life into the enterprise of cortical localization of function. Relatively few successful papers have emerged, however, in the domain of speech and language. In part, this is because the concepts of language invoked have tended to be rather primitive borrowings from secondary sources, and in part because the focus has tended to be on written rather than spoken language.

2.1.3 Neurobiological Constraints

The third strand, whose influence is rapidly increasing, but which until very recently had no discernible role in the mainstream study of speech and language, is constituted by the insights and constraints that derive from the neurobiology of the related brain systems. This is primarily driven by neuroanatomical and neurophysiological research with non-human primates, in particular our rather distant relative the macaque.

This link was historically weak and unexplored because of the general view that language is special to humans and that there is no 'animal model'. This may in critical respects still be true (cf. Hauser *et al.*, 2002), but it is now widely accepted that there must be structural parallels and evolutionary precursors for many aspects of the neural systems that instantiate language function in the human brain. The most specific parallels are likely to be found in the organization of the primate auditory processing system, where convincing evidence has emerged to suggest a basic 'dorsal/ventral' distinction in the organization of the major processing pathways emerging from primary auditory cortex, as well as a wealth of information about the structure and function of the auditory cortex itself.

The kind of picture that emerges, while bearing a partial resemblance to the Broca–Wernicke diagram (see Fig. 5.1), is more much specific about the functional and neural architecture of the system than anything that can we can say at present about the human system (see Fig. 5.2). This picture also contains an entire auditory processing system (the ventral route) that does not figure in the classical Broca–Wernicke diagram.[1]

In research with the macaque we can establish directly the detailed neuroanatomy of the pathways linking areas in auditory cortex (labelled 'core' and 'belt' in Fig. 5.2) to other cortical regions, while at the same time using single-cell recording techniques to specify the functional properties of the areas involved. Current views (e.g., Rauschecker and Tian, 2000) suggest that the ventral 'what' route, involving superior temporal lobe pathways, is specialized for processing and identification of complex auditory objects, including species-specific vocalizations, while the dorsal 'where' route, and the regions in auditory cortex from which it derives (CL and CM in Fig. 5.2) are more involved in the spatial analysis of auditory information.

[1] This reflects another weakness of the classic neurological approach to language function. The primary source of data is brain damage caused by stroke, which reflects the cortical areas served by different cerebral arteries. Areas that tend not to suffer stroke damage, or not to be selectively damaged, such as the anterior pole of the left temporal lobe, are less likely to have functions attributed to them.

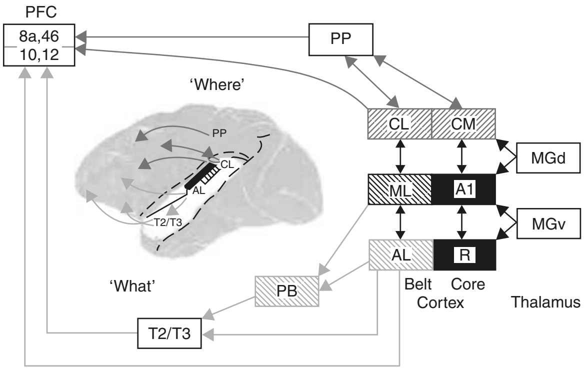


FIGURE 5.2 Schematic flow diagram of auditory processing streams in the primate brain. The dorsal 'where' route connects via posterior parietal cortex (PP) to prefrontal cortex (PFC). The ventral 'what' route connects either directly (from AL) or via temporal lobe structures (T2/T3) to a different set of PFC areas. (Reproduced from Rauschecker and Tian, 2000. Copyright © 2000. National Academy of Sciences, USA). (This figure appears in the colour plate section.)

It is tempting to adopt this analysis as a template for thinking about the organization of the speech and language processing system in the human brain, although two major caveats are in order. The first is that the human brain diverges in many respects from the macaque brain, most extensively in the anterior temporal lobe and frontal lobe areas that are critically involved in the systems postulated. The second caveat is that a system designed to support spoken language will need to make different and additional functional demands to those served by the macaque system. None the less, the emergence of a well-specified account of the neurobiological underpinnings of primate auditory processing has had important consequences. It provides a model for what a theory of these systems needs to look like, in terms of the specificity of both the functional and the neural account that is provided, and strongly suggests a very different approach to the characterization of human language function.

Classical cognitive and psycholinguistic approaches to the functional structure of the system for mapping from sound to meaning have always assumed that a single, unitary process (or succession of processes) is engaged to carry out this mapping. The neurobiological evidence suggests, however, that the underlying neural system is not organized along these lines, and that multiple parallel processing streams are involved, extending hierarchically outwards from auditory cortex. In this respect, speech and language analysis would be brought into closer alignment with long-held assumptions about the organization of primate visual processing systems – from which the dorsal-ventral 'what/where' dichotomy originally derives (see Learning and Memory summary (Chapter 9) for further discussion).

2.1.4 Speech Processing in the Human Brain

A fourth strand, closely related to the third, is the recent development of functional imaging studies of the perception of speech and other complex sounds. These studies are starting to build up a picture of the neural and functional organization of human auditory cortex along the lines indicated by studies of the macaque (for a recent review, see Scott and Johnsrude, 2003).

An important feature of these studies, and one shared by the most successful contributions so far, is the close involvement of concepts and techniques from mainstream research in psycho-acoustics and acoustic-phonetics. Without a precise understanding of the properties of the acoustic signal for

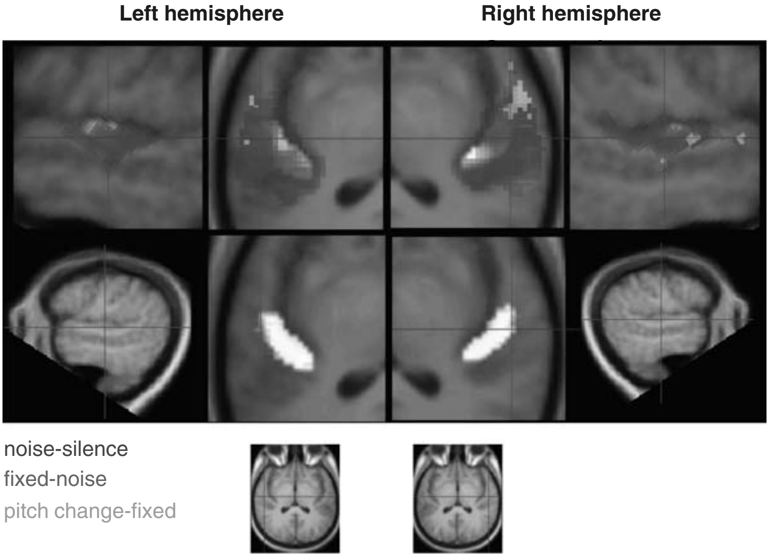


FIGURE 5.3 fMRI responses in auditory regions: The white area in the central panels denotes the average location of Heschl's gyrus (primary auditory cortex) in nine subjects. Noise, compared to silence, bilaterally activates a large auditory area (shown in blue). Extraction of pitch based on temporal cues activates an area at the lateral edge of Heschl's gyrus in both hemispheres (red) in the auditory belt region. Finally, introduction of a pitch difference between successive tones (melody) produces right-dominant activation, at the lateral tip of Heschl's gyrus (green/aqua). (This figure appears in the colour plate section)

speech, and a detailed functional analysis of the types of information that can be extracted, it is difficult to construct experimental contrasts that do not contain unfortunate confounds.

More generally, what is striking is the degree of functional and neural specificity that seems to be possible with the combination of fMRI and appropriate stimulus contrasts. Figure 5.3, based on a paper by Patterson and colleagues (2002), is a case in point. Here we see primary auditory cortex (Heschl's gyrus) precisely located in the listeners' brains, with a broad area, including Heschl's gyrus, activated by a basic contrast between sound and silence, and then more specific areas activated according to their different functional specializations. In this study, focusing on the perception of pitch and of melody (produced by variation of pitch), these two dimensions activate different cortical territories, with areas sensitive to melody showing additional hemispheric specialization. Similar findings are emerging from parallel studies contrasting speech and non-speech materials.

Two important design features of the neural and functional organization of human cortical speech processing are emerging from this research. The first is the apparent confirmation of the hierarchical structure of processing layers in primary auditory cortex. fMRI studies show that core areas, receiving direct input from lower level auditory centres in the thalamus, seem to respond equally well to all auditory inputs. These core structures project to surrounding belt (and 'parabelt' areas), which respond preferentially to more complex sounds, with some specialization emerging for different belt and parabelt areas. An intense focus of current research is the further differentiation of these areas, with the research summarized in Fig. 5.3 being one example of this.

The second design feature, and one that will have particular importance for the extension of this research to deal with the higher-order structure of the speech understanding process, is the accumulating evidence for multiple processing streams, emerging in parallel from primary auditory cortex and from surrounding areas specialized for the processing of complex auditory inputs such as speech.

The clearest results are from recent fMRI studies that examine the neuroimaging consequences of variations in speech intelligibility. By using techniques to vary intelligibility that maintain basic acoustic properties across different stimulus types, these studies avoid confounds present in earlier studies. This research show a clear processing progression moving anteriorly down the superior temporal lobe from primary auditory areas (e.g., Scott *et al.*, 2000; Davis and Johnsrude, 2003), a pattern of activation that looks very much like the kind of 'ventral' stream identified in research with macaques (see Fig. 5.2). There is also persuasive evidence for the hierarchical organization of these pathways.

Thus, in the research summarized in Figure 5.4 we see a central area of activation (coloured red to yellow) corresponding to regions surrounding primary auditory cortex on the left, which is sensitive to relatively low-level variations in the acoustic-phonetic properties of different types of disrupted speech.

Moving away from this central area, we see further areas of activation (coloured green to blue) that are not sensitive to acoustic-phonetic variation but which respond to changes in intelligibility – whether the sentence and the words in it can be successfully understood. Interestingly, these further areas of activation not only extend ventrally, moving forward down the superior temporal lobe, but also dorsally, in that there is a strong swathe of activation moving posteriorly, to the junction between the temporal and parietal lobes, consistent with a dorsal route connecting through to frontal areas (see also Figs 5.1 and 5.2). It would be a mistake, however, to restrict the range of possible processing streams just to the human analogues of the dorsal and ventral streams identified in non-human primates. There is accumulating evidence for a third stream of processing, also moving posteriorly but oriented downwards towards inferior temporal structures, which seem to be implicated in semantic processing of auditory inputs (e.g., Price, 2000). There is also some neuropsychological evidence to support the notion of a 'basal language area', and this may be a target for this potential third stream.

Of course, as we begin to develop a more differentiated view of the functional architecture of the human system, we will have to revise these initial attempts to classify and enumerate possible processing streams. We should remember that information flow in the brain is never unidirectional. Feedback to temporal lobe centres from more anterior

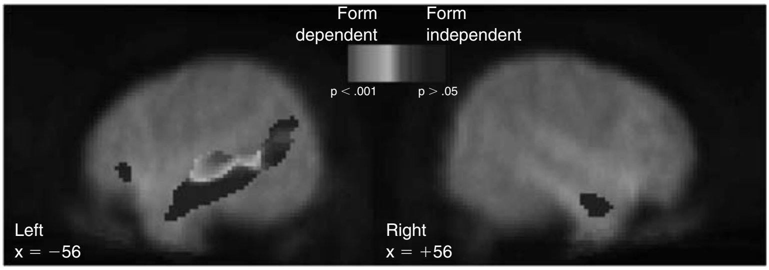


FIGURE 5.4 Activation as a function of intelligibility for auditorily presented sentences. The colour scale shows intelligibility-responsive regions which showed a reliable difference between different forms of distortion (orange to red) or no reliable difference between distortions (green to blue). See Davis and Johnsrude (2003). (This figure appears in the colour plate section)

or inferior processing centres may be just as important as the traffic in the opposite direction.

Finally, we should point out that auditory processing has intrinsic and critical dynamic properties. The speech signal is delivered rapidly over time. Speech and language processing systems in the brain track this temporal sequence with millisecond fidelity. However, fMRI is poorly adapted to track this millisecond-level dynamic, since it reflects changes in blood-flow to different parts of the brain, a process working on a timescale of seconds rather than milliseconds.

To track cortical responses with the necessary degree of temporal resolution requires techniques such as EEG and MEG. These measure, respectively, the electric and the magnetic fields generated by brain activity. These reflect directly the changes in neural activity accompanying different processing operations, and could be of great value in separating out their properties. They are also essential if we are going to be able to examine the timing of processing relations, and the role of feedback in modulating online analysis and interpretation.

2.1.5 The Cross-Linguistic Dimension

This final strand makes the point that it is language that is at issue here, not just speech perception, and that the content and structure of a given language system may itself be very variable. Human languages vary widely in how they organize the speech stream to communicate meaning, ranging from contrasts in the speech sounds that are specific to a given language, to the striking differences, at higher levels of the system, in the ways that larger units combine to form words and sentences. The study of cross-linguistic variation will be critical in research into specialized speech processing areas distributed around primary auditory cortex, and into the different processing streams thought to emanate from these areas.

One reason is that incorporating cross-linguistic distinctions into the research programme can provide useful empirical contrasts. For example, in the attempt to discern the processing properties of specific cortical areas, especially at the tricky boundary between speech and non-speech, it is useful to compare responses across sets of listeners whose languages differ in the phonetic distinctions they make.

A well-known study by Näätänen and colleagues (1997), for example, examines responses to the same vowel sound for a set of listeners where this was a speech sound in their language and for a different set of listeners where this was not a distinct vowel. Using MEG, these authors showed that early cortical responses to speech could be modulated by the linguistic role of the sounds being heard. fMRI experiments can exploit similar contrasts to give a more precise spatial localization of the processing areas involved in supporting this distinction.

The main reason, however, for working cross-linguistically is simply that we cannot determine the fundamental design properties of human speech and language as a neurobiological system on the basis of just one or two languages. If, for example, through the study of speakers of English, we come to a view about the core properties of different cortical processing streams, we would have no way of separating out general principles from the idiosyncrasies of English if that was the only language for which this kind of information was available. Nor could we even begin to adjudicate between contrasting views about the role of innate factors as opposed to developmental experience in determining the observed properties of these different processing routes.

To take a topical example, English is a language where a great deal of linguistic and communicative work is done by morphologically simple words – that is, words like *dog*, *smile* or *dark*, which are made up of a single meaning unit (or morpheme), with no further internal structure. In terms of organizing access to meaning, the processing system is confronted with a fairly straightforward challenge – how to map a relatively invariant surface unit onto an internal representation of its meaning and grammatical properties. However, many words, even in English, do not have a simple structure, and are made up of more than one phonological and morphological unit. In English, for example, the verb *smile* may occur in the complex past tense form *smiled*, where smile is combined with the past tense morpheme {-d}, or the word *dark* may occur as part of the abstract noun *darkness*, where it combines with the derivational morpheme {*-ness*}.

In other languages, such as Italian or Polish, essentially every surface form is complex, and stems almost always appear in combination with a derivational or inflectional affix. The contrast is even stronger for the Semitic languages, such as Arabic and Hebrew, where all surface forms are argued to be complex combinations of morphemes, and where words are formed not by chaining morphemes together in sequence (as in *dark* + *ness*), but by interleaving morphemic elements. Thus, the word *kataba* in Arabic, meaning 'write', is formed by interleaving the abstract root {ktb} with a word pattern which specifies the sequence of consonants and vowels, as well as the syntactic role of the resulting word. These more complex words present additional challenges to the processing system and may place differential demands on different processing pathways.

A case in point, returning to English, is the contrast between regular and irregular past tense forms. This played an important role in recent debates in cognitive science about the nature of mental computation and the structure of the language system (Marslen-Wilson and Tyler, 1998; Pinker, 1999). Moving away from the focus on rules and symbolic computation which dominated earlier stages of the debate, Tyler, Marslen-Wilson and colleagues focused on the status of irregular past tenses (such as *gave*, *bought*) as phonologically simple forms that can access stored lexical representations directly. The regular past tenses, on the other hand, because they are complex combinations of stems and affixes (as in *smiled*, *jumped* etc.), require further analysis processes.

Evidence from neuroimaging and neuropsychology (Tyler *et al*., 2002) implicates specialized morpho-phonological processing mechanisms in left inferior frontal cortex, possibly linked to dorsal as opposed to ventral processing streams. Further investigation of this possibility, however, requires research cross-linguistically, working with systems where contrasts in regularity can be dissociated from contrasts in phonological complexity.

A more dramatic type of cross-linguistic contrast that may be uniquely valuable in elucidating the underlying properties of cortical speech and language areas comes through the comparison between spoken languages and native sign languages, such as BSL (British Sign Language). These are languages where the primary medium is gestural (visuo-spatial) in nature, and which are learned as first languages, typically by the congenitally deaf children of deaf parents who are also signers.

A series of increasingly sophisticated studies show that, despite the difference in the sensory modalities involved, there is considerable overlap in the cortical areas activated when perceiving spoken and signed languages (e.g., Petitto *et al*., 2000; MacSweeney *et al*., 2002). These include not only inferior frontal regions (Broca's area), but also superior temporal regions classically associated with the processing of spoken language. Evidence from MacSweeney's study shows that in deaf native signers (although not hearing native signers) the areas activated in the perception of sign include secondary auditory cortex, in regions close to Heschl's gyrus (see Fig. 5.3). This reveals unsuspected plasticity in cortical regions normally associated with the processing of complex auditory inputs, but may also reflect underlying similarities in the types of analyses being conducted at this level of the system on communicative inputs, whether auditory or gestural in origin.

More generally, I would like to stress the importance of not only cross-linguistic but also developmental and genetic inputs to a full understanding of how and why the adult brain is structured to support the astonishing human capacities in the domain of speech and language.

3 IMPLICATIONS AND FUTURE DEVELOPMENTS

At the beginning of this document I claimed that over the next two decades the study of speech and language in the context of modern cognitive neuroscience, incorporating recent advances in neurobiology, provides a framework that should lead to major breakthroughs in the quality and specificity of our scientific understanding of these crucial human capacities. I then indicated some of the different strands that will need to be woven together to achieve these goals.

In effect, we need to link the concepts and techniques of experimental psycholinguistics, psychoacoustics and acoustic phonetics, in the service of a systematic experimental analysis of the cortical systems supporting human speech processing and language interpretation, using neuroimaging techniques to track these processes in space and in time, and working cross-linguistically to ensure the scientific generality of the account being constructed.

In this context, the clearest potential links with future developments in artificial cognitive systems in the domain of computational modelling and analysis of the neural and functional systems supporting speech and language. We will only know that we have fully understood a natural cognitive system when we can build a model of it.

4 OPEN QUESTIONS

The study of speech and language has many open questions. I list just a sample.

- What is the functional architecture of the human language processing system, and how are the conventional distinctions among knowledge-types (sounds, words, grammar, meaning) to be interpreted in a neurobiological framework?
- What are the underlying computational properties of the system? Will a uniform type of computational process be adequate to capture the characteristics of language representation and processing?
- How do we constrain these computational accounts to be neurobiologically realistic?
- What is an appropriate framework for linking studies of the damaged brain to neuroimaging studies of the intact brain? How will this link to potential translational research?
- What are the limitations of the non-human primate model as applied to the explanation of human systems? How should we interpret the notions 'where/what' in the context of human language?
- What is the functional neuroanatomy of the human speech and language system, and how far does it diverge from the standard macaque model?
- To what extent do cross-linguistic variations in language structure affect the properties of different cortical processing streams?
- What are the commonalities of language in the brain across spoken and sign languages?

References

Davis, M.H. and Johnsrude, I.S. (2003) Hierarchical processing in spoken language comprehension. *J. Neurosci*, 23[8]: 3423–3431.

Hauser, M., Chomsky, N. and Fitch, T. (2002) The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298: 1569–1579.

MacSweeney, M., Woll, B., Campbell, R., McGuire, P.K., David, A.S., Williams, S.C.R., Suckling, J., Calvert, G.A. and Brammer, M.J. (2002) Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain*, 125: 1583–1593.

Marslen-Wilson, W.D. (1987) Functional parallelism in spoken word-recognition. *Cognition*, 25: 71–102.

Marslen-Wilson, W.D. and Tyler, L.K, (1998) Rules, representations, and the English past tense. *Trends Cogn. Sci.*, 11: 428–435.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J. and Alho, K. (1997) Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385: 432–434.

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S. and Griffiths, T.D. (2002) The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36: 767–776.

Petitto, L.A., Zatorre, R.J., Gauna, K., Nikelski, E.J., Dostie, D. and Evans, A.C. (2000) Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. *Proc. Natl Acad. Sci.*, 97: 13961–13966.

Pinker, S. (1999) *Words and Rules: The Ingredients of Language*. London: Weidenfeld and Nicolson.

Price, C.J. (2000) The anatomy of language: contributions from functional neuroimaging. *J. Anat.*, 197 Pt 3: 335–359.

Rauschecker, J.P. and Tian, B. (2000) Mechanisms and streams for processing of 'what' and 'where' in auditory cortex. *Proc. Natl Acad. Sci. USA*, 97: 11800–11806.

Scott, S.K., Blank, C.C., Rosen, S. and Wise, R.J.S. (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123: 2400–2406.

Scott, S.K. and Johnsrude, I.J. (2003) The neuroanatomical and functional organization of speech perception. *Trends Neurosci.*, 26 (2): 100–107.

Tyler, L.K., Randall, B. and Marslen-Wilson, W.D. (2002) Phonology and neuropsychology of the English past tense. *Neuropsychologia*, 40: 1154–1166.

SECTION

3

Cognitive Systems in Action

Natural and artificial cognitive systems plan their actions. In doing so, they need to consider the broader context of others – other people or other computing devices.

Section Contents

This page intentionally left blank

Cognitive Systems Need to Plan, to Interact with Others and to Act: Introduction

Richard Morris and Lionel Tarassenko

Action is the business of doing things. People act. Computers act. They need to plan, and that planning reflects both our human existence as social creatures interacting with other humans and the fact that computers increasingly exist within networks in which their actions are influenced by, and in turn influence, the actions of other computers.

In acting within their worlds, animals and humans display superb abilities to choose and execute brief and more extended courses of action. Sometimes actions are deliberate, sometimes automatic. In the former case, planning and motivation are critical to the action selection process. Research into this process, and into the ability of both living and autonomous intelligent agents to select, execute and adapt their actions to a changing environment, encompasses multidisciplinary studies from psychology to neuroscience, and from computer science to information engineering.

In their chapter entitled Action Phil Barnarde and Peter Redgrave outline a formal framework for thinking about action selection and execution by animals. Their framework encompasses actions ranging from straightforward 'innate' motor sequences such as 'fight or flight' when an animal is faced with a predator, through to the complexities of the ostensibly simple action of a person reaching out to an object, picking it up, passing it in an appropriate way to another person, and then releasing it safely.

Different structures in the mammalian brain are involved in different actions, with a hierarchy of control levels, starting, at the lowest level, at the brain-stem and spinal cord, working through the basal ganglia and cerebellum all the way to the neocortex. The major challenge facing research into action is the apparently distributed nature of the control, given the simultaneous and critical influence of so many and such diverse structures. Formal theories, such as temporal difference learning, provide a framework for thinking about action selection.

Uta Frith and Sarah-Jayne Blakemore review the social dimension of cognition. They tackle a topic that neuroscientists have, until recently, been shy of addressing.

For Frith and Blakemore, social cognition is any cognitive process – implicit or explicit – that involves other people. One example is work on the ability to recognize faces, a social skill present in early infancy in a baby's reactions to its mother. It is also a skill that develops through life to the point where we can quickly recognize the slightest nuance in a friend or partner's face that may reflect their mood – such as anger.

Using similar experimental techniques to those of other aspects of contemporary neuroscience, including single-cell recording and

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

non-invasive brain imaging, we have a developing set of ideas about the specific areas of the brain involved in face recognition, what we can learn from the expressions of a face and more generally in 'mentalizing'. This work is complemented by related studies on perception of gaze and 'mirror neurons' that participate in the perception of action.

Brain imaging experiments, in particular, are entering hitherto uncharted territory by looking at such subtle processes as 'feelings' ranging from fear through to embarrassment and even empathy. Such studies may help to throw light on what we often think of as 'abnormal' behaviour in a person, or even extremes of distinctive social interaction as occurs in autism. Such research is guided by ideas about the human capacity to recognize that others have minds as well, and our ability to alter our behaviour towards others as a function of this recognition. One prominent idea is that autism is thought to impair the capacity to possess a 'theory of mind' and that this abnormality may arise through subtle genetic mutations that affect the anatomical development of relevant areas of the brain during gestation and early life.

Many see the computer on their desk as a lonely, lifeless tool, a very useful one to be sure, but not one with a rich social life. Not so. Increasingly, personal computers are harbingers of software that can act as 'intelligent agents' that seek out, find and process specific types of information for its owner. Nicholas Jennings and Anthony Cohn, in their chapter on Interaction, Planning and Motivation, take us into the world of the software agent – the worker bees of information technology. Soon, they surmise, we will all have agents that could plan our travel, respond to changed circumstances in doing this, and do it legally, safely and within the constraints that so many of us have in busy lives. To carry out this kind of task, these agents must interact with other agents and the computing systems in which they reside.

For an autonomous agent to act, no less than an animal fleeing a predator or a person preparing a meal, it must have motivation to do so. This motivation is partly a matter of mere 'push' – the energy to do it – but it is more than that, as the specifics may guide the choice of action, just as the meal a person prepares depends on the person or group for whom it is prepared and the occasion in question. Similarly, motivation and the planning that necessarily goes with it, influence the style of software tools that a computing engineer may develop for a particular range of tasks. A key development in current research is moving on from fixed action patterns and sequences to flexible systems in which it is acknowledged that you cannot predict accurately and completely what will happen when you start a task, but need instead to monitor things as the task progresses.

These topics may seem somewhat disparate: and, to date, research in the relevant domains of neuroscience and computer science has proceeded independently. But they clearly have common themes. People interact with people, and they need specialized mind/brain systems to do so. People interact with computers, and computers with other computers. As computing becomes all-pervasive, as people come to rely more and more on the handheld devices of their personal digital environments, we should anticipate that both natural and artificial social agents will need to discover more and more about each other to function effectively. At present, most human interaction with machines lacks any degree of social sophistication on the part of the machine, but this may change as their software starts to incorporate ideas emerging from the study of social cognition. When it comes to cognitive systems in action, there are many open questions ahead. These three challenging chapters identify some of these and show once again how much in common there is between life scientists and researchers in the physical sciences.

C H A P T E R

6

Action

Philip Barnard and Peter Redgrave

1 INTRODUCTION

Animals have superb innate abilities to choose and execute simple and extended courses of action. Many animals also exhibit envious adaptability to new and changeable environments, exploring efficiently, learning appropriately and exploiting resources near optimally. Understanding these capabilities has been the focus of wide variety of neural (i.e. implementational, in the sense of Marr), psychological (algorithmic) and computational studies. In this document we consider these studies, which have historically involved substantial interactions between psychology and neuroscience, and engineering and computer science.

We adopt a very broad definition of action, encompassing essentially any emitted or elicited behaviour of an organism. We thus include both deliberate actions, chosen either explicitly or implicitly to achieve particular goals, and habits, reflexes or tropisms that are exercised more automatically in response to external stimuli and internal states (though the actions may still achieve goals). One might also consider internal actions involved in the control of ongoing processing, such as the allocation of sensory attention to a particular stimulus or region of sensory space.

Modularity has played a critical role in attempts to understand the biological control of action and in building robots. We must therefore address it at the outset. At least three different sorts of modularity have been influential: the hierarchical abstraction of state, goals and actions; a division into separate systems for grossly separate functions (such as defensive behaviour); and the separate consideration of the specification, selection and execution of actions.

The first form of modularity concerns hierarchical abstraction. In cognitive and computational theory, this is often seen as proceeding from sensation, through levels that capture the abstract forms of individual

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

sounds, objects, movements or bodily experiences and their organization into external scenes or bodily configurations; extending to 'higher level' structures of human knowledge, including phonology, syntactic regularity, the nature of semantic reference underlying human language (see Chapter 5, Speech and Language), as well as other multimodal contingencies and types of metarepresentation.

As an example of hierarchical abstraction, consider a motor sequence that can be described in terms of the set of physical movements, for example, grasping a cylindrical object, picking it up, moving it to a point in space where someone else can grasp it, and releasing the object to complete the transfer. We can describe this sequence equally well in terms of a referential, or propositional meaning. For example, we could describe it as an action involving Jack giving Jill a cup of tea.

At another level, its description could include attributes that relate to the personal or social significance of that action. If the person receiving the tea is a stranger entering Jack's home, then we can think of the action as offering hospitality. Both the level of intention and the type of intention underlying goal-directed action can determine properties of its execution.

A second sort of modularity concerns a separation between systems involved in apparently different functions, such as between those involved in defence and those involved in homeostatic regulation. Each system might have a number of available courses of action, for instance, 'freezing', fleeing or fighting for defence. The overall choice of action involves arbitration both within and between systems. Later, we consider these separate systems in terms of modularity in 'sub-policies' or policy 'pieces'.

A third sort of modularity, the focus for the latter part of this chapter, divides control into specification, selection, execution and appraisal. Under specification (section 3), we consider affective and cognitive factors governing why certain actions might be preferred. As a way of organizing our analysis, we briefly describe a simple computational framework (reinforcement learning).

Under selection (section 4), we consider how the choice is made between actions of differing preference. Selection is obviously necessary between good and bad actions, but it is also required more generally when actions require simultaneous access to a restricted resource.

Under appraisal, which we consider along with specification, we discuss aspects of the evaluation and monitoring of actions. We can see appraisal as involving the execution of a particular internal action, which we consider as being similar to the execution of an external action. Under some conceptions, appraisal and other forms of explicit deliberation are closely tied to awareness. Execution in general concerns how selected actions are instantiated in the well-coordinated engagement of multiple effectors. Although execution is also critical and interesting, we focus here on the other topics.

Addressing the complexities of action seems to require that we build and analyse controllers in modular terms. However, modularity often leads to sub-optimal control. At the very least, biological control systems offer apparently seamless integration between control at different hierarchical levels and from different systems. This seamlessness and the resulting fluidity of control seem hard to replicate in artificial systems.

Any comprehensive treatment of action faces a formidable challenge. It must take into account – and requires theoretical and formal computational commitments on – the nature of interactions between multiple processes potentially acting on multiple levels of representation. These levels, in turn, could draw upon a range of storage and information processing capabilities. Ultimately, it will require a system-level theory of all these potential component resources and how they interact.

Despite a wealth and variety of experimental studies, we are at the early stages of our understanding. There are many fundamental disputes and gaps. For instance, there is no general agreement on even the crude question of the separate and joint roles of the three mammalian neural structures most closely involved; motor and premotor cortex, the basal ganglia and the cerebellum. Thus, any conclusions are necessarily preliminary.

This review starts with a general overview of the neural and computational framework that is involved in control. We then separately discuss specification, appraisal and selection. We highlight links with ideas from engineering, robotics, computer science, statistics and operations research as they occur. In section 5 we provide pointers to further interactions and open questions.

2 FRAMEWORK

Consider the example of a thirsty rat trained that, when a particular light comes on in a Skinner box, it should press a lever at one location in the box rather than pull a chain at a different location. The result of doing so is that water comes out of a spout at another place in the box. This seemingly simple and routine behaviour raises many issues.

Specification here involves learning that, under appropriate motivational conditions (i.e. thirst), the light requires a response followed by a move to the water spout. The problems of selection include how the animal chooses pressing the lever against other possible behaviours such as grooming, freezing or even attempted escape; and also chooses the light as a focus of attention against innumerable other stimuli in the environment.

The problems of execution include how the rat comes to move smoothly towards and depress the lever, and how it does so in a wealth of different conditions, such as different required forces on the lever, or different slipperiness of the floor. Appraisal and monitoring is required in establishing coherent patterns of behaviour in the first place, and reacting appropriately to things like reversals in the contingencies leading to reward.

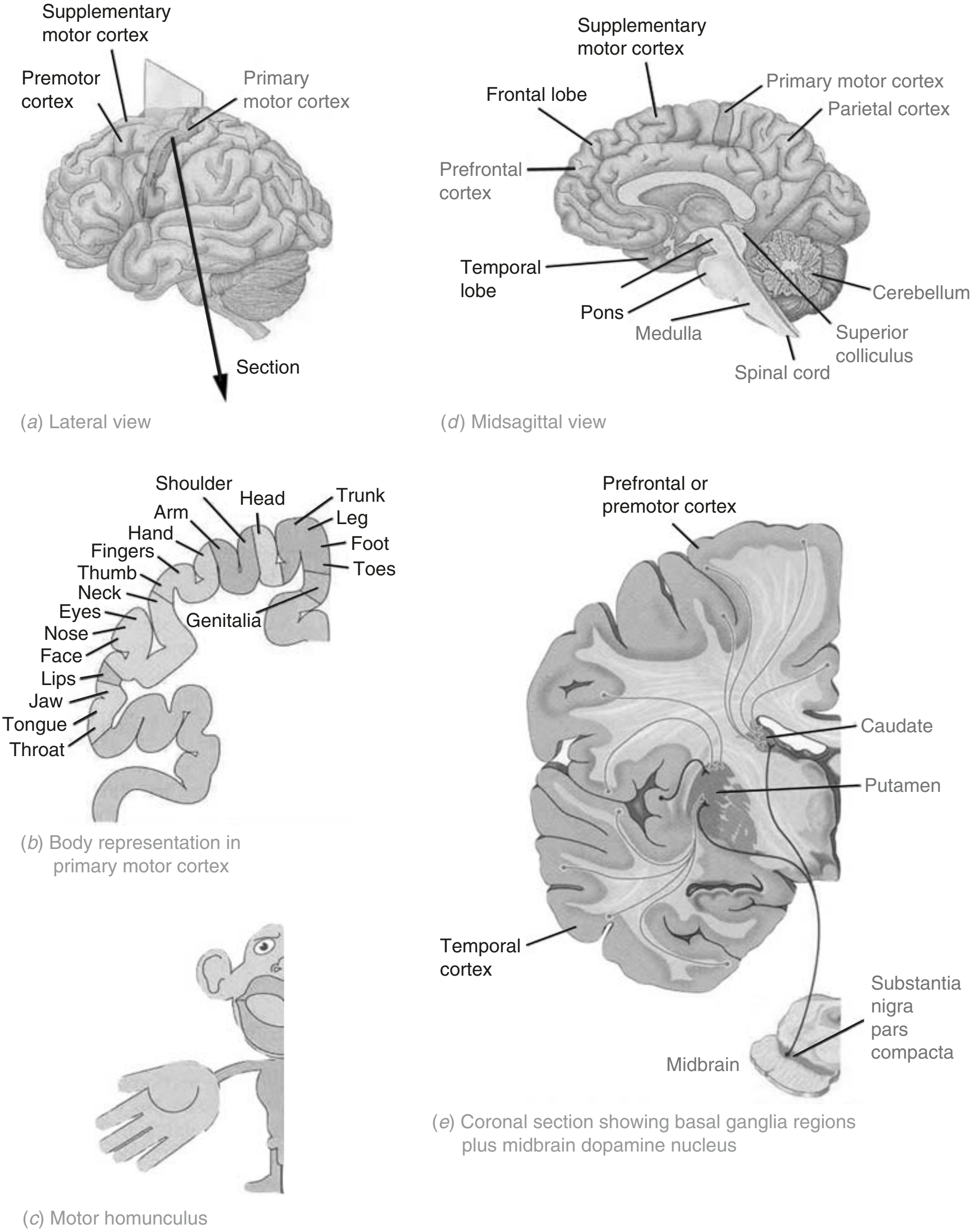
2.1 Implementation

Figure 6.1 shows a picture of the human brain, depicting various gross structures that are involved in action, and various ancillary areas with critical roles. We do not confidently know what these different structures do, and so we offer only the crudest characterization. Functional systems capable of specifying action – sometimes called command systems, and perhaps identifiable with the second sort of modularity described above – are distributed throughout all levels of the neuraxis and communicate directly with cortical and/or hindbrain pre-motor and motor areas.

The brain-stem and spinal cord are the lowest level control structures. They handle a variety of simple reflex behaviours (i.e. hard-wired stimulus-response mappings), many aspects of rhythmic movements such as swimming in fish, and in some sense provide the lowest-level primitives out of which complete actions are built. The midbrain includes hard-wired detection systems that connect predictive sensory events more or less directly to appropriate motor plant. An example of this is loom detectors in the superior collicullus of the frog, which are largely retained in humans.

Coordinated sequences of movements, rather than simple approach or withdrawal, necessitated the evolution of more flexibly organized motor control mechanisms in early versions of the cerebellum, basal ganglia and medulla. Limbic mechanisms, including the amygdala and the nucleus accumbens, allow arbitrary stimuli to predict good and bad outcomes, and thereby specify actions. For instance, the ventral tegmental area and substantia nigra pars compacta are nuclei containing neurons that deliver the neuromodulator dopamine to target structures.

Neuromodulators such as dopamine are not involved in standard, fast neurotransmission, but rather modulate ongoing activity and synaptic plasticity. There is substantial, though controversial, evidence that dopamine plays a key role in learning



(*a*) Lateral view

(*d*) Midsagittal view

(*b*) Body representation in primary motor cortex

(*e*) Coronal section showing basal ganglia regions plus midbrain dopamine nucleus

(*c*) Motor homunculus

FIGURE 6.1 The anatomy of action control. (*a*, *d*) lateral and medial views of the human brain showing cortical and subcortical action control structures. (*b*) Somatotopic organization of motor cortical specialization shown on a section of motor cortex along the precentral gyrus. (*c*) Motor homunculus showing the disproportionate representation of certain efferent structures on motor cortex. (*e*) Coronal section showing basal ganglia structures, including the caudate and putamen and the dopaminergic substantia nigra pars compacta. (Modified from Purves *et al.*, 1997, this figure appears in the colour plate section)

predictions, specifying actions in the light of their rewarding consequences, and also regulating action selection in the basal ganglia.

With its direct connections to the spinal cord, motor cortex can control behaviour relatively directly. It thus allows the full sophistication of cortical processing to be brought to bear on action control. Figure 6.1 shows a crude motor homunculus. It indicates how responsibility for different parts of the body is spatially arranged in motor cortex. The other structures in the figure are also spatially segregated, but seemingly in a more complicated manner. Finally, the ability of the prefrontal cortex to model aspects of the future over extended periods of time, and to specify controls in the light of sophisticated and context-sensitive predictions, makes it the substrate for highly flexible and effective control.

Each set of new structures has evolved, and been refined by evolution, to provide more sophisticated and flexible solutions to the same underlying set of problems about finding adequate food, water, heat, sex and avoiding harm and danger. The structures interact in complex ways, cooperating, competing and offering redundancy. Damage in one structure leads to a complex pattern of deficits arising from the control specified by other structures.

Figure 6.1 also shows the parietal cortex, which includes many areas involved in coordinate transformations. These are necessary when actions are specified in one frame of reference (the lever visible on the retina of the rat), but must be executed in another (the set of angles of the joints of the right foreleg necessary to reach the lever).

Finally, as a counter to cortical chauvinism, note that decorticate rodents, though not ones with compromised basal ganglia, are capable of basic forms of ingestion, grooming, orienting, defence and sexual behaviour, presumably based on control structures in the midbrain and brain-stem.

2.2 Computation

Computational ideas have played a prominent part in understanding optimality in the specification of action. The substantial field of reinforcement learning provides a formal description of the problem of an actor (such as the rat) learning to choose actions (such as pressing the lever) in the light of rewards (such as the water) or punishments (such as an electric shock). Reinforcement learning has rich links to engineering and economics, and notions from game theory of optimizing behaviour, ethological data on the fitness of animal action choice to environments, and to statistical notions of learning models of the world, and learning predictions of future outcomes. Various evidence implicates the basal ganglia and neuromodulatory systems, such as the dopamine system, in the implementation of reinforcement learning.

Reinforcement learning encourages us to think methodically about action specification. In particular, it forces us to consider two key aspects of the task, namely state and reward, and two key aspects of the actor, policy and value function.

The state is a complete description of relevant aspects of the environment and the actor. In general, this is hugely complicated, including aspects of the sensory milieu, internal variables such as hunger or thirst, and even the history of the interaction between the actor and the environment. As mentioned, it should also be hierarchically described. Defining a state in such a comprehensive manner is useful because state then encompasses everything that the actor needs to know to choose its action appropriately. In practice, actors can use highly reduced descriptions or representations of the state. It is a major task to construct representations of state that are computationally convenient for the tasks of action specification, selection and execution. It is a major task for the sensory processing and memory areas of the brain.

States change either on account of the actions of the actor (e.g. the rat pressing the lever), or autonomously as the environment changes (e.g. the light turning on). Many approaches to computational and biological reinforcement learning assume actors to build internal models, sometimes called forward models, of the way that the state changes, since these can be very useful for determining good actions. Cognitive aspects of the specification of action often depend on such models.

The reward quantifies the benefit (or, if negative, cost) of executing an action (such as drinking) at a particular state (which includes information about the state of hydration of the animal). Although a numerical reward value is essential for a clear definition of optimality, there are many disputes in the philosophical and psychological literature as to what constitutes a reward for an animal. In particular, there are challenges in creating a formal definition that is not circular, and can sensibly capture the idea that an animal might emit an action that is suboptimal according to its 'true' reward values. Following some early, and now largely abandoned, ideas about the role of dopamine and norepinephrine, another neuromodulator, various groups are seeking, or debating the existence of, a neural 'currency' of reward. There are also debates about whether and how punishment is quantified.

In many domains, it is also necessary to specify how rewards should accumulate over time, i.e. exactly what is to be optimized if animals perform a whole sequence of actions in a domain and experience a sequence of rewards and punishments associated with their actions. Even in the simple task for the rat, there is an element of this, in that the rat does not get rewarded directly for pressing the lever, only indirectly after performing the further actions required to take it to the water spout and drink. The most popular, though also debated, possibility is to use what amounts to the expected present value of the sum of future rewards, down-weighting or discounting far-off rewards compared to proximal ones. There is substantial evidence as to how humans and other animals discount future rewards. This has a convenient, though not necessarily psychologically accurate, link to economic theory via interest rates.

One internal aspect of the actor from the perspective of reinforcement learning is the policy. Formally, this indicates which action or actions the actor specifies in each state, or the relative preference between possible actions in each state. If, as in our discussion of modularity in the introduction, actions are considered in a hierarchical manner, then so will be policies, with (at least) high level policies (specifying such general objectives as drinking), intermediate level policies (such as approaching and pressing the lever and approaching the water spout), and low level policies governing the patterns of appropriate muscular activity required. In practice, though not in principle, most research in reinforcement learning has focused on intermediate levels, leaving notions of appropriate muscle activations to execution. Research has mostly ignored the specification of appropriate general objectives.

The last important idea from reinforcement learning is that of a value function. This indicates numerically either how good a state is, or perhaps how good it is to perform a particular action at a state. The quality of a state is determined by the sum of future rewards or punishments that are expected to be available from that state, i.e. exactly the value that needs to be optimized. The value function is particularly useful for chains of action extended over time.

Consider the case of the rat when the light is on. Pressing the lever is not immediately rewarding. The rat has to visit the spout to get water. However, the state that pertains after the light is on once the rat has pressed lever has a greater value than that of the state before the lever has been pressed, since it is reliably associated with the reward. The value function obviously depends on the policy, since, for instance, the rat gets no reward if it does not move to the water spout after pressing the lever. There is both neural and psychological evidence that animals learn value functions, predicting future rewards associated with present stimuli.

The engineering field of dynamic programming provides a set of powerful theoretical, conceptual and even practical ways of thinking about optimal action choice or optimal control in reinforcement learning problems described in this way. Some of these dynamic programming methods have close parallels in psychological and neural models of animal action specification. There are also many algorithmic ideas about the best ways to learn how to specify actions in the light of rewards, some of which are based on psychological data from animal conditioning experiments. We discuss various of these below.

Animals constantly balance multiple goals. Consider a highly simplified animal with three mutually exclusive behaviours: feeding, drinking and escape. The optimal policy must depend on state, so that the animal feeds when food is available and the animal is hungry, and drinks (or seeks water) if thirst becomes more pressing. The animal might also switch from feeding to drinking as it becomes increasingly satiated, or when the source of food is exhausted. On the other hand, survival is a priority so we might also expect even a relatively weak threatening stimulus to cause a switch from ingestion to escape. The idea, mentioned in our discussion of modularity, that separate systems are involved in programming these separate behaviours is conceptually attractive and consistent with anatomical and behavioural neuroscience data on the existence of separate command systems distributed throughout the neuraxis.

By contrast, reinforcement learning is usually considered in the narrow context of optimization for a single goal. It would treat behaviours of all these forms as arising uniformly from a single notion of value. One attractive approach to this form of modularity is to consider policies as being 'quilted' from competing (and perhaps cooperating) policies or sub-policies. Preference (and thus selection) is specified both within sub-policies (is pressing the lever or pulling the chain the best action to get the water?) and between sub-policies, as in the competition between ingestion and defence. Although it is clearly undesirable to consider sub-policies as being specified anatomically rather than functionally, we do not yet have a more appealing framework, so we retain this crude characterization. We consider both aspects of selection in section 4.

Computational ideas have also been important in action execution. For instance, the idea that muscle control satisfies optimality conditions has provided a powerful organizing force for a wealth of confusing literature. At the lowest levels of control, there are the problems of redundancy and noise. For the rat to touch the lever, its appropriate paw must be in the right place. However, its leg could take many paths to get the paw there. Those paths could be traversed faster or slower. Additional constraints are therefore required to pin down its actions fully. Possible constraints include minimizing movement time, or the amount of effort, or the possibility of ending at the wrong location in the face of the substantial noise in the control system.

These constraints are effectively different optimality conditions and are part of the computational characterization of the problem of execution. It is not clear how the task of satisfying these optimality conditions is allocated amongst the various neural structures described above, although disruption in smooth and accurate behaviour is consequent on various sorts of insult.

3 ACTION SPECIFICATION

Crudely speaking, there appear to be at least three main routes to action specification: innate reflexes or tropisms; learned (i.e. conditioned) Pavlovian and operant responses; and habits. In our modular scheme, these routes report preferences amongst actions; selection mechanisms arbitrate the resulting choices, potentially both between different routes and within at least some of the routes. The extent to which these psychologically defined routes to action specification are equivalent to anatomically defined command systems is unclear. In discussing action selection, we will assume that they indeed are.

Particular classes of appetitive stimuli, such as food and water, and aversive stimuli, such as shocks or threats, are called unconditioned stimuli in the behavioural literature. They are associated with reflex actions. For instance, animals have freeze, fight, flight actions available to them in response to aversive unconditioned stimuli, and presumably the capacity to select between these according to circumstance. Such reflexes are highly sensitive to internal, motivational states. For instance, a sated animal will not eat more of the food on which it has recently gorged, even when it is freely available.

Reflexes are a form of intrinsic policy or sub-policy, which is presumably arranged by evolution to try to maximize benefit, here characterized in terms of the rewards. Although reflexes may indeed be optimal in simple circumstances, their rigidity results in many anomalies and sub-optimalities of action specification.

The second route to action specification is called classical or Pavlovian conditioning. It consists of extending these innate reflexes to stimuli, called conditioned stimuli, such as lights, tones or flavours, by virtue of their predicting unconditioned stimuli. This results in conditioned responses, which are often closely related to the responses elicited by the unconditioned stimuli. It can also initiate reflexes such as approach and engagement, or withdrawal, associated with acquiring or avoiding those unconditioned stimuli.

Pavlovian responses are partly regulated by the same motivational factors as the unconditioned stimuli that are predicted. So sating an animal on a food will usually suppress approach to a light that predicts the delivery of food. However, something akin to a stimulus-independent reward value is also important, since, for instance, if a light predicts a tone that predicts food, then the animal might approach the light even if sated such that it would not approach the tone or eat the food. In fact, it has been argued on behavioural grounds (see, for example, Dickinson and Balleine, 2002) that there are two stimulus-independent systems by means of which value functions are specified, one reporting appetitive value, the other, aversive value. Such systems are candidates for representing the value functions we discussed above. A variety of neural studies implicates the activity of dopamine cells as reporting a form of prediction error (called a temporal difference prediction error) in the value function.

Even if Pavlovian conditioning nicely concerns predictions of future reward, the actions that it specifies, like other reflexes, can be suboptimal. This is clearest in omission schedules. For instance, pigeons will approach and peck a key whose illumination generally predicts the delivery of food, even if, on any trial on which they actually peck the key, the food is withheld. The consequence of the lighted key predicting the food is approach and engagement (i.e. pecking) even though this has a catastrophic effect on the delivery of reward.

The third route to action specification concerns the way that actions can be chosen explicitly to achieve goals and maximize, or approximately maximize, accumulated rewards. This is the realm of instrumental conditioning, and is the subject of the bulk of computational investigations. There is a wide variety of possible mechanisms for instantiating this sort of control, many of which are as fiercely debated in the computational literature as in psychology and neurobiology.

One basic division is between direct and indirect methods. Indirect methods specify policies at least partly through an explicit model of the environment, working out which action is best either by forward chaining (following through the consequences of an action and, recursively, those of its successors) or by backward chaining (working out which action achieves a goal; or, if none do so, from which sub-goal the goal can be reached, and, recursively, which action achieves that sub-goal). Direct methods are less flexible, though simpler. They involve learning policies that map the state to a suggested course of action without on-going reference to a model.

We can see the difference in the case of the rat with the option of pressing the lever. An indirect method would consult a model of the effect of pressing the lever (i.e. the availability of the water at the water spout), and would consider the utility of this action against other possible actions whose consequences it has also modelled, in light of its degree of thirst, etc. By contrast, a direct method specifies that pressing the lever may be appropriate, without reference to the actual consequence. Both methods depend on learning – indeed, one popular direct method uses the same temporal difference prediction error signal that we suggested was reported by the activity of dopamine cells – but acquire and use knowledge in different ways. Note that no equation should be made between rational behaviour – in terms, for instance, of maximizing rewards – and indirect control. A learned direct controller can be just as rational in the domain for which it is trained.

When specifically addressing selection mechanisms in humans, the distinction between direct and indirect means of control can be further elaborated. Key open issues in cognitive and computational neuroscience concern the mental and neural organization of the component systems and processes that contribute to cognitive control:

- how we might best construe representations that are utilized within those component systems
- how action selection is influenced or determined by attentional mechanisms working upon rich and multi-modal source of information about current external states of the environment and internal states of the body
- how information in memory interacts with those current states
- what kind of role might be served in our understanding of cognitive control by constructs such as phenomenology, awareness and intentions, and how we might best approach the problem of computationally realizing the generic mechanisms that underpin interactions among component systems.

One simple contrast is the long-standing distinction made in cognitive psychology between 'automatic' and 'controlled' processing (Schneider and Shiffrin, 1977). This essentially maps onto the contrast between direct and indirect control. The contrast implies that transitions among action steps differ in terms of when and how choices are made. As a result of learning history, successive steps in a sequence of actions or thoughts can undergo integration into higher level functional units. In automated control, one step is regarded as following on from another with 'minimal' executive processing at transitions within the unit.

Since the precise state of the environment and organism differ from one performance of an action to another, transitions must tolerate variation in states of the environment or organism. At the very least, there must be basic control mechanisms to detect when a current state significantly conflicts with a state compatible with the continued course of that action such that it warrants intervention of controlled processing. Determining what is significant depends on current states, including emotional states such as anxiety, or those preserved in memory. The term 'controlled' processing simply implies that something extra and qualitatively different happens at a transition between action steps than occurs in 'automated' control. At such transitions the immediate value of options, goals and plans can be subject to processes of evaluation, appraisal, conflict resolution, or systematic reformulation.

The second contrast concerns the actual contents of the states on which choice is based. Cognitive control can be thought of not only as involving generic processes, like planning, monitoring, appraisal, or conflict resolution. It can be distinguished by what the state itself actually represents or signifies. Either or both of these contrasts may be called into play when control of action is labelled as 'cognitive', and the same contrasts are often invoked when describing the control of actions involving movement and those confined to covert thought.

The interdependence of action control on multiple systems is clear. Humans routinely make simple action slips of the type that occur when we accidentally place tea in a pot when we intended to make coffee. In abstract terms, coffee and tea have similar characteristics. When running through an automatic sequence, the discrepancy can easily fail to trigger an intervention of controlled processing. Unsurprisingly, the precise distinctions that might best differentiate types or level of representation – how they are stored and definitions of generic processes required for the cognitive control of action – are, once again, hotly debated.

Treatments of cognitive control of action are rendered yet more intricate because they intersect with the role of other resources, such as transient emotional states, short-term or working memory, as well as long-term memory. Long-term memory is a vast resource that represents, or models, regularities in the co-occurrence of elements of information. So, for example, highly anxious individuals are more likely than less anxious individuals to attend selectively to, and act upon, threatening information.

Patients with certain forms of amnesia may lack representations of their past experience to support the cognitive control of action in their present context. Memory also preserves representations that abstract, or model, just those properties that related events or conceptual structures have in common. Abstraction into higher order representation is a key issue that links up with the intentional or willed control of action sequences involving movement.

Patients with conditions such as Parkinsonism, ideomotor apraxia, schizophrenia and frontal syndrome all exhibit selective deficits in their cognitive control of action sequences. Although they may well be capable of executing detailed sequences, patients with Parkinsonism may have problems getting out of a chair under intentional control but may be quite capable of doing so in response to significant environmental stimuli such as when someone shouts 'Fire!'. Likewise, individuals with ideomotor apraxia may be capable of hammering a nail in the meaningful context of the action sequence of hanging up a picture at home. When explicitly asked to hit a nail with a hammer in other contexts, such as a clinical test, they appear to know what to do but cannot do it in response to a de-contextualized instruction.

Patients with schizophrenia exhibit poverty of action, perseveration, and carry out actions that are inappropriate to the context. Patients with substantial damage to their frontal lobes, while exhibiting a range of specific deficits, including perseveration and capture errors, none the less initiate a range of responses or actions that can be legitimately labelled as under the control of cognitive systems or processes. They also exhibit deficits in planning, particularly on rather complex tasks with an element of novelty, such as how best to organize a trip around an unfamiliar shopping centre. Such deficits are often linked to problems in forming or using highly abstract 'meta-representations'.

Concerns with phenomenology, awareness, volition and intention naturally enter into many treatments of attention, working memory and the cognitive control of action. To fail to mention them would be a significant omission, whilst attempting to summarize the more philosophical points succinctly would be a mammoth task. We shall therefore, conclude our discussion of cognitive control with a few points.

Awareness is often loosely linked to the operation of the mechanisms of attention. The attribute of being 'focally aware of' has been linked to just those objects, sounds, bodily sensations that have been selected by attentional mechanisms from the wider external scenery or bodily state of the organism. The current contents of working memory have also been associated with focal awareness.

As with our earlier example of construing the goals of an action of different levels, our awareness of actions, and their underlying intentions, relates more broadly to richer forms of representations of the environment and organism. A good example of this comes from blindsight. Patients with this condition are not aware of the presence of entities in their blind field. Even if they were thirsty and a glass of water was in that field, they would not spontaneously drink it. None the less, there is good evidence that when such patients are encouraged to guess where an entity is, and to act in relation to it, they can perform a well-formed motor sequence that is matched to the physical properties of the object and its spatiality. This implies that information of states of both the environment and organism is available in some non-conscious form at least to lower-level action specification structures.

4 ACTION SELECTION

Animals have multiple output channels (hands, feet, teeth, etc.) that can act on the environment and on internal processes. In principle, each effector could be controlled independently. It is typically not the case that when a competition for use of one set of muscles is resolved this automatically denies access to all other muscle groups. Thus, with a few notable exceptions, most of us can actually walk and chew gum! The need to choose between actions comes from competing demands for the same channels.

The discussion of action specification has thus left us with two critical selection problems, within and between what we have called sub-policies or command systems. Crudely, within a sub-policy, competition comes between actions with the same or similar goals. Between sub-policies or command systems, competition comes between competing goals, possibly specified at different levels of a hierarchy. It may be that one structure in the brain, such as the basal ganglia, arbitrates between sub-policies (e.g. selecting defensive behaviour) and, for some sub-policies arbitrates between actions within a sub-policy (e.g. pressing the lever rather than pulling the chain). Other aspects of action choice may also emerge from the selection process, such as the vigour or enthusiasm with which they are executed. The combined competition ultimately leads to the specification of a single policy.

The raw material of selection is preference (i.e. specification). For conditioned operant responses, we discussed how preferences emerge through a desire to maximize summed future rewards. Indeed, there are suggestions that we can measure relative preferences directly as firing rates in some action-influencing areas of cortex (Glimcher, 2003). For other sub-policies, such as reflex control, for the choice between sub-policies (such as defence versus ingestion), and specially for cognitive control, we know much less about the provenance or expression of preferences. It is attractive to suppose that there is a common currency by which qualitatively different policies can be evaluated, perhaps one associated with appetitive and aversive motivational systems. However, little evidence directly supports this.

Remember that selection is formally necessary only when there is competition for some restricted resource. However, it may take place at a more basic level, so ignoring, for instance, the possibility of choosing between defensive actions in the additional light of satisfying ingestive goals. Remember also the hierarchical nature of policies and sub-policies. Since many aspects of the problems of selection are the same at all levels of a hierarchy, it is attractive to believe that copies of a standard selection circuit could be used, just with different inputs and outputs for different levels of control.

Desirable characteristics for effective selection include:

- Clean-switching: a competitor with a slight edge over the rest should see the competition resolved rapidly and decisively in its favour.
- Avoidance of distortion: the presence of activated but not selected competitors should not interfere with expression of the selected system's (winner's) outcome.
- Avoidance of dithering: following a selection, the execution of a winning outcome often changes the state of the agent so that it is no longer so desirable (e.g. drinking reduces thirst which reduces the desirability of drinking).

This can allow a close competitor, such as eating, to dominate. The same then happens with the second selection, causing a switch back to the first competitor, and so on – this oscillation is dithering. Clean-switching and avoidance of distortion can be provided by circuits that implement 'winner-take-all' functionality. Dithering can be avoided by endowing the switching mechanism with some form of persistence or hysteresis, such as the incorporation of a non-linear positive feedback loop into the switching circuit to maintain, or enhance support for the winner.

4.1 Selection architectures

A variety of architectures has been proposed to deal with the selection problem in both artificial and biological systems. We will now describe and consider some of main solutions as possible templates for interpreting patterns of connectivity that could implement selection within the vertebrate's brain (Fig. 6.1).

Brooks developed the 'subsumption architecture' (Fig. 6.2*a*) as a robust architecture to control the behaviour of autonomous mobile robots (Brooks and Stein, 1994). It consists of an hierarchically organized set of layers. Each layer has a specialized sensory input. If activated, the input automatically links to motor output that generates a specific behaviour, and is thus like a command system or sub-policy. Higher layers implicitly rely on the appropriate operation of those below. Conflicts between layers are handled according to a fixed priority scheme.

Layered architectures of this type allow rapid responses to environmental contingencies. They can provide appropriate action selection for robots with a limited number of behavioural goals. However, since prioritization is 'built-in', the system is inflexible. It becomes difficult to determine appropriate dominance relationships as control systems become more complex. It has no means for instantiating the subtleties of action specification discussed above.

In distributed selection architectures, all competitors are reciprocally connected (Fig. 6.2*b*). Each has an inhibitory link to every other (recurrent reciprocal inhibition) and an excitatory link to the shared output resource for which they are competing. Such networks display a form of positive feedback, since increased activity in one competitor causes increased inhibition on all others, thereby reducing their inhibitory effect on the first. Recurrent reciprocal inhibition can therefore support winner-take-all functionality, making it an attractive means for implementing action selection.

The relative strengths of incoming excitatory links, and of the inhibitory links between

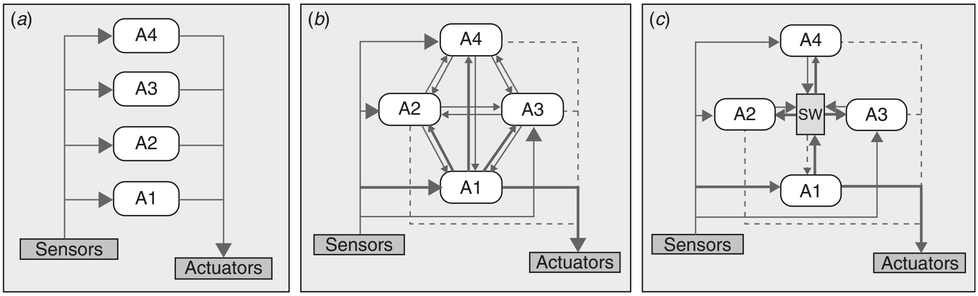


FIGURE 6.2 Connectional architectures capable of selecting between competing command systems (A1–A4). (*a*) Subsumption architecture used by Brooks (Brooks and Stein, 1994) to control behavioural output of autonomous mobile robots. (*b*) Distributed selection mediated by reciprocal inhibitory connections between competitors. (*c*) Central section in which individual command systems output is under control of a central selector or switching mechanism (SW). Relative preference of inputs from each command system to the central selector determines output control. Thickness of the lines in (*b*) and (*c*) represents the strengths (preferences) of the signals. (This figure appears in the colour plate section)

competitors, can be tuned to support the sort of complex pattern of dominance relationships required by the action specification we discussed earlier. Reciprocally inhibiting networks are widespread in the central nervous system. However, connection costs are likely to preclude it from being the direct arbiter of selection between potentially competing functional units distributed widely throughout the brain.

Finally, there may be a 'centralized selection mechanism' involved in closed loop connections with all competing systems (Fig. 6.2*c*). Inputs associated with each action are excitatory and convey a measure of preference (as in action specification within a sub-policy) or urgency to the central selector. The return connection from the selector is tonically active and inhibitory. Depending on the comparative preferences, tonic inhibitory output is withdrawn exclusively from the action that is most preferred.

There are several reasons why both biological and artificial control systems may benefit by exploiting centralized selection for overall behavioural control. First, centralized selection requires only two connections for each competitor (to and from the selection mechanism) resulting in a total of 2*n* connections. This is a considerable saving over the $n(n - 1)$ connections required by a fully connected distributed architecture. Moreover, adding a new competitor to the central selector entails adding only two further connections compared to the 2*n* connections required for reciprocal inhibition between all competitors. Secondly, insofar as selection is separated from perceptual and motor control problems, independent adaptations and modifications to selection can be made with less serious consequences for other components.

In summary, while selection probably occurs throughout the brain – much of it distributed with emergent properties, where there is a specific need to arbitrate between spatially distributed functional units – it is clear that a central selection device could play an important role. In this regard, it is probably significant that the central selection system illustrated in Fig. 6.2*c* can be viewed as a cartoon template of the architecture of the basal ganglia.

4.2 The Vertebrate Solution?

A recurring theme throughout an extensive literature is that the basal ganglia, with their particular channelled architecture (Alexander *et al.*, 1986; Middleton and Strick, 2000), are involved in selection. It has thus recently been proposed that the basal ganglia act as a central selection mechanism in the vertebrate brain (Figs 6.2*c* and 6.3) (Redgrave *et al.*, 1999). Distributed selection architectures (Fig. 6.2*b*) may operate within basal ganglia circuitry (see below), where their useful switching properties may be exploited, while minimizing the undesirable overheads incurred by fully connected reciprocal inhibition.

Competition happens between different sub-policies (command systems), and between different actions within each sub-policy. It has been suggested that the basal ganglia are divided into limbic, associative and

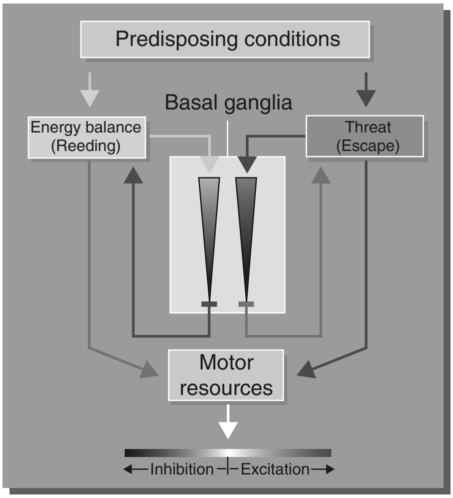


FIGURE 6.3 A conceptual model of selection by the basal ganglia arbitrating between two command systems competing for access to a shared motor resource. In this model, grades of red and blue represent the relative strengths of excitatory and inhibitory connections respectively. (Dark colours = greater preference, this figure appears in the colour plate section)

motor domains, each with ordered influences on the next. One speculative, though attractive, idea is that this provides a substrate for hierarchical aspects of selection, with the limbic domain resolving overall goals, the associative domain resolving appropriate actions and the motor domain resolving specific patterns of muscular activity.

Connections to and from the basal ganglia are arranged in a particular manner. Inputs, from throughout the cortex and limbic structures, are excitatory and physically active. Return connections to command systems are mostly inhibitory and tonically active (Fig. 6.3). The suggestion is that relative preferences of competing inputs to the basal ganglia are directed by intrinsic processing (Fig. 6.4) to inhibit the inhibitory output of the preferred channel (Fig. 6.3), while at the same time maintaining or increasing the tonic inhibitory control of others. In this manner the action, or even the sub-policy with the most salient input, is selected by 'disinhibition' from the basal ganglia (Chevalier and Deniau, 1990), a process that permits its sensory/cognitive input to connect with its efferent projections to the motor plant. As mentioned above, it is conceivable, though not certain, that some sub-policies (e.g. some defensive systems) have their own competitive mechanisms (between freezing, fleeing and fighting) to decide between possible actions. Others might rely directly on the basal ganglia itself.

For the basal ganglia to be involved in selecting between actions within the sub-policies of reflex and direct and indirect control discussed above, it needs the sort of rich information about the internal and external state of the actor, and predictions of the affective consequences of stimuli and actions, that are the mark of these forms of control. Indeed, the striatum receives a variety of contextual information, and has a critical and extensive innervation from neuromodulatory systems such as the dopamine system.

In mammals, there is a large projection to the basal ganglia from cortical areas subserving primarily cognitive functions. This indicates that the mammalian basal ganglia may not be confined to the selection of behaviours, actions and movements, but rather could play a comparable selective role in cognition, i.e. arbitrating between multiple cortical systems competing for a share of limited memorial or attentional processing.

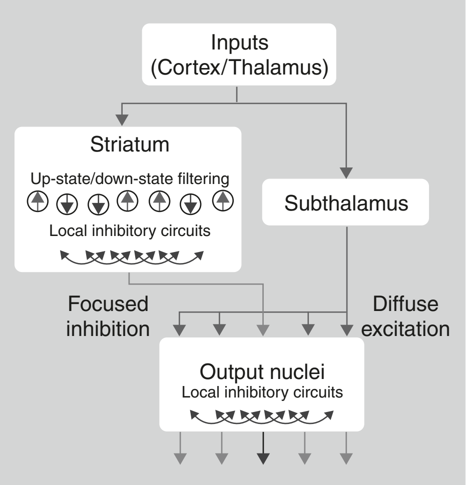


FIGURE 6.4 Intrinsic processing within the basal ganglia can be viewed as a sequence of selective processes that confer winner-takes-all functionality (Wickens, 1993). The processes include:

- The bi-stable 'up–down' state of the membrane of striatal medium spiny neurons (Wilson, 1995) may operate as a first pass input filter in which weakly supported 'bids' for selection get no further. Reciprocal inhibition between striatal medium spiny neurons (Tunstall *et al.*, 2002) and the wider projecting but highly influential GABAergic interneurons (Koos and Tepper, 1999) may refine selections based on local and longer-range reciprocal inhibition respectively.
- The superimposition of focused inhibitory input from the striatum onto diffuse excitatory input from the subthalamus at the level of the output nuclei will produce an inhibitory 'off' centre encircled by an excitatory 'on' surround at the level of the basal ganglia output nuclei. (Mink, 1996; Gillies and Willshaw, 1998)
- Output selections may be sharpened by powerful reciprocal inhibitory connections between output cells in substantia nigra and the entopeduncular nucleus/internal globus pallidus.(Mailly *et al.*, 2001)

(This figure appears in the colour plate section)

5 LINKS AND OPEN QUESTIONS

In action and planning, there has already been substantial interaction between computational and engineering methods and ideas and ethological, psychological and neural findings. There is a ready flow of ideas across the apparent chasm. For instance, theories about the course of operant conditioning from mathematical psychology lay at the heart of the engineering field of stochastic learning automata.

Temporal difference learning has psychology and engineering as joint parents. It was then shown to offer an explanation for an otherwise puzzling set of neurophysiological results on the activity of dopamine cells during conditioning. It has also proved to have a sound underpinning in dynamic programming, and, at least according to some sources, to support much work in planning in artificial intelligence.

Equally, cognitive control methods were strongly influenced by ideas from Old-Fashioned Artificial Intelligence, such as production systems and semantic networks. More recently, cognitive control methods have been caught up in the more general debate between connectionist and symbolic architectures for psychological modelling. This has led to a range of interesting, though computationally limited, suggestions.

A direct test of the idea of centralized selection and the basal ganglia has been made through implementing a high-level computational model of intrinsic basal ganglia circuitry and its interactions with simulated thalamocortical connections (Gurney *et al.*, 2001*a*,*b*). The computational model was then exposed to the rigours of 'real world' action selection by embedding it within the control architecture of a small mobile robot (Prescott *et al.*, 2002). These investigations established that a computational model based on basal ganglia architecture can effectively switch between competing channels depending on the dynamics of relative input preference, and in the robot, can effect appropriate and clean switching between different actions generating coherent sequences of behaviour. These studies confirmed that the basic template connective architecture of basal ganglia can achieve selection in the artificial context of computer simulation and robot control.

5.1 Open Questions

One important open question concerns the precise roles played by the three major neural control structures we have discussed, the cortex, the basal ganglia and the cerebellum. We have emphasized various aspects of their roles, notably specification for the cortex, selection for the basal ganglia and (though in less detail here) smooth execution for the cerebellum. However, these functions, and therefore the required behaviours of the structures, overlap. Even functions such as timing are likely to be multiply represented. There are also competing conceptions, for instance Doya's (1999) suggestion that they subserve different aspects of learning, with the basal ganglia performing reinforcement learning, the cerebellum supervised learning and the cortex unsupervised learning.

A second open question afflicts biology and engineering alike. This is the adaptive specification of hierarchical control structures. Such things have been considered in the cognitive terms of chunking and automatization, and, in computer science, in terms of options, macros and subroutines. However, there is currently a dearth of good ideas for how they can be automatically induced.

One major difference between biological and engineering control in this arena is that computers are not limited to storing only a few entities in active, working memory. So, for instance, computers can explicitly use a forward model to consider the consequences of many more actions. However, a powerful curse of dimensionality plagues this sort of explicit use of a model. Hence the same concerns apply for only slightly larger systems. Both disciplines are also exploring methods for handling the consequences of error or mis-prediction in action outcome. Neither yet seems to have an appealingly formal treatment.

Related to this, we have seen that a bewildering array of structures and processes are involved in specifying, selecting and executing actions. The key open issues concern not whether control mechanisms are centralized or not, but the details of how they are distributed among the component resources and processes of a wider system, and how these are best defined and bounded.

As we noted in the opening section, a major challenge for research requires theoretical and formal computational commitments on the nature of interactions between multiple processes potentially acting on multiple levels of representation which, in turn, potentially draw upon a range of storage capabilities. This will require a system-level theory of these potential component resources and how they interact.

Formal logics and process algebras for describing and instantiating the interaction of multiple, concurrently operating systems have been developed within computer science and engineering, but research has only just begun to address their utility for modelling neural and mental architectures.

A third open question, which is more particular to biological action, concerns the relationship between reflexes and the direct or indirect control of actions, and the way in which a whole policy is quilted from sub-policies. The advantage of having a set of innate, usually useful, actions in response to stimuli is obvious, although it is regrettably hard to recapitulate for mobile robots the millions of years of evolution that led to their specification. However, how direct and indirect methods for controlling actions interact, how this ultimately leads to the birth of new habits, and how motivational factors influence both, is under explored.

In sum, the study of many aspects of the control of action already involves strong links between biological and physical sciences. As better understanding develops of the intrinsic characteristics and limitations of animal and engineering control systems, the interplay can only become richer.

References and Further Reading

Alexander, G.E., DeLong, M.R. and Strick, P.L. (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Ann. Rev. Neurosci.*, 9: 357–381.

Baddeley, A.D. (2000) Working memory. In A.E. Kazdin (ed.), *Encyclopedia of Psychology*. Washington, DC: American Psychological Association.

Brooks, R.A. and Stein, L.A. (1994) Building brains for bodies. *Autonomous Robots*, 1: 7–25.

Chevalier, G. and Deniau, J.M. (1990) Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci.*, 13: 277–280.

Dayan, P. and Balleine, B.W. (2002) Reward, motivation and reinforcement learning. *Neuron*, 36: 28S–298.

Dickinson, A. and Balleine, B.W. (2002) The role of learning in motivation. In C.R. Gallistel (ed.), *Learning, Motivation and Emotion*, Volume 3 of *Steven's Handbook of Experimental Psychology*, 3rd edn. New York: Wiley.

Doya, K. (1999) What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural Netw.*, 12: 961–974.

Gillies, A.J. and Willshaw, D.J. (1998) A massively connected subthalamic nucleus leads to the generation of widespread pulses. *Proc. R. Soc. Lond. B Biol. Sci.*, 265: 2101–2109.

Glimcher, P.W. (2003) *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics.* Cambridge, MA: MIT Press.

Gurney, K., Prescott, T.J. and Redgrave, P. (2001a) A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern.*, 84: 401–410.

Gurney, K., Prescott, T.J. and Redgrave, P. (2001b) A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biol. Cybern.*, 84: 411–423.

Head, H. (1920) *Studies in Neurology.* London: Henry Frowde/Hodder & Stoughton.

Houk, J.C., Davis, J.L. and Beiser, D.G. (eds) (1995) *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press.

Koos, T. and Tepper, J.M. (1999) Inhibitory control of neostriatal projection neurons by GABAergic interneurons. *Nature Neurosci.*, 2: 467–472.

Mailly, P., Charpier, S., Mahon, S., Menetrey, A., Thierry, A.M., Glowinski, J. and Deniau, J.M. (2001) Dendritic arborizations of the rat substantia nigra pars reticulata neurons: spatial organization and relation to the lamellar compartmentation of striato-nigral projections. *J. Neurosci.*, 21: 6874–6888.

Middleton, F.A. and Strick, P.L. (2000) Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Res. Rev.*, 31: 236–250.

Mink, J.W. (1996) The basal ganglia: focused selection and inhibition of competing motor programs. *Progr. Neurobiol.*, 50: 381–425.

Prescott, T.J., Gurney, K., Montes-Gonzalez, F., Humphries, M. and Redgrave, P. (2002) The robot basal ganglia: action selection by an embedded model of the basal ganglia. In R. Faulls (ed.), *Basal Ganglia VII*. New York, NY: Plenum Press.

Purves, D., Augustine, G.J., Fitzpatrick, D., Katz, L.C., LaMantia, A. and McNamara, J.O. (1997) *Neuroscience*. Sunderland, MA: Sinauer Associates, Inc.

Redgrave, P., Prescott, T. and Gurney, K. (1999) The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89: 1009–1023.

Schneider, W. and Shiffrin, R.M. (1977) Controlled and automatic human information processing: I. Detection, search and attention. *Psychol. Rev.*, 84: 1–66.

Schultz, W. (1998) Predictive reward signal of dopamine neurons. *J. Neurophysiol.*, 80: 1–27.

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning*. Cambridge, MA: MIT Press.

Tunstall, M.J., Oorschot, D.E., Kean, A. and Wickens, J.R. (2002) Inhibitory interactions between spiny projection neurons in the rat striatum. *J. Neurophysiol.*, 88: 1263–1269.

Wickens, J. (1993) *A Theory of the Striatum.* New York: Pergamon Press.

Wilson, C.J. (1995) The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In: J.C. Houk, J.L. Davis, D.G. Beiser (eds) *Models of Information Processing in the Basal Ganglia*, pp. 29–50.

CHAPTER

7

Social Cognition

Uta Frith and Sarah-Jayne Blakemore

1 INTRODUCTION

Human beings are social creatures. We crave social communication and suffer profoundly if temporarily isolated from society. So much so that punishment in most cultures involves some kind of isolation from others.

Much of the brain must have evolved to deal with social interactions. We share many social competencies with other animal species: choosing a mate, competing with rivals, nurturing babies, making alliances and so on. Some aspects of social communication are thought to be unique to humans, for instance the desire to teach, the development of self-awareness and awareness of others, and the ability to outwit others. Nevertheless, a few non-human species

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

share some of these advanced social abilities in rudimentary form.

In this report, we first try to identify the scope of social cognition and offer a definition. In the section 7.2 we describe recent research on the mechanisms of social cognition and its component processes in the brain. In the section 7.3 we discuss how pathology affects social cognition. Our fourth section poses some burning questions from interactions in everyday life, which we hope can be solved in the next five to ten years.

1.1 What do We Mean by Social Cognition?

Social cognition means different things to different people. Most generally, social cognition is defined as any cognitive process that involves other people. These processes can be involved in social interactions at a group level or on a one-to-one basis. When we use the term *cognition* we refer to unconscious mechanisms in the mind (the brain) that bring about representations (a neural implementation of experience). We can be consciously aware of these representations but mostly we are unaware of them. We know for instance that our own perspective and the perspective of another person on the same event can be quite different. However, when we act in everyday life, we often have to judge other people's perspectives implicitly, which occasionally leads to misinterpretation of others' actions as insults if we are not made aware of the different viewpoint.

Within social psychology, the traditional understanding of social cognition is taken to mean the study of social knowledge, social structure, group behaviour, social influence, processing biases, whether and how social category (sex, age, race) defines people, stereotyping, memory for social information, and attribution of motives. This work has produced a solid body of knowledge and has contributed to a better understanding of prejudice, peer pressure, group behaviour and bullying.

Within evolutionary biology, social cognition includes processes such as learning and memory in a social context, with respect, for example, to territoriality in animals, dominance and subordination within the social structure and the complexities of living in a group leading to social pressures and stress. Work with social animals such as non-human primates, mice, rats and birds has lead to important advances. Birds for instance have been shown to be capable of tactical deception (Emery and Clayton, 2001). Researchers have already started to sequence genes in social insects (Bourke, 2002).

Within developmental psychology, it is often assumed that the factors governing cognitive performance in terms of interactions with others are a product of individual cognitive abilities and social competence. This is exemplified by Piaget's work on moral development (Piaget, 1972), where he proposed that social agreement was needed for a true understanding of wrong-doing and its punishment. It is also exemplified by Vygotsky's work on learning in a social context (Vygotsky and Vygotsky, 1980), where negotiating with peers helps problem-solving. The study of the development of infants has recently received a great boost through new behavioural techniques. As we shall discuss throughout section 7.2, this work has revealed very early sensitivity to other people.

In the clinical or psychopathological context social impairments are common and contribute a great deal to the burden of mental illness or disability. Autism is one developmental disorder that is defined by social and communication impairment. Here a deficit in one aspect of social cognition, an intuitive ability to attribute thoughts and feelings to others ('theory of mind'), has been demonstrated (Baron-Cohen *et al.*, 1985). In certain types of schizophrenia too such a deficit has been pinpointed (Bentall *et al.*, 2001). Psychopathy has been recently interpreted as a deficit in another aspect of social cognition, a failure in intuitive empathy (Blair *et al.*, 1996). Researchers are currently investigating the brain basis of these cognitive deficits.

It is clear that the field of social cognition represents a huge diversity of interests. We require a broad notion of cognition, incorporating emotional processes, for instance those that underlie empathy. Given this diversity, what we mean by social cognition may be in danger of encompassing everything the mind (brain) does! Even though we believe that social influences are pervasive, a wide definition is not useful. We clearly have to set boundaries on social cognition for this report. We will discuss only those processes of social interaction and communication that are required when talking about the effect of one person on another. These processes must not be so vague as to be impossible to explain by computational and/or neural mechanisms.

It is only relatively recently that the search for the biological basis of social cognition has started, from genes to brain processes. We still do not know just how biological factors interact with environmental variables to produce individual differences and pathology. Clearly, the study of such processes needs to be influenced, if not carried out, by scientists from a variety of disciplines.

How special is social cognition as compared with other cognitive processes? It could be that social cognition is simply a very complex example of how cognitive functions have to be organized to deal with complex processing demands. However, the idea that there are specific social processes is attractive. Or can we explain the more complex phenomena of social cognition by basic cognitive processes, such as visual perception, memory and attention? Is face processing, for instance, any different from the perception of other complex stimuli with emotional overtones? Although general cognitive processes such as visual perception, memory and attention are vital to social competence, in this report we focus mainly on processes that appear to be special to social interaction.

1.2 What are the 'Building Blocks' of Social Cognition?

Are there different ways of processing objects in the world according to their social or non-social nature? This question highlights the importance of identifying simple underlying functions that contribute to social cognition. Given the appropriate neural model, these simple functions can then be mapped to neural substrates. Congenital abnormalities in these substrates can serve to identify endophenotypes that relate to disorders of social cognition. The term endophenotype refers to what is the 'inside phenotype' rather than overt behaviours, which are likely to be the product of different endophenotypes. These then can lead the search for the genetic basis of specific social functions. One example is the recently identified allele of the serotonin transporter gene that affects the amygdala's responses, and thus contributes to individual variation in anxiety (Hariri *et al.*, 2002).

Cognitive processes are invisible. However, we can measure their effects in behaviour. And to some extent their biological basis is visible through well-controlled brain imaging. It may seem over-ambitious to work out how connections can be made between highly sophisticated social behaviour and fundamental neurophysiological mechanisms. However, there are already examples. Not so long ago it would have seemed foolish to look for the neurophysiological basis of social acts such as deception and double bluff. Now it is mainstream cognitive neuroscience. We anticipate that the genetic basis for different aspects of social cognition will be illuminated. This is feasible by studying individuals born without the ability to develop normal social communication (e.g. autism).

1.3 What can be Measured?

The study of social cognition uses the same measures as any other area of cognitive science. Some measures are especially helpful in the study of communication and emotion. These include: autonomic responses, e.g. heart rate, pupil dilation, galvanic skin response; brain activity, neuronal activity, scalp electrical impulses, cerebral blood flow; and non-verbal behaviour, e.g. facial expression, gesture, posture. Unlike most other areas of cognitive science, the measures used in the study of social cognition also include qualitative observation of interactions and dialogues, and introspection.

The tools used in the empirical study of social cognition are the same as in other areas of cognitive neuroscience. They include computational modelling, psychophysical techniques, eye movement monitoring, neuronal recording and brain imaging techniques, such as electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), brain stimulation (TMS) and neuropsychological tests. (For more details of these techniques, see the Foresight Research Review *Advanced Neuroscience Technologies*.)

2 MECHANISMS OF SOCIAL COGNITION: HOW DOES THE BRAIN DEAL WITH THE SOCIAL WORLD?

Animals, including humans, have evolved to live in a complicated and often hostile social world, just as much as to live in a complex and often dangerous physical environment. For most animals, survival depends upon their ability to identify the movements, eye gaze and social signals of other creatures, to distinguish whether they are prey, predators or mates and to predict their future actions. As social animals, humans behave largely on the basis of their interpretations of the actions of others. We are continually, and implicitly, reading, analysing and decoding multiple social signals from people around us.

One widely held idea is that we need cognitive systems specialized in dealing with different aspects of the physical and social worlds to maintain our ability to adapt and survive. This specialization allows us to evade (usually) a falling rock on the one hand and a treacherous enemy on the other. How do you know an enemy is treacherous? How do you know you can trust someone? And how do you convince others to trust you? Although we can talk about these questions, reflect on possible mechanisms and devise computational models to emulate them, these skills are deeply intuitive. They are triggered by certain stimuli of which we are not necessarily aware. Often perception turns into action so quickly that deliberation and rational thought have no chance to intervene.

Although we like to think of ourselves as rational creatures, especially when dealing with our fellow humans, our instant reactions are as bound by our evolutionary history as are those of other animals. One social ability that may well be unique to humans is that we can also reflect on our negotiations with the social world. Nevertheless, this ability is also a result of neural processes.

In the following sections we discuss mechanisms that recent research has pinpointed as candidates for explaining our social and communicative competence. These mechanisms are thought to be crucial for reading faces, detecting eye gaze, recognizing emotional expressions, perceiving biological motion, and detecting goal-directed actions and agents. We share these abilities, which appear in humans in the first year of life, with many animals.

Another set of abilities appears later in human development, that is around 18 months, after the end of infancy. By age 5, the major phase of this development appears to be complete, but, of course further learning and refinement happens throughout life. These abilities are observed only in very few non-human animals and even then only in rare instances. They include imitating the intentional actions of others, attending to the same object when directed to do so by another person, and attributing mental states, such as desires and beliefs, to oneself and to other people. It is not always appreciated that the latter group of abilities is just as implicit as the former, and just as pervasive in everyday social understanding and interaction. These pre-eminently human abilities are indispensable prerequisites for teaching and for the informal transmission of cultural knowledge between peers and across generations.

One hypothesis about the evolution of the mechanisms involved is that they build upon the earlier-appearing mechanisms of social cognition that we share with other animals, i.e. those concerned with the perception of faces, eye gaze, biological motion, goal-directed action and agency. However, over and above these mechanisms, a qualitatively different type of mechanism may have evolved. To speculate wildly, this might coincide with the spectacular success of *homo sapiens*, which eclipsed that of other humanoids, such as Neanderthal man. Social rather than physical prowess might have helped *homo sapiens* to dominate others.

2.1 Reading Faces

Babies are born with a very basic, but impressive, capacity to recognize faces. At birth the brain seems to be equipped with information of what a face should look like. Newborn babies prefer to look at drawings of whole faces than at drawings of faces with scrambled features. Within a few days of birth, babies learn to recognize their mother's face. They will look at a picture of their mother's face longer than a picture of a stranger's face.

This remarkable early ability to recognize faces is probably controlled by different brain pathways than later, more sophisticated, face recognition. The early recognition of faces might have evolved because it produces an automatic attachment, or 'imprinting', of new-born babies onto the people they see most (Morton and Johnson, 1991). Early face recognition probably relies on subcortical structures such as the superior colliculus. These structures, below the cerebral cortex, are part of a pathway in the brain that allows us to make movements quickly and automatically on the basis of what we see. It would certainly be useful for newborn babies to imprint onto the faces they see most.

Research on monkeys has shown that a region in the fusiform gyrus contains cells that respond to particular faces in certain orientations – for example, to a profile but not to a front-on view of a particular person's face (Perrett *et al.*, 1992). Research with brain imaging has demonstrated that an equivalent region in the human brain, called the fusiform face area (FFA), responds selectively to faces but not to other visual objects, such as buildings, scenes or objects (Kanwisher, 2000). Only from about two or three months of age do these cortical brain regions start to take over a baby's face recognition ability.

Recent research has demonstrated that human babies are born with the inherent ability to recognize a large number of faces, including faces from other species (monkeys). Only after about 10 months of age do we lose this ability, a process that depends on which types of face we are naturally exposed to (Pascalis *et al.*, 2001). This is analogous to the well-known finding that after about 10 months of age we lose the ability to identify all different sounds. Again this process depends on the sounds we are exposed to during the first 10 months of life (Kuhl, 1994). These findings are important because they highlight the fact that development is, in part, an experience-dependent process that is influenced by the species-specific environment. The rule seems to be fine-tuning rather than indiscriminate adding of information.

2.2 Recognizing Emotional Expressions

Within social psychology, research has demonstrated the ubiquity of facial expression. All cultures use the same expressions for basic emotions such as anger, happiness and sadness (Ekman, 1982). The brain reads facial expressions extremely rapidly. A large number of studies with PET and fMRI, in which subjects observed expressions in different faces, have shown that the amygdala is particularly important for analysing fearful and sad faces, and that this processing often occurs without awareness of the face (Morris *et al.*, 1998; Dolan, 2002). Impairments in emotion recognition are clearly detrimental to social interaction. Imagine not realizing when someone is angry! Normally the effect of seeing an angry face, even for a split second, is to stop you in your tracks or to run away. On the other hand, even within the normal range of individual differences, the response of the amygdala appears to depend to some extent on the personality of an individual, being especially correlated with extraversion and neuroticism (Canli *et al.*, 2001).

2.3 Eye Gaze

The ability to respond to the direction of eye gaze has high evolutionary significance (Emery, 2000). Human babies are automatically drawn to look where another person is looking and prefer direct eye contact (Farroni *et al.*, 2002). The involuntary tendency to look into the same direction as another individual has obvious benefits: the target that another attends to is also likely to be of interest to you. In conjunction with the ability to read emotional expressions, it can allow instant response selection, e.g. approach or avoidance.

A critical neural system implicated in the detection of eye gaze is located in the superior temporal sulcus (STS). Cells in this region in the monkey brain respond to eye gaze direction information from other monkeys or humans (Perrett *et al.*, 1992). In humans, recent studies have found that simply viewing eye gaze stimuli, or stimuli that display animate motion cues, activates a homologous region of the STS amongst other regions (Calder *et al.*, 2002). Direct eye gaze usually indicates threat. This is clearly not the case in humans, who use eye gaze to indicate a wide variety of emotions and intentions, positive as well as negative. Support for this notion comes from a recent fMRI study demonstrating that parts of the brain's reward networks are activated by eye gaze when the eyes belong to someone the subject finds attractive (Kampe *et al.*, 2001).

Although emotion is one major aspect of social cognition, and some social cognitive processes can only be studied within a wider context of emotion processing, there are none the less many aspects of social cognition that are not part of emotion. We deal with these processes in the rest of this section.

2.4 Joint Attention

The attention of babies can be directed to an object by another person simply by looking at them directly and drawing their attention to the object using gaze direction. Attention can also be drawn to an object or event by pointing. In the first year of life, this works only when the object is already in the field of vision. From the middle of the second year of life the attention can be drawn to an object that initially is out of view. This form of triadic attention – that is, interaction between two people about a third object – is thought to be one of the earliest signs of an implicit theory of mind (see section 2.9). Young children with Williams syndrome show the earlier form of dyadic joint attention rather than the later triadic form, thus revealing a dissociation between two social mechanisms that at first glance appear to be similar (Laing *et al.*, 2001).

Joint attention may not be unique to humans. There is evidence, both anecdotal and empirical, that dogs can glean information from joint attention cues (such as pointing and gaze direction) given by humans. This is intriguing because there is no evidence that non-human primates can use this kind of cue from humans. Recently, a rigorous study investigated this evolutionary anomaly. Hare and colleagues (2002) compared the ability of chimpanzees, wolves, dogs and puppies to glean by human pointing information about where an object was hidden. The chimps were no good at this, nor were the wolves, demonstrating that the ability is not inherently canine. However, puppies, with little experience with humans, and therefore unlikely to have learned the significance of pointing, were nevertheless able to use human pointing information to find objects. This demonstrates that the ability of dogs to use joint attention information has been bred over years of domestication. Possibly, the importance of selective mating on social cognition also needs to be considered in human societies.

2.5 Sensitivity to Biological Motion

Among all sensory inputs, one crucial source of information about another creature is their pattern of movement. There are various types of motion in the natural environment. It is essential to detect motion of biological forms in order to predict the actions of other individuals. Here we refer to biological motion as distinct from mechanical, Newtonian motion. Biological motion is self-propelled and non-linear in that it may undergo sudden changes in acceleration, velocity and trajectory. Two different brain regions are concerned with processing biological and mechanical movements.

The Swedish psychologist Johansson (1973) devised an ingenious method for studying biological motion without interference from shape. He attached light sources to the main joints of actors and recorded their movements in a dark environment. He then showed the moving dots to naive perceivers who, rapidly and without any effort, recognized the moving dots as a person walking. Using the same technique, several researchers have demonstrated that observers can recognize not only locomotion, but also the gender of the person, their personality traits and emotions, and complex actions such as dancing represented by moving dots (Koslowski and Cutting, 1978; Dittrich *et al.*, 1996). The ability to distinguish between biological and non-biological movement develops early: 3-month-old babies can discriminate between displays of moving dots with biological motion and displays in which the same dots move randomly (Bertenthal, 1993). This suggests that the detection of biological motion becomes hardwired in the human brain at an early age.

Single-cell studies in macaque monkeys have revealed that STS cells selectively respond to depictions of the face and the body in action (Perrett *et al.*, 1985; Oram and Perrett, 1994; Perrett *et al.*, 1992). STS neurons continue to respond to biological movements even when part of the action is occluded (Jellema and Perrett, 2002). This has been interpreted as demonstrating the contribution of the STS to the visual recognition, representation and understanding of others' actions (Emery and Perrett, 1994). The STS receives information from both dorsal and ventral visual streams – involved in vision for action and vision for identification, respectively – rendering it an interface between perception for identification and perception for action. This combination of visual information would be useful for recognizing the movements of other animate beings and for categorizing them as threatening or enticing. Furthermore, the emotional value of this information is likely to be stored in memory and will enter into predictions about future actions of the agent in question.

Several studies with brain imaging have investigated the neural processing of biological motion in humans. Most of these studies compared brain activity while subjects observed Johansson point-light walkers with brain activity while subjects observed visual stimuli made of the same dots but moving in non-biological ways, such as showing coherent motion (Grossman *et al.*, 2000) and rigid object motion (Grèzes *et al.*, 2001). These studies demonstrated activation of the ventral bank of the STS, often more pronounced in the right hemisphere than in the left hemisphere (Allison *et al.*, 2000). Other neuroimaging studies have detected activation in the right posterior STS in response to seeing hand, eye, and mouth movements (Puce *et al.*, 1998).

2.6 Perception into Action: Mirror Neurons

The notion that actions are intrinsically linked to perception goes back to the nineteenth century when William James, in his ideomotor theory of action, claimed that 'every mental representation of a movement awakens to some degree the actual movement which is its object' (James, 1890). The implication is that observing, imagining or in any way representing an action excites the motor programmes used to execute that same action (Jeannerod, 1994; Prinz, 1997).

Interest in this idea has grown in the past decade due to the discovery of mirror neurons in monkeys (Rizzolatti *et al.*, 1996; Gallese *et al.*, 1996). These neurons are well represented in an area known as ventral premotor cortex (F5) and respond to goal-directed actions of an agent. They respond to an action being carried out by the animal itself (execution) and by the mere observation of the same action being carried out by an experimenter. Mirror neurons appear to distinguish between biological and non-biological motion, responding only to the observation of hand–object interactions and not to the same 'action' performed by a mechanical tool, such as a pair of pliers (see Fig. 7.1) (Rizzolatti *et al.*, 2001).

Mirror neurons provide a perfect example of what we mean by a social cognitive mechanism where there is neurophysiological activity in response to your own and another person's action. Some of the other mechanisms we discussed earlier are conceivable in machines that passively view other animals and categorize their appearance, their eye gaze and their movements. Mirror neurons open up another class of mechanism that may be fundamental to a number of higher level social processes, where the actions of other agents are interpreted in such a way that they directly influence one's own actions. This is the case in the attribution of intentions to others and oneself, and the ability to imitate others as well as to teach others.

There is a large body of evidence that in humans several brain regions are activated both during action generation and during observation of the actions of others (Stephan *et al.*, 1995; Grafton *et al.*, 1996; Decety *et al.*, 1997; Hari *et al.*, 1998). In some brain regions, there is a highly specific overlap between action observation and action execution. Action observation activates premotor cortex according to the body schema that is represented in this region (Buccino *et al.*, 2001).

In an fMRI experiment, subjects observed actions performed by the mouth, hand and foot. These actions were either performed in isolation or with an object, such as chewing food, grasping a cup or kicking a ball. The results demonstrated that watching mouth, hand and foot movements alone, without objects, activates the same functionally specific regions of premotor cortex as making those respective movements. Furthermore, when actions were directed to objects, the parietal cortex became activated. Again, functionally specific regions of the parietal cortex were activated according to the object-directed action being performed.

Observing a movement has measurable consequences on the peripheral motor system (Fadiga *et al.*, 1995). Fadiga and colleagues stimulated left primary motor cortex of human subjects using TMS while

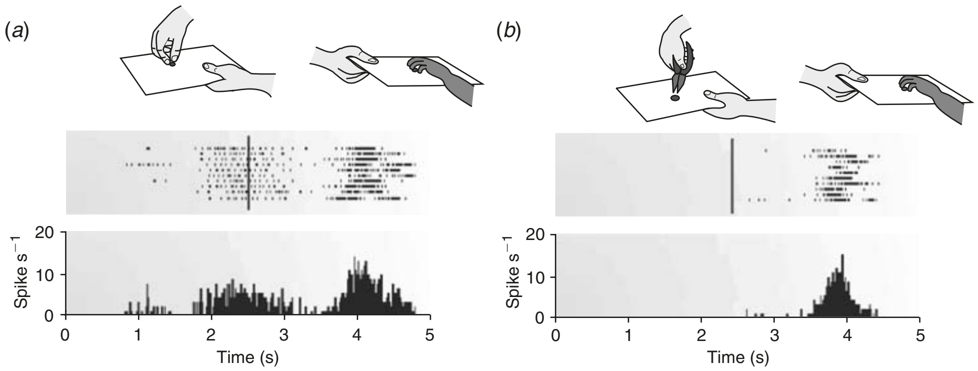


FIGURE 7.1 Mirror neurons respond both during a goal-directed action (for example when the monkey grasps a peanut, shown in both (*a*) and (*b*), and to the sight of a human or another monkey making a goal-directed action (shown in the first half of (*a*). (Reproduced with permission from Rizzolatti *et al.*, 2001, this figure appears in the colour plate section)

the subjects observed meaningless actions and grasping movements (and other control tasks). Motor evoked potentials (MEPs) were recorded from the subjects' hand muscles. It was found that during action observation there was a selective increase of MEPs from the hand muscles that would be used for observed movements.

Rizzolatti and colleagues argue that the mirror system facilitates action understanding, suggesting that we understand other people's actions by mapping observed action onto our own motor representations of the same action. It has been proposed that the mirror system might have evolved to facilitate communication, empathy and the understanding of other people's minds (Gallese and Goldman, 1998). Simulating other people's actions would trigger an action representation from which we could infer the underlying goals and intentions on the basis of what our own goals and intentions would be for the same action. Thus the mirror system is a possible neural mechanism for simulation of other people's actions.

2.7 Detecting Agency

2.7.1 Distinguishing the Self and Other Agents

Given the overlapping brain network that processes action execution and observation, a key question concerns how we are able easily to distinguish the actions we produce from those generated by other people. How do we know who the agent of an action is? Because humans are constantly interacting with others it is crucial to know *who did what*.

The perception of the self as agent is simply 'the sense that I am the one who is causing or generating an action' (Gallagher, 2000). According to Gallagher, a low-level sense of agency, the 'minimal self', is present from birth. Evidence comes from an experiment analysing the behaviour of newborn babies during self-stimulation and external stimulation of the face (Hespos and Rochat, 1997). Here an increase of 'rooting' responses was noted following external as compared to self stimulation.

One mechanism that has been proposed to contribute to the recognition of self-produced action involves the use of internal models (Miall and Wolpert, 1996). It has been proposed that a forward model (an internal representation of the world and the body's kinematics) is used to predict the consequences of self-generated movements using a so-called efference copy of the motor command (Frith *et al.*, 2000). This prediction is then used to determine whether a movement or sensation is self-produced or externally generated by cancelling the results of self-generated sensations (Blakemore *et al.*, 1999). There is evidence that the perceptual attenuation of the sensory consequences of movement is accompanied by, and might be due to, a reduction in activity in regions of the brain that process the particular sensory stimulation being experienced (Blakemore *et al.*, 1998). This predictive system is one mechanism that facilitates the distinction between self and other.

There is accumulating evidence that the parietal cortex plays a role in the distinction between self-produced actions and observed actions generated by others. The right inferior parietal cortex is activated when subjects mentally simulate actions from someone else's perspective but not from their own (Ruby and Decety, 2000). This region is also activated when subjects lead rather than follow someone's actions (Chaminade and Decety, 2002) and when subjects attend to someone else's actions rather than their own (Farrer and Frith, 2002). Patients with parietal lesions have problems in distinguishing their own and others' actions (Sirigu *et al.*, 1999).

2.7.2 Knowing that Something is an Agent Like You

The detection of intentional contingencies, or agency, may be based either on type of motion or on interaction between objects. Movement that is self-propelled is perceived as belonging to an agent with intentions. A second feature that yields attribution of agency to an object is the presence of non-mechanical contingency or causation at a distance. An object that follows another object or reacts to its movement is perceived as driven by internal intentions or goals. Such animacy and contingency features lead to attributions of mental states such as intentions and emotions to simple 2D shapes (Heider and Simmel, 1944). This phenomenon has of course long been exploited in cartoons and virtual reality games.

Even at 12 months of age, infants respond to a blob without any kind of animal or biological features, such as a face, if this blob responds contingently to the infant's actions by beeping or a light flashing. The infants' responses are just like the responses of adults and just like their responses to real people. For example, they look in the same direction as the blob (Johnson *et al.*, 1998; Johnson, 2003). A recent fMRI study investigated activations associated with the perception of intentional contingencies in simple 2D displays involving two moving meaningless shapes. The left STS and the right superior frontal cortex were activated by intentional contingencies, but only when subjects were specifically paying attention to these contingencies (Blakemore *et al.*, 2003). As we shall see (section 2.9), these are a subset of the regions activated during the inference of high level mental states (desires, beliefs).

2.8 Imitation

Motor imitation involves observing the action of another individual and matching one's own movements to those body transformations. The finding that very young babies can imitate certain facial gestures suggests an innate, or early developed, system for coupling the perception and production of movements (Meltzoff and Moore, 1977). This research emphasizes another aspect of the early social responsiveness of the infant but it is not clear how the mechanisms involved relate to later intentional imitation of action. In an enlightening series of experiments, infants of 18 months were exposed either to a human or to a mechanical device attempting, but failing to achieve, various actions, such as pulling apart a dumb-bell (Meltzoff, 1995). The children tended to imitate and complete the action when it was made by the human but not when made by the mechanical device. This demonstrates that preverbal infants' understanding of people, but not inanimate objects, is within a framework that includes goals and intentions, which can be gleaned from surface behaviour alone.

Another experiment showed that children of this age are capable of using what we might call 'common sense' to avoid slavish imitation (Gergely *et al.*, 2002). They imitated an exact movement sequence when the adult pressed a button with the forehead with both hands free. However, they did not imitate when the adult pressed the button with her forehead while holding a shawl around her using both hands. In this case the children generally used their hands to press the button, presumably inferring that the woman would have done so too, had her hands been free. These experiments suggest that imitation might serve, through development, as an automatic way of interpreting the behaviours of others in terms of their underlying intentions and desires.

Recent functional imaging studies have attempted to explore the neural correlates of imitation in the human brain. When we observe another person's actions, brain regions are activated that are involved in motor execution. However, they are more activated when we are told that we should imitate the action later than when we do not have this instruction (Decety *et al.*, 1997). Other brain imaging studies have implicated several different neural structures in imitation, depending on which aspect of an action is imitated (Iacoboni *et al.*, 1999; Chaminade *et al.*, 2002; Koski *et al.*, 2003) and who imitates whom (Decety *et al.*, 2002).

2.9 Theory of Mind

Humans have an inherent ability to understand other people's minds. An experimental paradigm to study this process was first introduced in the early 1980s (Wimmer and Perner, 1983) and has generated much research in developmental psychology (see Wellman *et al.*, 2001). Children at around 4 years start to develop an explicit understanding of the content of other people's minds. They use this understanding in the manner of a theory to predict their behaviour. Hence the term 'theory of mind'. At this age children are aware that people can have different beliefs about states of affairs in the real world, and for good reasons. For instance, they may be told a lie by someone else or they may not be present when vital information is provided about a change in the state of affairs.

However, the implicit attribution of mental states to others is present in children at a much younger age and evidence for the implicit awareness of intentions and desires in others is plentiful from around 18 months. For instance, they understand pretend play and they engage in joint attention. In adults too, there is evidence of both implicit and explicit mentalizing abilities. To investigate neural systems involved in mentalizing, brain imaging studies have used a wide variety of tasks and stimuli, both verbal (stories) and non-verbal (cartoons), which do or do not require an understanding of other people's desires and beliefs. The comparison of mentalizing and non-mentalizing tasks consistently activates at least three brain regions. These are the medial frontal lobe (Brodmann areas 8/9/32), the STS and the temporal poles, adjacent to the amygdala (Frith and Frith, 1999, 2003).

One very implicit mentalizing task involves showing participants animations of moving shapes. As long ago as 1944, Heider and Simmel established that ordinary adults feel compelled to attribute intentions and other psychological motives to animated abstract shapes, simply on the basis of their movement patterns (see section 2.7.2). Castelli *et al.* (2000) showed such animations in a PET study. They contrast sequences where the movements of two triangles were scripted to evoke mental state attributions (e.g. one triangle surprising the other or mocking the other), and sequences where the triangles moved randomly and did not evoke such attributions. This comparison showed activation in the same system as in other studies with different mentalizing tasks.

Imaging experiments have also used interactive games that involve implicit on-the-spot mentalizing. In one such study, researchers scanned volunteers while they played a Prisoner's Dilemma-type game with another person (McCabe *et al.*, 2001). In this game, mutual cooperation between players increased the amount of money that could be won. In the comparison task the volunteers believed they were playing with a computer that used fixed rules. A comparison of brain activation during the game task and the comparison task revealed activity within the medial prefrontal cortex.

The same region was also activated when subjects played 'Stone–Paper–Scissors', a competitive game in which success depends upon predicting what the other player will do next (Gallagher *et al.*, 2002). Again, the comparison condition was created by telling the volunteers that they were playing against a computer. In fact, the sequence of the opponent's moves was the same in both conditions. Participants described guessing and second-guessing their opponent's responses and felt that they could understand and 'go along with' what their opponent, but not the computer, was doing. The medial prefrontal cortex was activated only when the volunteers believed that they were interacting with another person.

What is the involvement of the brain regions that are reliably activated during mentalizing? At present we have only conjectures (Frith and Frith, 2003). It is tempting to conclude that the STS plays a role in mentalizing because it is sensitive to biological motion (see section 2.5). There are exciting speculations about the role of the medial prefrontal cortex, which has direct connections to the temporal pole and to the STS (Bachevalier *et al.*, 1997).

The medial prefrontal region activated by mentalizing studies is the most anterior part of the paracingulate cortex, where it lies anterior to the genu of the corpus callosum and the anterior cingulate cortex (ACC) proper. Although the ACC is an ancient structure that belongs to the limbic lobe, the existence of an unusual type of projection neuron (spindle cell) found in sub-areas of the ACC in humans and some other higher primates (pongids and hominids), but not monkeys, suggests that the ACC has undergone changes in recent evolution (Nimchinsky *et al.*, 1999). Furthermore, in humans these cells are not present at birth, but first appear at approximately 4 months of age (Allman *et al.*, 2001). It remains to be seen whether the recent evolutionary changes observed in ACC are relevant to the other regions where activations associated with mentalizing are observed.

2.10 Deception

Understanding someone else's beliefs, and how these beliefs can be manipulated and maintained, is what we mean by having a 'theory of mind' and underlies the ability to deceive people. The full-fledged ability does not develop until about 5 years, after which time children start to tell lies rather than just physically manipulating situations to hide things from other people (Sodian, 1991).

Functional neuroimaging studies have recently attempted to investigate deception. The confined and artificial context of the brain scanner makes this a difficult task. So researchers have devised tasks in which subjects are instructed to withhold truthful responses and answer with their opposites to questions concerning recent autobiographical events (Spence *et al.*, 2001), or to lie about a card's identity (Langleben *et al.*, 2002) or past events (Lee *et al.*, 2002). These studies have found activations in components of the mentalizing system when subjects are lying. However, whether brain scans can act as lie detectors remains to be seen. Traditional lie detectors measure physiological arousal. Some people at least can train themselves to suppress such responses, just as most people can train themselves not to show telltale signs when lying, such as blushing or evading eye contact. The ingenuity of people to outwit each other and to use bluff and double bluff is an instance of advanced mentalizing ability. Lie detectors may reveal the implicit processes underlying such devious manipulations, but this will be a challenging task.

2.11 Interpretation of Complex Emotions

Complex emotions are the stuff of comedies, tragedies, poetry, novels, films and indeed of everyday life. They have been explored for centuries in many art forms, particularly the theatre. In contrast, mechanisms in the mind and in the brain underlying complex emotions have hardly been studied.

Complex emotions differ from the simple emotions that we might recognize in another person's face. In section 2.1 we mentioned that even split-second exposure to faces expressing fear, sadness, anger and disgust seems to instantly activate amygdala function (Morris *et al.*, 1998; Dolan, 2002; Morris *et al.*, 2002), which may be part of a hard-wired response to threat. Complex emotions – jealousy, envy, pride, embarrassment, resentment, low self-esteem, disdain, empathy, guilt – are different and involve more than an amygdala response. They often imply awareness of another person's attitude to oneself, and an awareness of the self in relation to other people. If so, they are likely to involve the mentalizing system of the brain. These emotions are truly social emotions and probably unique to humans. Research attempting to understand the cognitive and neural processes underlying these emotions and their decoding is only just beginning.

A recent fMRI study scanned the brains of subjects while they were thinking about embarrassing scenarios (Berthoz *et al.*, 2002). Subjects read short vignettes in which social transgressions occurred. These could be accidental or deliberate. In comparison to matched stories in which no transgression occurred deliberate and accidental transgressions both elicited activity in the same three regions that are activated in mentalizing tasks: the medial prefrontal cortex, temporal poles and STS. Activity was also seen in the orbitofrontal cortex, a region involved in emotional processing.

When subjects are asked to make explicit judgements about the trustworthiness of someone based on their eyes, the right STS is activated (Winston *et al.*, 2002). Bilateral STS is activated by faces that subjects found trustworthy compared with faces they did not find trustworthy. Similarly Wicker *et al.* (2002) found that the STS was activated when subjects had to make an explicit emotional attribution about a face.

2.12 Empathy

We need to distinguish between instinctive empathy, or sympathy, and intentional empathy. Instinctive sympathy, accompanied by autonomic responses, is a basic emotional response that is contagious. It is not complex in the sense that the person feeling it has to be aware of their feelings. When somebody is sad and crying, you also become sad and feel an urge to cry. Empathy as a complex emotion is different. It requires awareness of the other person's feelings and of one's own reactions. The appropriate reaction may not be to cry when another person cries, but to reassure them, or even to leave them alone.

At around the age of 2, children start showing sympathy responses when perceiving that another person is upset or in pain (Perner, 1991). Research on empathy has mainly been conducted in the context of lack of empathy – callousness, inability to respond to a victim's distress. Because we are interested in highlighting potential mechanisms of social cognition, we will pick out a few recent brain imaging studies which at least attempt to arrive at such mechanisms.

In a recent fMRI study, subjects were asked to make empathic and forgiving judgements based on hypothetical scenarios (Farrow *et al.*, 2001). Several regions in the superior medial frontal cortex were activated by empathic judgements (subjects had to give an explanation as to why somebody might be acting in a certain way) and forgiving judgements (subjects had to think about which crimes seem most forgivable given a certain situation) compared with the baseline social reasoning judgements.

2.13 Morality

Not so long ago, it would have been considered absurd to search for a brain mechanism underlying morality. Of course, the development of morality does involve cultural input and explicit teaching. The existence of a code of laws has been a major leap in the cultural evolution of social interactions. However, neuroscience has started to tackle the question of a universal sense of morality without which this cultural achievement might not have occurred. Paradigms for studying this question include the ability to make intuitive moral judgements, regardless of any existing code of law.

Even young children seem able to distinguish what is right or wrong in simple stories where conventional rules are broken, and those where 'moral' rules are broken (Smetana *et al.*, 1999). These two kinds of rules are not usually distinguished explicitly. Yet, 4-year-olds can indicate that if permission is given it is alright to break a conventional social rule, talking in class, for example, but not alright to break a rule that prevents harm being done to others, such as hitting another child. Amazingly, even those children who had poor models around them and had themselves been maltreated were unerring in this judgement. This paradigm has not yet been used in scanning studies.

In adults, moral judgements have been found to activate brain regions that are involved in mentalizing, including the medial frontal cortex and the right posterior STS. These regions were activated by morally upsetting stimuli compared with unpleasant pictures that had no moral connotations – a picture of a man assaulting a woman compared with a picture of an injured body, for instance (Moll *et al.*, 2002).

In another study, fMRI was used to scan subjects while they were evaluating moral dilemmas (Greene *et al.*, 2001). An example of a moral dilemma is the train dilemma: a runaway train is heading towards five people who will be killed if the train proceeds on its current course. The only way to save them is to turn the train onto an alternate set of tracks where it will kill one person. Should you turn the train to save five people at the expense of one? Evaluating these problems involves emotional processing, resolving conflict, accommodating cultural beliefs and putting oneself in someone else's shoes. In the study by Greene *et al.*, subjects responded to different types of dilemma, some that were moral, some not; some involved people, others did not. The results showed that the medial frontal cortex was activated by dilemmas that were moral and personal more than by dilemmas that were neither.

This work, though preliminary, demonstrates that we can study the mechanisms of social cognition, even in complex and culturally influenced human interactions that involve the ability to tell right from wrong.

2.14 The Future of Research in Social Cognition

Most of the research we have described in this section is evolving and in its first stages. As well as optimizing and continuing research on the issues discussed in this section, there are many directions in which research could move in the coming years.

2.14.1 Social Competence

There has been a change in people's perception of infants. Infants are more cognitively competent than was previously thought. This has influenced how we perceive and interact with babies. It has even been suggested that the rise in IQ observed from one generation to the next may partly be a result of the type of early interaction that babies have enjoyed in more recent generations. Social competence is achieved from the first moments of life, via parent–child and peer–baby interaction. Since there are individual differences in social competence, and since the opportunity for social experiences varies, there might well be room for improvement. Both formal and informal training might be considered. Perhaps virtual reality games could be a useful tool. Research is required to inform more about individual differences, their causes and whether success through training might be achieved.

Social competence continues to develop throughout childhood and into adulthood. Everyone experiences both positive and negative interactions. Adolescence is a little understood and yet critical time where social roles change and social awareness is still developing. In adolescence, the incidence of antisocial behaviour increases ten-fold (Moffitt, 1993). Understanding the cognitive and neural maturation during this crucial period is vital.

Demographic changes also turn the spotlight on the aged, of whom there will be an increasing number. We are less likely to live in extended communities than before. There are, therefore, fewer opportunities to interact with people from different generations. Political pressures and practical necessity may well lead to new research on social communication in older age groups.

Peer interaction seems to offer some of the most daunting challenges as well as some of the most rewarding activities known to us. It would be useful to develop tools for training in domains that are important when interacting with new people, e.g. emotion reading and mentalizing skills. These tools might involve virtual reality and robots. Educational interventions might improve many areas of social communication, including the management of one's reputation and the understanding of negotiations and compromise. Such interventions should be based on the results of research in social cognition and should be tested rigorously.

2.14.2 Robot Communication

Computers do not currently behave like humans. It feels very different communicating with a cash machine to obtain money than having the same 'conversation' face to face with a bank clerk. The absence of nonverbal communication – raising an eyebrow, smiling, etc. – is a distinctive feature in computer communication. Yet these processes are fundamental to human communication.

We may soon be able to build an artificial being that can understand, and react to, social signals from other people. A robot that can provide social interaction with humans might be useful for people who are isolated for reasons such as health, old age or habitat. It is easy to imagine this as a pet, a living doll or a willing slave, but it could also be a companion and confidant. There is already talk of a robot that can mimic human emotions. This is the first step to understanding emotion. There are robots that can imitate complex human actions they see, such as bouncing a ball on a tennis racquet (Miyamoto and Kawato, 1998). Algorithms exist that make a robot 'know' whether another robot is imitating it correctly. This is exciting progress as imitation may be important in learning, communication and social understanding (see section 2.8).

Researchers are already developing robots that produce human-like emotional expressions, such as smiling and frowning, and which react to facial expressions of people. This may lead to computer agents that interact with the user in different ways according to their emotion. The development of robotic toys that respond to eye gaze, and change their behaviour according to the player's focus of attention, and the development of cars that respond to the driver's attention or aggression, may be key in interdisciplinary research in social cognition. Can we imagine a day when social cognition electronic detectors, analogous to glasses for social short-sightedness, indicate social *faux pas* and aid us in social situations?

A relevant question concerns human–computer interaction and how software is designed to help us to interact better with a computer. This is a vital research question in this climate of high reliance on computers.

2.14.3 Cultural Evolution

A massively under-researched area in cognitive science is how context and culture affect cognition and the brain. This is particularly relevant to cognition during childhood, adolescence and old age, where development seems to depend both on preprogrammed, biological and adaptive responses to novel contexts. Humans are excellent at adapting to their context: cultures are built on this adaptability. Can robots adapt to novel contexts? Can robots who learn to imitate observed actions generalize those actions to similar actions in different contexts? Can robots generalize what they learn by imitation? Can they extract the relevant information, such as goals and intentions, as can young infants (Meltzoff, 1995)? Or do robots simply do slavish imitation?

2.14.4 Psychopharmacology

As we discover the physical (neurotransmitter) mechanisms bound up with social interaction, drugs will be developed that alter social cognition. Narcotics that increase social confidence are available, and popular in the cases of ecstasy, cocaine and marijuana. Other drugs (e.g. antidepressants, Prozac) are often taken by mildly depressed or chronically worried people and result in an increase in self-confidence and social ease. Alcohol has been used for centuries to aid social interaction. On the other hand, alcohol is also a major cause of violence. Can we imagine a time when it is normal and acceptable to pop a pill every morning to facilitate social interactions during the day? Research on this kind of drug, attitudes towards taking such a drug and the ethics of these issues needs to be tackled.

3 WHEN SOCIAL COMMUNICATION FAILS

Biologically caused abnormalities that lead to mild or severe developmental disorders are surprisingly common. They occur in a sizeable proportion of children, estimated at between 5 and 10%. However, severe forms of mental retardation occur in only a very small fraction of cases, 0.3%. Developmental disorders do not just affect children, but more often than not persist lifelong and very often they involve a degree of social impairment. These disorders tend to have a genetic origin but other causes exist as well, for instance, viral illness can attack the brain at a young age.

3.1 Autism

Autism is characterized by difficulties in communication, social interaction and play. It is therefore is a key disorder for the study of social cognition (see Frith, 2003). The cause of autism is most likely a genetic fault that affects brain development from early prenatal life. The signs and symptoms appear only gradually and can often be fully recognized only from the second and third year. This suggests that at least in some cases the early appearing social functions (e.g. detecting agency, biological motion) are intact and only the later appearing ones are affected (e.g. joint attention, mentalizing). However, in other cases, the early appearing social mechanisms may also be disturbed, resulting in more severe and more general social impairments.

Autism comes in many degrees. It can occur together with high intelligence, then usually labelled Asperger syndrome. Some features of autism, such as stereotyped movements and obsession with routines, are not in the social domain at all. However, failure in social communication is the core feature of autism, the feature that unites the many varieties of the autistic spectrum, as it is now called. Because of the wide variety of different forms of autism and their gradual recognition, the number of diagnosed cases has risen enormously in the last decades. According to recent studies autistic disorder is estimated to affect six people per 1000, that is 0.6% of the population.

One striking feature about individuals with autism is that they tend to be more interested in objects than in people. A deficit in the recognition of faces has been recently identified and related to abnormal brain activation in the FFA (see section 2.1). Other social mechanisms, such as a deficit in imitation (Charman *et al.*, 1997) and the inability to recognize emotional expressions are also hypothesized, but still lack systematic investigation. One mechanism that researchers have studied systematically is theory of mind, or mentalizing. This seems to be impaired in all individuals with autistic disorder (see section 2.9), and hence is a strong candidate for an endophenotype of autism.

The normally developing child shows implicit mentalizing from about 18 months. Failure to mentalize can only be observed reliably from that age. Early signs of mentalizing failure in autism are: delays in joint attention; absence of declarative gestures, such as pointing to something without necessarily wanting to get it; and a failure to understand pretence. Imaginative social play (pretend play) is normally pervasive in early childhood. It implies the ability to tell the difference between a real state of affairs and a pretended one. Its absence in autism was a key observation that led to the hypothesis of a mentalizing deficit (Baron-Cohen *et al.*, 1985).

The ‘mind blindness’ hypothesis, tested in different laboratories for the past 20 years or so, has received a large amount of empirical support. It is a specific hypothesis and is consistent with the observation that social competence is globally absent in people with autistic disorder. An example is the poor understanding of deception which coexists with good understanding of sabotage, the latter requiring the ability to distinguish between goodies and baddies and the motivation to win in a competitive game (see Fig. 7.2) (Sodian and Frith, 1992).

Tasks of explicit mentalizing – i.e. predicting someone’s behaviour on the basis of that person’s belief, even if it clashes with the real state of affairs – are an important tool in the study of mentalizing failure in autism. Children with autism who have sufficient verbal ability to follow the scenarios show a delay of about five years before they can pass these tasks (Happé, 1994). However, this slow acquisition of an explicit theory of mind does

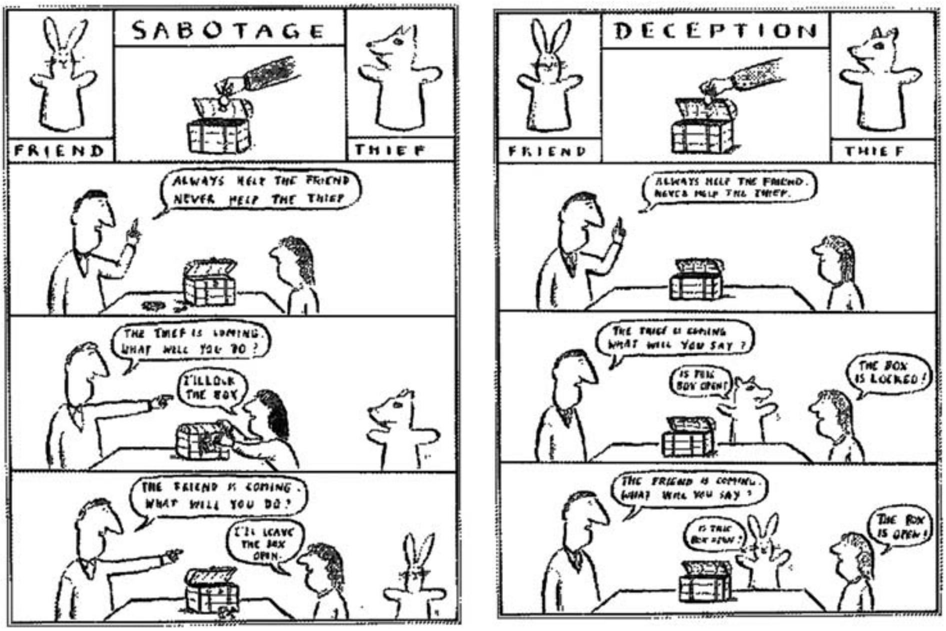

SABOTAGE
FRIEND
THIEF
ALWAYS HELP THE FRIEND NEVER HELP THE THIEF
THE THIEF IS COMING. WHAT WILL YOU DO ?
I'LL LOCK THE BOX
THE FRIEND IS COMING. WHAT WILL YOU DO ?
I'LL LEAVE THE BOX OPEN.
DECEPTION
FRIEND
THIEF
ALWAYS HELP THE FRIEND. NEVER HELP THE THIEF.
THE THIEF IS COMING WHAT WILL YOU SAY ?
IS THE BOX OPEN?
THE BOX IS LOCKED!
THE FRIEND IS COMING. WHAT WILL YOU SAY ?
IS THE BOX OPEN ?
THE BOX IS OPEN!

FIGURE 7.2 Sabotage and deception (Axel Scheffler): example of an experimental design used to study deception as an instance of mental state manipulation. Children with autism were perfectly capable of using sabotage to prevent a 'baddie' from getting at a reward, but were unable to tell a lie to achieve the same aim.

not replace the missing intuitive mentalizing ability. Even very able adults with Asperger syndrome show slow and error prone responses in mentalizing tasks. The brain activation normally shown during mentalizing (see section 2.9) is reduced in individuals with Asperger syndrome and there is weak connectivity between the components of the mentalizing network of the brain (Castelli *et al.*, 2002). Evidence from dementia patients has shown that patients with atrophy of the medial and dorsolateral frontal cortical areas have impaired performance on theory of mind tasks (Gregory *et al.*, 2002).

Post-mortem studies of autistic brains as well as structural imaging studies are rare. So far they suggest that parts of the social brain are anatomically different in autism. These tiny anatomical abnormalities are likely to have a genetic basis and may be a cause for the 'derailment' of normal mental development early in life. More precise data will undoubtedly become available in the near future.

3.1.1 *Compensating for Mind Blindness*

Lacking the innate basis for mind reading does not preclude the ability to learn about mental states. Where intuition is weak, an individual can gradually accumulate an awareness and understanding of mental states. This is achieved by explicitly teaching the person about mental states using logic, memory and detailed explanation of events that happen and what they mean. For an autistic person to learn about other people's intentions and beliefs, the implications of actions, facial expressions, gestures and words need to be spelled out, even if they seem obvious. In this way, able individuals with autism can eventually learn to perform well on mentalizing experiments.

It might be useful to think of teaching someone with autism or Asperger syndrome about the minds of others as like teaching most people about complex mathematics. A small number of people have an intuitive grasp of complex mathematical concepts, which they seem to 'see' without much, if any, conscious effort. Most people lack this intuitive grasp of mathematics but that does not preclude them from learning many of those mathematical concepts, although this requires effort and motivation as well as explicit teaching.

In the same way that learned mathematical ability differs from innate, intuitive mathematical ability, learned mentalizing differs from innate, intuitive mentalizing. Compensatory, explicit learning takes a long time, and is susceptible to more mistakes than intuitive understanding. The knowledge about mental states that results from compensatory learning remains fragile, with frequent errors in the attribution of particular mental states. For example, some autistic children who learn to pass theory of mind tasks in the laboratory cannot generalize their new knowledge of mental states under the stressful demands of real life social situations. In the laboratory they are under no time pressure and can work out logically what the answer should be. Even autistic people who have learned when and how to attribute mental states in principle still lack intuitive mentalizing. They often find it very difficult to make consistent mental state attributions in different, and especially novel, contexts. Nevertheless, to have the ability for non-automatic attribution of mental states is a good enough accomplishment in most cases, and can improve social and communication skills.

3.2 Antisocial Behaviour

Antisocial behaviour is salient and perceived in all societies as intolerable. A mechanism for cheater detection may have evolved as part of the social brain (Tooby and Cosmides, 1990). Deviant behaviour comes under various labels, such as oppositional defiant disorder, conduct disorder, attention deficit disorder and, in adults, antisocial behaviour disorder. These labels at present confound cases with a primarily biological and those of a purely environmental origin.

Of course, biology and environment always interact. Thus, children who grow up in an abusive environment are likely to attribute hostile causes to actions in others. This mechanism may be responsible for the so-called cycle of violence over generations. However, predisposing genes seem to be a prerequisite. An important longitudinal study in New Zealand showed that only those maltreated people who also had a certain predisposing gene later became severely antisocial (Caspi *et al.*, 2002).

Antisocial behaviour is not only a developmental phenomenon, with its roots in childhood experiences and genes, it can also occur out of the blue, as a result of brain damage. Phineas Gage was a young railroad construction supervisor in Vermont, USA, when, in September 1848, a large explosion occurred and a steel rod entered his skull through his left cheek. The rod destroyed his eye, traversed the frontal part of the brain and left the top of the skull at the other side. As a result of this accident, Gage, who could still walk and talk, started to have changes in his personality and mood. In particular, he became extremely anti-social, impulsive, rude and extravagant. Phineas Gage became a classical case in the textbooks of neurology. The part of the frontal lobes which had been damaged, including the orbitofrontal cortex (OFC), is associated with inhibition of inappropriate behaviour, rational decision-making and the processing of emotion. Since Phineas Gage, several patients with OFC lesions have been studied extensively and the same kinds of social impairments have been found (Damasio, 1994). These patients generally have spe-cific deficits in detecting emotional expression and in decision-making that involves emotional evaluation, which demonstrate the importance of relatively low-level emotional cues for understanding other people.

3.3 Psychopathy

The most serious form of antisocial behaviour disorder in childhood leads to psychopathy or antisocial personality disorder in adulthood. This disorder may well have a genetic basis. One ongoing twin study on psychopathic traits shows that they are highly heritable, about 70% of the variance being explained by genetic influences. The rest of the variance was accounted for by non-shared environmental factors, i.e. child-specific environmental influences such as birth order and parent-child interaction (E. Colledge, personal communication).

What kind of neurodevelopmental disorder is psychopathy? Blair (1995) proposed that psychopathy results if there is a fault in the brain system that normally enables instinctive empathy and triggers a violence inhibition mechanism. This idea built on evidence that certain emotional expressions trigger innate brain mechanisms located in circuits involving the amygdala. These circuits can become active in quite subtle situations, causing instinctive reactions to fearful events without any need for awareness of the event (see section 2.1). We do not like to see other creatures suffer or be afraid. When we see fear or hurt in someone's eyes, and we are the cause of it, we tend to stop what we are doing. This is like a reflex that Konrad Lorenz described in fighting dogs and other animals. There are certain signals, so-called submission cues, that tend to make the winning animal stop and shrink back from doing further damage Blair argued that the same reflex exists in humans, and that this reflex is vital for intuitive moral knowledge.

What would be the consequence for development in the case of a fault in the instinctive 'empathy circuit'? Blair *et al.* (2001) predicted and confirmed that such children find it difficult to recognize expressions of fear and sadness. They also have problems in learning moral imperatives, such as not to hurt others, and are unable to distinguish between rules that govern social conventions, and rules that are motivated by a deeper moral sense (see section 2.13).

Although psychopaths lack instinctive sympathy and feel no guilt at having caused harm to another person, they may nevertheless have excellent mentalizing skills (Blair *et al.*, 1996). Not all people with an inability to feel instinctive empathy are excessively violent. If they have no motive to offend, then no harm may come to others. However, individuals who perpetrate violence without pity or remorse are dangerous. From this point of view, and contrary to popular opinion, psychopathy is not the same as being a violent type (Mitchell and Blair, 2000). A violent person, like the fighting dog, may still respond to the distress cues of a victim, stop their action and feel guilt. Aggression can be increased with loss of inhibition, and is a well-known consequence of alcohol intake in some individuals.

3.4 Social Cognition Impairments in Schizophrenia

People with schizophrenia and other mental illnesses have significant social problems. One symptom that carries severe social penalties is a delusion of persecution, in which a person holds a bizarre and paranoid belief with extraordinary conviction, despite experiences to the contrary and counter-arguments. Persecutory delusions are symptoms commonly associated with schizophrenia, but they also occur in other psychiatric disorders including depression, bipolar disorder and schizoaffective disorder.

Within the cognitive approach to psychopathology, it has been argued that processes involved in social inference – that is, the processes by which we interpret the actions of other people and events involving others – play an important role in the development of paranoid delusions (Frith, 1992; Bentall *et al.*, 2001). One such social inference process that may be associated with paranoid delusions involves inferring the causes of social interactions. Individuals readily attribute causes to external events. Bentall and his colleagues have argued that paranoid beliefs may be a product of abnormal causal attributions (Bentall *et al.*, 2001). Overall, research findings support this proposal. Paranoid patients tend excessively to believe that powerful others influenced the course of life (Kaney and Bentall, 1989). Furthermore, patients with persecutory delusions over-attribute negative events to external causes (e.g. Kaney and Bentall, 1989) and to the actions of other people (e.g. Kinderman and Bentall, 1997).

A second type of social inference process has been proposed to underlie delusions of persecution. This involves attributing intentions to other people ('theory of mind'; section 2.9). Frith (1992) has argued that a dysfunctional theory of mind may be implicated in psychotic symptoms including persecutory delusions. Impairments in understanding other people's intentions could lead to the belief that others have malevolent intentions (Corcoran *et al.*, 1995; Frith and Corcoran, 1996; Corcoran *et al.*, 1997).

4 SOME BURNING QUESTIONS FROM EVERYDAY SOCIAL INTERACTIONS

Social competence is one of the most basic and crucial building blocks of society. One aim of the research must be to improve social competence, especially in those people who lack it for genetic or environmental reasons. The ten questions below suggest that there is indeed room for improvement. We pose them here since we believe that future research in social cognition might lead to some answers.

- Are social interactions a major cause of stress (divorce, migration, class)? How can this stress be relieved or prevented?
- How can we use our ability to manipulate other people's beliefs in order to manage our reputations (often maligned as spin)? We all do it, not only politicians.
- How can we be made aware of the insidious nature of prejudice (race, gender, class)?
- What is the effect of negative role models (violent films, fighting games) on imitative behaviour? How can they be counteracted with positive role models?
- How can individual differences in social competence be best assessed? What are the sensitive cognitive tasks that can be used for such measurement?
- When should we be responsible for our actions? If a brain scan shows that someone's brain reacts differently to most people's brains during a certain (e.g. a decision-making) task, does that mean that that person was not responsible for their actions? Do any of us have free will? When the genetic and neural basis of psychopathy is uncovered, what shall we do with people with the genetic potential to become a psychopath? If you can decipher someone's future actions based on their intentions, as determined by some objective measure, could you stop them from executing that action if it were harmful?
- How can we evaluate environmental factors and cultural evolution vs. biological factors, all impacting upon an individual? Can robots, virtual reality and psychopharmacology play a role in this?
- How can we identify more successfully than to date suitable endophenotypes for disorders that are especially important to society, such as depression and schizophrenia as well as developmental psychopathologies?
- To what extent can social cognition be reduced to simple factors, which might be unconscious or even be modelled in experimental animals? Could transgenic mice be useful in the study of social cognition? Is there a mouse without a social gene?
- How best to combine, in a transdisciplinary manner, social psychology and cognitive neuroscience?

Social psychologists have already studied the first five questions successfully for a number of years, but interdisciplinary work might advance this field now. The second five questions await synthesis of some of the more recent research findings on the social brain. The time seems right to identify mechanisms of social cognition and specify their nature in computational and neural terms.

4.1 Transdisciplinary Research

To tackle any of these questions we urgently need research crossing many different disciplines. Research also needs to be easily accessible and applicable to everyday life if it is to have an impact. Application to the real world could also be facilitated by collaboration across disciplines, involving research on at least four levels; the genetic, the neuronal, the cognitive, the social as well as the educational and political levels. It may also be the right time to use often neglected research methods in parallel, such as qualitative reports and interviews, and observational methods in a real-life context.

4.2 Public Engagement with Science

The most obvious and urgent way to facilitate the impact of scientific research on society is to promote public engagement with science. It is fair to say that the public and media have a poor understanding of science – statistics, significance, reproducibility, sample size etc are little understood concepts. And yet these are easy concepts when you master them, so why are they understood so poorly? The public (i.e. us!) are susceptible to scares (BSE, MMR vaccine). How people react is based on how they weigh up risk and cost-benefit. One high-profile article in a newspaper about the possible risks of autism from vaccination might make you wary of vaccinating your baby despite the fact that the evidence presented is highly controversial. On the other hand, you may risk danger of radiation from your mobile phones, but love your mobile too much to abandon it. Irrational decisions such as these are common and understandable.

To give considered information about the slow and error prone process of scientific theories and findings is a huge task. To further the public engagement with science there needs to be a dialogue between scientists, the media, policy-makers and the general public. Scientists need to give more importance to explaining their research. The media, the policy-makers and the general public have a right to better informed and to be better trained in scientific principle such as probabilities, sampling and significance. This multidirectional dialogue is a perfect example of social cognition at work.

We know that people make decisions based on emotion rather than probability (Kahneman and Tversky, 1982). This often leads to irrational decisions, many of which disregard scientific evidence. This may largely be because of a mistrust of science and scientists. The public needs to be able to trust the people who provide information about science – scientists, politicians, funding bodies and so on. This is at the heart of social cognition, which may one day be able to inform how to best manage one's reputation.

References

Allison, T., Puce, A. and McCarthy, G. (2000) Social perception from visual cues: role of the STS region. *Trends Cogn. Sci.*, 4 (7): 267–278.

Allman, J.M., Hakeem, A., Erwin, J.M., Nimchinsky, E. and Hof, P. (2001) The anterior cingulate cortex. The evolution of an interface between emotion and cognition. *Ann. NY Acad. Sci.*, 935: 107–117.

Bachevalier, J., Meunier, M., Lu, M.X. and Ungerleider, L.G. (1997) Thalamic and temporal cortex input to medial prefrontal cortex in rhesus monkeys. *Exp. Brain Res.*, 115 (3): 430–444.

Baron-Cohen, S., Leslie, A.M. and Frith, U. (1985) Does the autistic child have a theory of mind? *Cognition*, 21: 37–46.

Bentall, R.P., Corcoran, R., Howard, R., Blackwood, N. and Kinderman, P. (2001) Persecutory delusions: a review and theoretical integration. *Clin. Psychol. Rev.*, 21 (8): 1143–1192.

Bertenthal, B.I. (1993) Infants' perception of biomechanical motions: intrinsic image and knowledge-based constraints. In C. Granrud (ed.), *Visual Perception and Cognition in Infancy.* Hillsdale, NJ: Erlbaum, pp. 175–214.

Berthoz, S., Armory, J.L., Blair, R.J.R. and Dolan, R.J. (2002) An fMRI study of intentional and unintentional violations of social norms. *Brain*, 125: 1696–1708.

Blair, R.J.R. (1995) A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, 57: 1–29.

Blair, R.J., Colledge, E., Murray, L. and Mitchell, D.G. (2001) A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies *J. Abnormal Child Psychol.*, 29 (6): 491–498.

Blair, R.J.R., Sellars, C., Strickland, I. *et al.* (1996) Theory of mind in the psychopath. *J. Forens. Psychiat.*, 7: 15–25.

Blakemore, S.J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C. and Decety, J. (2003) The detection of contingency and animacy from simple animations in the human brain. *Cerebral Cortex*, 13(8): 837–844.

Blakemore, S.J., Frith, C.D. and Wolpert, D.W. (1999) Spatiotemporal prediction modulates the perception of self-produced stimuli. *J. Cogn. Neurosci.*, 11 (5): 551–559.

Blakemore, S.J., Wolpert, D.M. and Frith, C.D. (1998) Central cancellation of self-produced tickle sensation. *Nature Neurosci.*, 1 (7): 635–640.

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L. *et al.* (2001) Action observation activates premotor and parietal areas in somatotopic manner: an fMRI study. *Europ. J. Neurosci.*, 13: 400–404.

Bourke, A.F.G. (2002) Genetics of social behaviour in fire ants. *Trends Genet.*, 18: 221–223.

Calder, A.J., Lawrence, A.D., Keane, J., Scott, S.K., Owen, A.M., Christoffels, I. and Young, A.W. (2002) Reading the mind from eye gaze. *Neuropsychologia*, 40 (8): 1129–1138.

Canli, T., Sivers, H., Whitfield, S.L., Gotlib, I.H. and Gabrieli, J.D. (2002) Amygdala response to happy faces as a function of extraversion. *Science*, 296: 2191.

Caspi, A., McClay, J., Mofitt, T., Mill, J. *et al.* (2001) Role of genotype in the cycle of violence in maltreated children. *Science*, 297: 851–854.

Castelli, F., Frith, C., Happe, F. and Frith, U. (2002) Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125 (Pt 8): 1839–1849.

Castelli, F., Happé, F., Frith, U. and Frith, C.D. (2000) Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement pattern. *Neuroimage*, 12: 314–325.

Chaminade, T. and Decety, J. (2002) Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport.*, 13 (15): 1975–1978.

Chaminade, T., Meltzoff, A.N. and Decety, J. (2002) Does the end justify the means? A PET exploration of the mechanisms involved in human imitation. *Neuroimage*, 15 (2): 318–328.

Charman, T., Swettenham, J., Baron-Cohen, S., Cox, A., Baird, G. and Drew, A. (1997) Infants with autism: an investigation of empathy, pretend play, joint attention, and imitation. *Dev. Psychol.*, 33 (5): 781–789.

Corcoran, R., Mercer, G. and Frith, C.D. (1995) Schizophrenia, symptomatology and social inference: investigating 'theory of mind' in people with schizophrenia. *Schizophrenia Res.*, 17 (1): 5–13.

Corcoran, R., Cahill, C. and Frith, C.D. (1997) The appreciation of visual jokes in people with schizophrenia: a study of 'mentalizing' ability. *Schizophrenia Res.*, 24 (3): 319–327.

Damasio, A. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Decety, J., Chaminade, T., Grezes, J. and Meltzoff, A.N. (2002) A PET exploration of the neural mechanisms involved in reciprocal imitation. *Neuroimage*, 15 (1): 265–272.

Decety, J., Grèzes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F. and Fazio, F. (1997) Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain*, 120: 1763–1777.

Dittrich, W.H., Troscianko, T., Lea, S.E. and Morgan, D. (1996) Perception of emotion from dynamic point-light displays represented in dance. *Perception*, 25 (6): 727–738.

Dolan, R.J. (2002) Emotion, cognition, and behavior. *Science*, 298 (5596): 1191–1194.

Ekman, Paul (1982) *Emotion in the Human Face*, 2nd edn. Cambridge: Cambridge University Press.

Emery, N.J. (2000) The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.*, 24: 581–604.

Emery, N.J. and Clayton, N.S. (2001) Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, 414: 443–446.

Emery, N.J. and Perrett, D.I. (1994) Understanding the intentions of others from visual signals: neurophysiological evidence. *Curr. Psychol. Cogn.*, 13: 683–694.

Fadiga, L., Fogassi, L., Pavesi, G. and Rizzolatti, G. (1995) Motor facilitation during action observation: a magnetic stimulation study. *J. Neurophysiol.*, 73: 2608–2611.

Farrer, C. and Frith, C.D. (2002) Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage*, 15 (3): 596–603.

Farroni, T., Csibra, G., Simion, F. and Johnson, M.H. (2002) Eye contact detection in humans from birth. *Proc. Natl Acad. Sci. USA*, 99 (14): 9602–9205.

Farrow, T.F., Zheng, Y., Wilkinson, I.D., Spence, S.A., Deakin, J.F., Tarrier, N., Griffiths, P.D. and Woodruff, P.W. (2001) Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*, 12 (11): 2433–2438.

Frith, C.D. (1992) *The Cognitive Neuropsychology of Schizophrenia.* Lawrence Erlbaum.

Frith, C.D. and Corcoran, R. (1996) Exploring 'theory of mind' in people with schizophrenia. *Psychol. Med.*, 26 (3): 521–530.

Frith, C. D. and Frith, U. (1999) Interacting minds – a biological basis. *Science*, 286: 1692–1695.

Frith, C.D. and Frith, U. (2003) Development and neurophysiology of mentalizing. *Phil. Trans. R. Soc. Lond. Biol. Sci.*, 1431: 459–474.

Frith, C.D., Blakemore, S-J. and Wolpert, D.M. (2000) Abnormalities in the awareness and control of action. *Phil. Trans. R. Soc. Lond. Biol. Sci.*, 355: 1771–1788.

Frith, U. (2003) *Autism. Explaining the Enigma*, 2nd edn. Oxford: Blackwell.

Gallagher, H.L., Jack, A.I., Roepstorff, A. and Frith, C.D. (2002) Imagining the intentional stance. *Neuroimage*, 16: 814–821.

Gallagher, S. (2000) Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.*, 4 (1): 14–21.

Gallese, V. and Goldman, A. (1998) Mirror neurons and the simulation theory of mind reading. *Trends Cogn. Sci.*, 2: 493–501.

Gallese, V., Fadiga, L., Fogassi, L. and Rizzolatti, G. (1996) Action recognition in the premotor cortex. *Brain*, 119: 593–609.

Gergely, G., Bekkering, H. and Király, I. (2001) Rational imitation in preverbal infants. *Nature*, 415: 755.

Grafton, S.T., Arbib, M.A., Fadiga, L. and Rizzolatti, G. (1996) Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Exp. Brain Res.*, 112 (1): 103–111.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. and Cohen, J.D., (2001) An fMRI investigation of emotional engagement in moral judgment. *Science*, 293: 2105–2108.

Gregory, C., Lough, S., Stone, V., Erzinclioglu, S., Martin, L., Baron-Cohen, S. and Hodges, J.R. (2002) Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer's disease: theoretical and practical implications. *Brain*, 125 (Pt 4): 752–764.

Grèzes, J., Fonlupt, P., Bertenthal, B., Delon-Martin, C., Segebarth, C. and Decety, J. (2001) Does perception of biological motion rely on specific brain regions? *Neuroimage*, 13: 775–785.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V. *et al.* (2000) Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.*, 12: 711–720.

Happe, F.G. (1994) An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J. Autism Dev. Disord.*, 24 (2): 129–154.

Hare, B., Brown, M., Williamson, C. and Tomasello, M. (2002) The domestication of social cognition in dogs. *Science*, 298: 1634–1636.

Hari, R., Fross, N., Avikainen, E., Kirveskari, E., Salenius, S. and Rizzolatti, G. (1998) Activation of human primary motor cortex during action observation: aneuromagnetic study. *Proc. Natl Acad. Sci. USA*, 95: 15061–15065.

Hariri, A.R. *et al.* (2002) Serotonin transporter genetic variation and the response of the huyman amygdala. *Science*, 297: 400–403.

Heider, F. and Simmel, M. (1944) An experimental study of apparent behavior. *Am. J. Psychol.*, 57: 243–249.

Hespos, S.J. and Rochat, P. (1997) Dynamic mental representation in infancy. *Cognition*, 64 (2): 153–188.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C. and Rizzolatti, G. (1999) Cortical mechanisms of human imitation. *Science*, 286: 2526–2528.

James, W. (1890) *Principles of Psychology.* New York: Holt.

Jeannerod, M. (1994) The representing brain – neural correlates of motor intention and imagery. *Behav. Brain Sci.*, 17: 187–202.

Jellema, T. and Perrett, D.I. (2002) Neural coding for visible and hidden objects. *Attention Performance*, XIX: 356–380.

Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. *Perception Psychophysics*,14: 201–211.

Johnson, S. (2003) Detecting agency. *Proc. R. Soc. Lond. Ser. B* (in press).

Johnson, S., Slaughter, V. and Carey, S. (1998) Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Dev. Sci.*, 1: 233–238.

Kahneman, D. and Tversky, A. (1982) On the study of statistical intuitions. *Cognition*, 11 (2): 123–141.

Kampe, K.K., Frith, C.D., Dolan, R.J. and Frith, U. (2001) Reward value of attractiveness and gaze. *Nature*, 413 (6856): 589.

Kaney, S. and Bentall, R.P. (1989) Persecutory delusions and attributional style. *Br. J. Med. Psychol.*, 62 (Pt 2): 191–198.

Kanwisher, N. (2000) Domain specificity in face perception. *Nature Neurosci.*, 3 (8): 759–763.

Kinderman, P. and Bentall, R.P. (1997) Causal attributions in paranoia and depression: internal, personal, and situational attributions for negative events. *J. Abnorm. Psychol.*, 106 (2): 341–345.

Koski, L., Iacoboni, M., Dubeau, M.C., Woods, R.P. and Mazziotta, J.C. (2003) Modulation of cortical activity during different imitative behaviors. *J. Neurophysiol.*, 89 (1): 460–471.

Koslowski, L.T. and Cutting, J.E. (1978) Recognising the sex of a walker from point-lights mounted on ankles: some second thoughts. *Percept. Psychophys.*, 23: 459.

Kuhl, P.K. (1994) Learning and representation in speech and language. *Curr. Opin. Neurobiol.*, 4 (6): 812–822.

Laing, E., Hulme, C., Grant, J. and Karmiloff-Smith, A. (2001) Learning to read in Williams syndrome: looking beneath the surface of atypical reading development. *J. Child Psychol. Psychiat.*, 42 (6): 729–739.

Langleben, D.D., Schroeder, L., Maldjian, J.A., Gur, R.C., McDonald, S., Ragland, J.D., O'Brien, C.P. and Childress, A.R. (2002) Brain activity during simulated deception: an event-related functional magnetic resonance study. *Neuroimage*, 15 (3): 727–732.

Lee, T.M., Liu, H.L., Tan, L.H., Chan, C.C., Mahankali, S., Feng, C.M., Hou, J., Fox, P.T. and Gao, J.H. (2002) Lie detection by functional magnetic resonance imaging. *Human Brain Mapping*, 15 (3): 157–164.

McCabe, K., Houser, D., Ryan, L., Smith, V. and Trouard, T. (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl Acad. Sci. USA*, 98 (20): 11832–11835.

Meltzoff, A.N. (1995) Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Dev. Psychol.*, 31: 838–850.

Meltzoff, A.N. and Moore, M.K. (1977) Imitation of facial and manual gestures by human neonates. *Science*, 198: 75–78.

Miall, R.C. and Wolpert, D.M. (1996) Forward models for physiological motor control. *Neural Networks*, 9: 1265–1279.

Mitchell, D. and Blair, R.J.R. (2000) The psychopath: an individual with an emotional impairment? *The Psychologist*, 13: 356–360.

Miyamoto, H. and Kawato, M. (1998) A tennis serve and upswing learning robot based on bi-directions theory. *Neural Networks*, 11: 1331–1344.

Moffitt, T.E. (1993) Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychol. Rev.*, 100 (4): 674–701.

Moll, J., de Oliveira-Souza, R., Eslinger, P.J., Bramati, I.E., Mourao-Miranda, J., Andreiuolo, P.A. and Pessoa, L. (2002) The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *J. Neurosci.*, 22 (7): 2730–2736.

Morris, J.S., deBonis, M. and Dolan, R.J. (2002) Human amygdala responses to fearful eyes. *Neuroimage*, 17 (1): 214–222.

Morris, J.S., Ohman, A. and Dolan, R.J. (1998) Conscious and unconscious emotional learning in the human amygdala. *Nature*, 393: 467–470.

Morton, J. and Johnson, M.H. (1991) CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychol. Rev.*, 98 (2): 164–181.

Nimchinsky, E.A., Gilissen, E., Allman, J.M., Perl, D.P., Erwin, J.M. and Hof, P.R. (1999) A neuronal morphologic type unique to humans and great apes. *Proc. Natl Acad. Sci. USA*, 96 (9): 5268–5273.

Oram, M.W. and Perrett, D.I. (1994) Responses of anterior superior temporal polysensory (STPa) neurons to biological motion stimuli. *J. Cogn. Neurosci.*, 6: 99–116.

Pascalis, O., de Haan, M. and Nelson, C. (2001) Is face processing species-specific during the first year of life? *Science*, 296: 1321–1322.

Perner, J. (1991) *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Perrett, D.I., Hietanen, J.K., Oram, M.W. and Benson, P.J. (1992) Organization and functions of cells responsive to faces in the temporal cortex. *Phil. Trans. R. Soc. Lond. Ser. B*, 335: 23–30.

Perrett, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D. and Jeeves, M.A. (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond. Ser. B*, 223: 293–317.

Piaget, J. (1972) *The Psychology of the Child*. New York: Basic Books.

Prinz, W. (1997) Perception and action planning. *Eur. J. Cogn. Psychol.*, 9: 129–154.

Puce, A., Allison, T., Bentin, S., Gore, J.C. and McCarthy, G. (1998) Temporal cortex activation in humans viewing eye and mouth movements. *J. Neurosci.*, 18 (6): 2188–2199.

Rizzolatti, G., Fadiga, L., Fogassi, L. and Gallese, V. (1996) Premotor cortex and the recognition of motor actions. Brain Research. Cognitive.{?9} *Brain Res.*, 3: 131–141.

Rizzolatti, G., Fogassi L. and Gallese, V. (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Rev. Neurosci.*, 2 (9): 661–670.

Ruby, P. and Decety, J. (2001) Effect of the subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neurosci.*, 4: 546–550.

Sirigu, A., Daprati, E., Pradat-Diehl, P., Franck, N. and Jeannerod, M. (1999) Perception of self-generated movement following left parietal lesion. *Brain*, 122: 1867–1874.

Smetana, J.G. *et al.* {?10} (1999) Maltreated and non-maltreated preschoolers' conceptions of hypothetical and actual moral transgressions. *Dev. Psychol.*, 35: 269–281.

Sodian, B. (1991) The development of deception in young children. *Br. J. Dev. Psychol.*, 9: 173–188.

Sodian, B. and Frith, U. (1992) Deception and sabotage in autistic, retarded and normal children. *J. Child Psychol. Psychiat.*, 33 (3): 591–605.

Spence, S.A., Farrow, T.F., Herford, A.E., Wilkinson, I.D., Zheng, Y. and Woodruff, P.W. (2001) Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, 12 (13): 2849–2853.

Stephan, K-M., Fink, G., Passingham, R.E., Silbersweig, D., Ceballos-Baumann A.O., Frith, C.D. and Frackowiak, R.S.J. (1995) Functional anatomy of mental representation of upper extremity movements. *J. Neurophysiol.*, 73: 373–386.

Tooby, J. and Cosmides L. (1990) On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. *J. Personality*, 58 (1): 17–67.

Vygotsky, L. and Vygotsky, S. (1980) *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wellman, H.M., Cross, D. and Watson, J. (2001) Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.*, 72 (3): 655–684.

Wicker, B., Perrett, D.I., Baron-Cohen, S. and Decety, J. (2002) Being the target of another's emotion: a PET study. *Neuropsychologia*, 41 (2): 127–138.

Wimmer, H. and Perner, J. (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13 (1): 103–128.

Winston, J.S., Strange, B.A., O'Doherty, J. and Dolan, R.J. (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, 5 (3): 277–283.

CHAPTER

8

Motivation, Planning and Interaction

Nicholas R. Jennings and Anthony G. Cohn
with contributions from
Maria Fox, Derek Long, Michael Luck, Danius T. Michaelides,
Steve Munroe and Mark J. Weal

1 INTRODUCTION

Artificial intelligence (AI) has long concerned itself with developing autonomous agents. These are computer programs or automatons that can operate in various environments, be it the physical world or an artificial one. The early pioneers of AI set lofty goals for designing and building intelligent entities with capabilities like those of humans. Moreover, a significant proportion of the early work drew much of its inspiration and many of its metaphors from studying humans and the ways and means by which they seemed to exhibit intelligent behaviour.

Early optimism foundered when it became apparent that each component of an artificial intelligent agent was a challenge in its own right. This realization caused the research to fragment into many sub-fields. These worked more or less in isolation for

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

many years. In the past 10 years, however, the holistic approach has made a comeback. Many researchers are now attempting to build intelligent agents that combine various components. In most cases, though, today's researchers have much more realistic aims about the likely competence of the agents they devise.

While there is still much debate about exactly what constitutes agenthood, many researchers accept the following characterization: 'An agent is an encapsulated computer system that is situated in some environment, and that is capable of flexible, autonomous action in that environment in order to meet its design objectives' (Jennings, 2001).

This definition raises a number of points that require elaboration. Agents are:

- clearly identifiable problem-solving entities with well-defined boundaries and interfaces
- situated (embedded) in a particular environment over which they have partial control and observability – agents receive inputs related to the state of their environment through sensors and act on the environment through effectors
- designed to fulfil a specific role – agents have particular objectives to achieve
- autonomous – agents control both their internal state and their own behaviour
- capable of exhibiting flexible problem-solving behaviour in pursuit of their design objectives – being both reactive, able to respond in a timely fashion to changes that occur in their environment, and proactive, able to opportunistically adopt goals and take the initiative (Wooldridge and Jennings, 1995).

When adopting an agent-oriented view, it soon becomes apparent that most problems require or involve multiple agents. These are needed to represent the decentralized nature of a problem, the multiple loci of control, the multiple perspectives or competing interests. Moreover, the agents need to interact with one another: either to achieve their individual objectives or to manage the dependencies that ensue from being situated in a common environment.

These interactions can vary from simple information passing, through traditional client–server interactions, to rich social interactions that involve the ability to cooperate, coordinate and negotiate about a course of action. Whatever the nature of the social process, two points differentiate agent interactions from those in other software engineering paradigms.

First, agent-oriented interactions generally occur through a high-level, declarative, agent communication language that is often based on speech act theory (Mayfield *et al.*, 1995). Consequently, interactions take place at the 'knowledge level' in terms of which goals to follow, at what time and by whom (Newell, 1982).

Secondly, agents are flexible problem-solvers. They operate in an environment over which they have only partial control and observability. This means that their interactions need to be handled in a similarly flexible manner. Agents therefore need the computational apparatus to make context dependent decisions about the nature and scope of their interactions, and to initiate, and respond to, interactions that were not foreseen at design time.

Against this background, this chapter deals with three key issues related to designing and building intelligent agents:

- How can an intelligent agent determine what its aims and objectives should be at any given time?
- How can an agent plan a series of actions to achieve its aims and objectives?
- How can an agent interact with humans and other software agents to complete its objectives?

We deal with each issue in the sections on 'motivation' (section 2), 'planning' (section 3) and 'interaction' (section 4). Given the trend towards more complete agents, it is also timely to look at recent insights from studies of human beings, the ultimate integrated agent, and to examine how they affect the development of this artificial counterpart. To this end, we draw out links and potential synergies with research in the neurosciences.

2 AGENT MOTIVATION

Much of computing, especially AI, is conceptualized as taking place at the 'knowledge level', with computation defined in terms of 'what' to do, or 'goals'. Computation can then achieve those goals, as is typical in planning, for example (section 3).

We do not usually consider the reasons for the goals arising. However, these reasons may have a substantial influence over how to achieve the goals. If goals determine what to do, these reasons, or 'motivations', determine 'why' and consequently how.

The best illustration of the role of motivation in computing is perhaps in relation to autonomous agents. In essence, these possess goals that are generated within, rather than adopted from other agents (Luck and d'Inverno, 1998). Agents generate these goals from motivations, higher-level non-derivative components that characterize the nature of the agent. The goals can be considered to be the desires or preferences that affect the outcome of a given task of reasoning or behaviour.

For example, 'greed' is not a goal. It does not specify a state of affairs to achieve. Nor is it describable in terms of the environment. However, greed may generate a goal to rob a bank. The motivation of greed and the goal of robbing a bank are clearly distinct, with the former providing a reason to do the latter, and the latter specifying how to satisfy the former. In a computational context, we can imagine a robot that explores its environment in an effort to construct a map, but must sometimes recharge its batteries.

These motivations of 'curiosity' and 'hunger' lead to the generation of specific goals at different times. The balance of importance of these goals changes as time passes.

Similarly, when undertaking a reasoning task, the nature and degree of reasoning that is possible must depend on the need for it. For example, in a critical medical emergency, a coarse but rapid response may be best. Experimental trials of new medical treatments, on the other hand, require repeatability and accuracy, often regardless of the time it takes. Motivation is the distinguishing factor between these two scenarios.

This view is based on the generation and transfer of goals between agents. More specifically, agent-based computing generally operates at the knowledge level where goals are the currency of interaction. Goals specify what the agent must achieve without specifying how. In that sense, goals enable individual agents to choose the best means available to them in deciding how to achieve goals.

Although this gives a large degree of freedom in the dynamic construction of multi-agent systems, virtual organizations etc., it provides little by way of direction, guidance or meta-level control that may be valuable in determining how best to achieve overarching aims. Motivations address this both by providing the reasons for the goal, and by offering constraints on how best to achieve the goal when faced with alternative courses of action.

Motivation as an area of study is thus important in two key respects. First, and most intuitively, it is critical in understanding human and animal behaviour. Computational models can aid in testing relevant hypotheses. Secondly, from a computer science, and altogether more pragmatic, perspective, it could offer a substantially higher level of control than is currently available. This will become more important for agent systems that need to function in an autonomous yet persistent manner while responding to changing circumstances.

Autonomous agents are powerful computational tools. However, without the constraints that motivations could provide, agents may lack the required behavioural control. A combination of these two aspects may also permit computational agents to better understand and reason about another, possibly animate, agent's motivations, both in application to computational multi-agent systems and to the human–computer interface.

2.1 Motivation in Psychology, Ethology and Computer Science

'Motivation' does not refer to a specific set of readily identified processes, though for practical purposes motivation can be discussed in terms of 'drives' and 'incentives', the push and pull of behaviour (Halliday, 1983). Drives are internally generated signals that tell an organism that it has violated a homeostatic balance such as hunger, thirst etc. There are also circadian drives such as sleep and wakefulness.

Incentives originate outside of an organism and can vary in their attractiveness to the organism arousing more or less motivation. Incentives can be both positive and negative. For example, a positive incentive usually causes 'approach' behaviours such as a person deciding to buy a car because of its attractive specifications. A negative incentive causes avoidance behaviours, such as a shy person avoiding social interaction.

Motivation has long been seen as a key concept in the organization of behaviour within the psychological and ethological sciences. In computer science, however, the focus is to use motivations to provide an effective control mechanism to govern the behaviour and reasoning of autonomous agents.

In cognitive psychology, researchers come close to an understanding of motivation that is valuable in a computational context. Kunda informally defines motivation to be 'any wish, desire, or preference that concerns the outcome of a given reasoning task' (Kunda, 1990). Kunda also suggests that motivation affects reasoning in a variety of ways, including accessing, constructing and evaluating beliefs and evidence, and decision-making. Much experimental work has set out to explicate these thoughts but research is just beginning to put them in a computational context.

One early example takes motivation to be 'that which controls attention at any given time' (Simon, 1979). Simon explores the relation of motivation to information-processing behaviour, but from a cognitive perspective. Sloman has elaborated on this work and has shown how motivations are relevant to emotions and the development of a computational theory of mind (Sloman, 1987; Sloman and Croucher, 1981).

We can consider problem-solving as finding actions that achieve the current goals. In this way, goals provide the reason and context for behaviour. But how are the goals to be chosen? Typically an agent chooses a goal if the environmental circumstances support the necessary preconditions for that goal. That is, the external context determines goal selection. In real biological agents, however, the same environmental cues often elicit different behaviour. This can be attributed to an agent's current motivations. Computational agent-based systems often lack such an internal context.

Clearly, this is inadequate for research on modelling autonomous agents and creatures. This requires an understanding of how to generate and select such goals. Additionally, it is inadequate for research that aims to provide flexibility of reasoning in a variety of contexts, regardless of concerns with modelling artificial agents. Such flexibility can be achieved through the use of motivations that can lead to different results, even when the goals remain the same (Luck, 1993).

In his development of Simon's ideas, Sloman argues explicitly for the need for a 'store of "springs of action"', motives (Sloman and Croucher, 1981). For Sloman, motives represent to the agent what to do in a given situation and include desires, wishes, tastes, preferences and ideals.

The key to Sloman's concept of motives is their role in processing. Importantly, Sloman distinguishes between two types of motives. First-order motives directly specify goals, whereas second-order motives generate new motives or resolve conflicts between competing motives. These are termed 'motive generators' and 'motive comparators'.

According to Sloman, a motive produced by a motive generator may have the status of a desire. This relatively early work presents a broad picture of a two-tiered control of behaviour. Motives occupy the top level.

They provide the drive or urge to produce lower-level goals that specify the behaviour itself.

In subsequent work, the terminology changes to distinguish between 'non-derivative motivators' and 'derivative motivators', rather than between motivators and goals themselves. Nevertheless, the notion of derivative and non-derivative mental attitudes makes it clear that there are two levels of attitude, one which is in some sense innate, and which gives rise to the other, which is a result of the first.

In a different context, the second of Waltz's 'Eight Principles for Building an Intelligent Robot' requires the inclusion of 'innate drive and evaluation systems to provide the robot with moment-to-moment guidance for its actions' (Waltz, 1991). In elaborating this principle, Waltz explains that the action of a robot at a particular time should not be determined just by the current sensory inputs, but also by the 'desires' of the robot, such as minimizing energy expenditure, laziness, and maintaining battery power, hunger.

Moffat and Frijda (1995) use a similar concept which they term 'concerns'. These are 'dispositions to prefer certain states and/or dislike others'. In their model, an agent selects the most relevant information perceived through its sensors. The relevance of an event comes from the agent's concerns. Thus, for example, if an agent detects food in its environment, and, if this event is relevant to its hunger concern, this may generate a goal to move towards the food and eat it. The most relevant event causes a signal which in turn instantiates the relevant goal.

2.2 Computational Modelling of Motivations

Research into robotics, artificial life and autonomous agents and creatures provided the impetus for a growth of interest in computational modelling of motivations. Researchers have developed different representations for motivations and mechanisms for manipulating them (Balkenius, 1993; Halperin, 1991).

Responses to a particular stimulus can vary depending both on the internal state of the agent and/or the external situation – i.e. the environment. If the external situation remains constant, differences in response must result from changes in the internal state of the responding agent. These differences are due to the motivations of the agent.

We can think of an agent as having a fixed range of identifiable motivations of varying strength. We can regard these motivations as being innate. Certain behaviours may be associated with one or more motivations. For example, sexual courtship might be associated with the motivation for reproduction. Executing the courtship behaviour may enable an agent to procreate with the partner, which will typically mitigate the motivation to reproduce. These behaviours are known as 'consummatory behaviours'. Other behaviours, such as courtship displays, known as 'appetitive behaviours', make the conditions of the consummatory behaviour come true.

This is a somewhat simplified view of motivation. Although much behaviour occurs in functional sequences with appetitive behaviours leading to consummatory ones, there can be complex interactions between motivations and behaviours (Hinde, 1982). For example, a single situational cue could relate to many motivations that could release many activities, or cause an action that leads to other behaviours, or even decrease some motivations so that others would increase. In addition, there are inhibitory relationships between behaviours in animals as well as relationships that can increase the strength of other behaviours. Moreover, the combination of motivations may lead to different or variable behaviours.

These are all difficult issues that we must address in attempting to construct accurate behavioural models of real and artificial agents. We must address them even if the concern is not with providing such accuracy, but in constructing simple yet adequate models that allow effective control of behaviour.

At present, most work on motivation is under one of two main areas: developing lifelike or believable agents; and developing models of motivation. In the former, much of the effort concerns human–computer interface in aiming to build applications with substantial user interactions, mediated by computational agents (section 4). Motivation here provides richer behaviour for artificial agents: it enables engaging and sophisticated interaction, both through the agents and through better user models.

Some of this research also aims to develop agents for entertainment systems, such as games, in which self-motivated agents can provide more effective autonomous behaviour (Elliott and Brzezinski, 1998). The latter work concentrates on the fundamental aspects of developing appropriate, accurate and effective models to focus the attention of a reasoning system on salient aspects, for example.

2.3 Modelling Motivation

We can define autonomous agents to be agents with a higher level control that is provided internally by motivations. Thus we can specify motivations of 'curiosity', 'safety', 'fear', 'hunger' and so on. In a simple agent, we might associate the motivation of 'safety' with the goal of avoiding obstacles which, in turn, is associated with the actions required to achieve this result.

Motivations also vary over time according to the internal state of an agent. For example, if an agent spends a long time without food, then the hunger motivation increases. When the agent feeds, the hunger motivation decreases. The common view is that each motivation has an associated strength or intensity value (Sloman, 1987). This strength can be either variable, depending on external and internal factors, or fixed at some constant value.

We can regard an autonomous agent as embodying a set of motivations, that is a function of the kind of agent we are considering. At a particular point in time, each motivation in this set of motivations is a function of an instance of a particular kind of agent and its environment together. In order to act on motivations, the strength may have to exceed a threshold value to force action. Alternatively, the value with the highest strength may be used to which motivation is currently in control.

More sophisticated mechanisms are also possible (Sloman, 1987; Beaudoin and Sloman, 1993; Moffat *et al.*, 1993; Moffat and Frijda, 1995; Norman and Long, 1995a,b). In addition, other representations for motivations and mechanisms for manipulating them have been developed at both subsymbolic and symbolic levels (Maes, 1989a,b, 1991; Halperin, 1991; Schnepf, 1991).

2.4 Open Questions

While there are several models of motivation, they are limited. The use of motivation has been seen in limited one-off applications. For example, Sloman and colleagues have done much to develop models of attention and affect, but without real application. Norman and Long have studied alarms as a means of concentrating attention through motivation on particularly important issues at any time.

In terms of application, some research has addressed the use of lifelike agents as part of the interface in virtual theatre, teaching and various presentation systems (Andre and Rist, 2000). However, they have typically been with much weaker models of motivation. Perhaps more interesting are applications in which the effect of the motivational component is not just to give more realistic or rich interactions, but to provide effective constraints on action and behaviour for better functionality. These examples are fewer, but include the work on planning, and on machine discovery, goal selection and autonomous agent interaction. For example, recent research describes an AI planning framework that continually generates goals in response to changes in both the environment and the agent's motivations (Coddington, 1998, 2001; Aylett and Coddington, 2000).

Each newly generated goal has an associated value indicating its importance or priority. The value is determined by the strength of the motivations that influenced the creation of the goal. This allows the planner to choose to achieve high priority goals in preference to those with low priority. Changes to the environment that arise as a consequence of executing actions in the plan change the agent's motivations. Executing an action will support or undermine the agent's motivations to a lesser or greater degree. It is therefore possible to predict the future strength associated with motivations by examining the sequence of actions in a plan. When generating solution plans, this strategy allows the agent to choose the solution plan containing a sequence of actions that best supports the agent's motivations.

In a similar fashion, Luck describes the use of motivations in biasing processes underlying machine discovery to address issues relating to urgency, importance and context (Luck, 1993). Different resources may be available, depending on the nature of the reasoning task, and on the motivations of the reasoning agent. The reasoning, or learning, undertaken needs to reflect these demands, for example by providing rapid but coarse solutions, or slower but more accurate ones.

Researchers have applied the same ideas in consideration of reasoning in relation to autonomous interaction. We can view this as a reasoning process with uncertain outcome. Clearly, the potential impact and applicability is strong, but will require much more work, especially in these latter areas. We discuss some possibilities for further consideration below.

2.4.1 *Utility*

In many agent architectures, the notion of 'utility' helps an agent to decide what to do at any given time. Motivation and utility perform similar tasks for agents – they tell an agent what to do – but motivation is a more wide-ranging concept than utility.

Utility is the value placed on a course of action in a given situation. Good courses of action have high utility, and poor actions have low utility. An agent examines its options and chooses those with high utility. Motivation on the other hand, whilst also performing this task, is involved more intimately with the agent's decision-making process.

A motivated agent has a dynamically changing internal environment, provided by motivations, that can influence its decisions. For example, in the presence of food, an agent may or may not choose to eat depending on the state of its internal environment, specifically its hunger motivation.

Utility is the end result of a process of deliberation about the options available to an agent, given its external and internal environments. In many systems, the designer calculates the utility for a given action in advance. While this may change with different circumstances, it is typically hard-wired. By contrast, motivated agents should be able to calculate utility on-the-fly, based on weightings provided by their current motivational state.

The application of motivation mechanisms to mediating agent interactions is perhaps the most obvious immediate avenue to explore. The scope here is great, with motivations providing a means by which the reasons underlying such interactions can contribute to, and constrain, the formation, operation and dissolution of multi- agent systems and virtual organizations, for example.

To focus on negotiation in particular, one current difficulty is in eliciting user preferences and incorporating them into the negotiation process. One possibility for motivation is to offer a way to represent an agent's negotiation objectives, and to enable the developer to employ sophisticated means by which the agent can trade off various desires against one another to reach an acceptable outcome. Despite the desirable properties of taking a motivational approach to negotiation there is little, if any, work under way to explicitly address this issue.

2.5 Motivation and Neuroscience?

In investigating the notion of motivation, it may help to examine research in the neurosciences on motivation and emotion.

Whereas we have so far used motivation to refer to the process of influencing decision-making and providing the mechanism of goal generation, it is recognized in neuroscience that motivation is connected to the affectual or emotional states of an organism. As such, motivation can play a more varied role.

There has already been much work to understand the structures and pathways of the emotional and motivational subsystems of the brain and how these influence cognition. For example, Damasio's influential development of the 'somatic marker' theory argues that emotions and motivational arousal are crucial components in the system that enables an agent to choose between alternative courses of actions in a given context (Damasio, 1994).

The somatosensory cortex imbues incoming sensory information with a 'valence' – with this an agent can judge the desirability of the current situation. Furthermore, there exists an 'as if' loop which the agent can use to postulate future states and can assess them in terms of their valence. In the theory, the valence associated with a given situation is imprecise and fuzzy: it is primarily used to exclude choices, rather than enabling an agent to home in on the best choice.

The fundamental limitation in research on computational models of motivation from the perspective of an agent and artificial intelligence is that it is largely based on plausible but not necessarily valid models. Some work in the field of artificial life by contrast, where different disciplines come together, has attempted to use valid models based on neuroscience, ethology etc. However, this has largely been concerned with simple artificial creatures.

Research in neuroscience can inform the computational mechanisms being developed for artificial agents, both as a means of enhancing interactive experiences, and in achieving more effective behaviour. In particular, current models are limited and probably naive, suggesting limited applicability and functionality. The impact of motivation in reasoning and behavioural tasks is also largely unrecognized. Thus there is a strong potential for providing flexible mechanisms for guiding behaviour.

The picture emerging from neuroscience, by contrast, is that what we have been calling the motivational system is a diffuse, loose and dynamically interacting set of systems that enable an organism to recognize, assess, track and make decisions about world states that are important to it.

Research in motivated software agents has mainly focused on a simple homogeneous concept of motivation. Work on real, biological systems teaches us that the processes of motivation are multifaceted and heterogeneous.

For computational systems to exhibit the rich variety of responses we associate with biological organisms, they will probably need motivational and emotional mechanisms of a complexity approaching that of the real biological systems studied in the neurosciences. The objectives of neuroscience differ from those of computing. However, computing, especially agent-based computing, has gained from using psychological and biological models. Neuroscience can experimentally verify questions concerning computational models of motivation in relation to neural function. It can also uncover correlations between intelligent behaviour and corresponding neuronal activation. More generally, it can help us to understand more precisely the role of motition, its impact, and its value to computational intelligence and to effective agent behaviour.

3 AGENT PLANNING

Planning research has been an active area of artificial intelligence for more than three decades. Despite such a long history, the most dramatic advances have been in the past five or six years. Agent planning has undergone an important shift away from a focus on theoretical underpinnings towards an empirical emphasis on systems and potential applications. This has made the subject an exciting and relevant one, with a greatly improved opportunity to play a significant role in the development of intelligent autonomous systems. In this section we review planning research, the current state of the art and where the field is going. Finally, we speculate on the likely direction of the field in the next decade.

3.1 Overview

Planning is the problem of finding ways to achieve goals. A plan is a collection of instructions indicating what must be done, by what and to what, and when. A planner is actually only one component in the solution of many of these problems.

To solve these problems it is typically necessary to add a component that can abstract a representation of the problem from sensors that capture the situation in which the problem arises. Solving problems may also require an 'executive', a component that will achieve the execution of the plan. In addition, there might be need for intelligent fault diagnosis to determine what has gone wrong if a plan fails to execute. Some of these components can be humans: for example, if a planner is expected to plan deliveries, then the executive will be a collection of drivers under the direction of a manager. The sort of issues that planning addresses include:

- the efficient deployment of a collection of trucks and drivers to make a set of deliveries
- making best use of a collection of machine tools and operators to satisfy diverse orders for machined parts
- the management of the safe evacuation of friendly nationals from a politically unstable province
- the collection of as much scientific data as possible using an unmanned mobile laboratory on a distant planet
- the safe start-up of a chemical process plant.

Automatic planners have been or are being applied to all of these problems. The objective of planning research is to find efficient ways to generate plans for problems like these and to make those plans as cost-effective as possible. Planning researchers are concerned with the associated problems of making planning technology accessible, by making it easier to describe a problem to a planning system, and of better integrating planning systems with the related components such as sensors, executives and execution monitors.

3.2 AI Planning

The idea of AI, to construct some sort of surrogate human, is some way away from the reality. A more realistic concept of AI is as a collection of technologies that can tackle problems at which human problem-solving has proved effective but difficult to characterize. Among these are problems, such as image interpretation and robot control, that form the context in which planning can play a vital role in supplying an intelligent framework for autonomous action, particularly in application areas where human intervention is difficult, such as in deep space, deep sea or other hazardous environments. Planning is a technology for the future and, as we will see in this section, researchers in the field are rising to the challenge.

AI planning is concerned with constructing a programme of activity that will achieve a desired goal from the current situation. To do this, a planner has a description of the activities that may be enacted, the desired goal and the current state. The description of the activities must include information about the circumstances in which the activities may be executed and the consequences of applying them. Using the description of the consequences, the planner can then predict the effects of individual actions and, by composition, of sequences of actions. The job of the planner is to find efficient compositions of actions to achieve the goal.

Planning concerns constructing plans before they are executed. This depends upon the accurate predictive power of the planner's models. Planning is not appropriate where the relevant part is unpredictable or capricious.

Alternative technologies, such as reactive behaviour in which actions are selected and executed one-by-one, are more appropriate where situations can change unpredictably or where the effects of actions are highly uncertain. However, it is often possible to abstract models of the relevant part of the world and its behaviour to a point where uncertainties about the outcome of actions can be ignored.

For example, the action of driving from London to Cambridge is an abstraction of a complex chain of underlying actions that will include depressing accelerator, clutch and brake, changing gear, signalling, switching lanes and turning and so on, with several of these actions performed concurrently at various times. At the level of gear changes and use of accelerator, it is impossible to construct a plan – the world is simply unpredictable. At the abstracted level, however, the world is sufficiently predictable that a plan can be usefully constructed, say to drive to Cambridge, to take a room in a hotel, eat dinner and so on, to achieve a goal such as to attend a meeting in Cambridge first thing the following day.

During execution of this plan, the abstract actions will have to be interpreted in terms of the more primitive behaviours and these will often be executed using a more reactive style – gear changes, braking and accelerating are typically responses to road conditions executed within the framework of an overall objective.

In everyday usage, ‘planning’ often refers to organizing a collection of activities into an executable schedule. For example, project planning typically involves allocating time and personnel to complete known tasks efficiently. The AI research community describes this as problem ‘scheduling’. Researchers have traditionally considered this to be a separate problem from AI planning, which is concerned with selecting the actions that must be executed, rather than scheduling resources to those actions.

Many potential applications of planning technology require a system to solve both action selection and scheduling of resources, while it is often not possible to schedule activities efficiently without considering alternative activities that might achieve goals. This interaction has led the research communities in planning and scheduling to ever-closer integration. It is now common to find planning systems that carry out scheduling as part of their problem-solving.

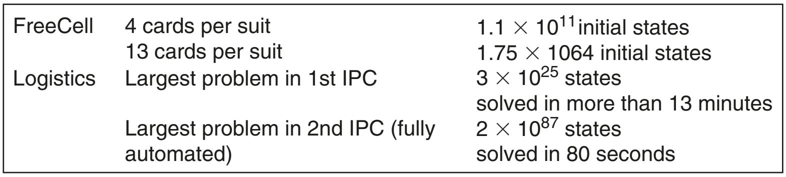
3.3 The State of the Art

Modelling for planners traditionally rests on a number of simplifying assumptions that define classical planning. First, it assumes that it is possible to predict completely and accurately the evolution of action sequences applied to a completely known initial situation, as though there were no external influences. Secondly, planning is the task of constructing a single completed plan that achieves the goal, before executing any part of it.

The formulation of classical planning also assumes that planners know the goals before planning starts – planners do not set their own goals and goals do not change as execution progresses. This makes classical planning a poor technology for realistic problems in which goals arise continually and important things can happen outside the control of the planner.

Finally, in classical planning, in which reasoning with numbers within the planning problem itself is excluded, the quality of the plan is determined solely by the number of actions in the plan. This is, of course, a simplistic measurement of a plan’s quality: recent work that relaxes this assumption, as well as some of the other restrictive assumptions of classical planning, is discussed below. Even under these simplifying assumptions, plan generation is computationally very hard, which explains its longstanding research interest.

The space of reachable configurations of the problem domain is exponential in the size of the problem description. The description is schematic, compressing the representation by generalizing actions to describe their effects on arbitrary objects. So the task



FreeCell	4 cards per suit	1.1×10^{11} initial states
	13 cards per suit	1.75×1064 initial states
Logistics	Largest problem in 1st IPC	3×10^{25} states solved in more than 13 minutes
	Largest problem in 2nd IPC (fully automated)	2×10^{87} states solved in 80 seconds

FIGURE 8.1 The table indicates the sizes of state-spaces for planning problems. FreeCell is a problem based on the well-known solitaire card game, introduced as a planning benchtest in the 2nd International Planning Competition (IPC). The logistics problem is commonly used as a benchtest for planning systems.

It is difficult to estimate the size of the state-space in all problems. For FreeCell the number of essentially distinct initial states gives an impression of the size of the state-space. In the logistics domain, it is easier to compute the number of different reachable states.

of a planning algorithm is to find a path between the initial situation and one satisfying the goal and to do this while exploring as little of the space as possible. This makes plan generation different from finding a shortest path in a graph – in this instance, the graph of states of the problem world linked by the actions allowing transitions between them. The graph is too big to be built explicitly (Fig. 8.1), so a plan generation algorithm must intelligently build only that part of it that contains the solution.

Intelligent exploration of a problem's state-space depends on the ability of the planning algorithm to exploit powerful heuristics or control knowledge to guide its search. The discovery of informative heuristics, which can be very effective in directing search towards a solution, had led to many recent strides forward in planning.

Planning research traditionally relied on simple and relatively unstructured models of the problem. This placed the research emphasis on developing algorithms and powerful heuristic control methods. Although it has been recognized that a model typically contains hidden structure that a planner can exploit, the tendency has been to persist with construction and communication of traditional models. This led researchers to supplement the traditional model with problem-specific control rules, or to use automated analysis to extract the hidden structure and make it accessible to the planner's reasoning mechanisms.

The recent successes of planners using problem-specific control rules raises the question of how much modelling can influence search efficiency. Other research communities in AI have focused on exploring the extent to which modelling choices can expedite the solution of a problem. Certainly, the more that human expertise is embedded in the model, the less discovery the solver has to make. This is true not only of planners, but also of AI systems more generally. However, the burden on the human expert can be prohibitive.

The correctness of the reasoning system depends on the correctness of the model, so modelling errors can be catastrophic. The traditional approach in planning has been to limit this burden as far as possible, providing a standard means for modelling action-centred behaviour and placing the problem-solving emphasis on automated techniques.

3.3.1 Recent Developments

Modern systems have seen steady relaxation of the simplifying assumptions of classical planning. The development of a richer expressive power than was used in 'first generation' planning systems to describe the behaviour of actions was considered from relatively early days and achieved in least commitment planners, as well as more recent systems such as IPP. This is based on the highly influential Graphplan planning system developed by Blum and Furst (1997).

Graphplan had a dramatic effect on the planning community. It produced vastly improved behaviour compared to the then current technology. Graphplan provided the foundations for several other planning systems and remains a powerful and influential tool in many current systems.

The Graphplan approach is to construct a data structure, called a 'plan graph'. In this graph, vertices are organized into successive layers, alternately representing facts and actions. A fact vertex represents a proposition that can be true after a number of distinct applications of an action, corresponding to the layer in which it occurs. An action vertex represents an action of which all preconditions could be satisfied by the preceding fact layer. Edges link fact vertices to the actions that achieve them. Edges also link actions to the precondition facts that they require.

As well as showing which facts are, in principle, achievable and at what stage in a plan they might be made true, the graph records an important additional detail. Where there are pairs of mutually exclusive facts in the same layer and where there are pairs of mutually exclusive actions in the same layer, each of these conditions is recorded using an edge linking the affected pairs.

A plan is found by searching the plan graph for a subgraph in which a single fact layer includes all the goal facts. For each fact included in the subgraph there is an achieving action, unless it is in the initial layer, which represents the initial state. For each action included, all of its preconditions are included. In the original Graphplan algorithm, searching was through an iterated depth-first search conducted backwards from the goal facts, which, if a plan exists, guarantees to find an optimal plan. This has proved expensive for some problems, however, and other search strategies have proved to be more effective.

The plan graph is a powerful way of encoding information about the states that can be reached by execution of actions from the initial state. It has proved extremely valuable in several other planning algorithms. One which has been particularly successful is the FF system, developed by Hoffmann and Nebel at the University of Freiberg.

In this system, a plan is found by forward search. A choice is made between possible actions leaving the current state and, once made, the choice is executed as a commitment. Backtracking is possible during certain phases of activity, but in general FF does not allow backtracking.

The important element of this strategy is to make a successful choice of action. To do this, FF heuristically estimates which of the states it can reach is closest to the goal state. The distance is estimated by counting the number of actions in a relaxed plan, constructed using a Graphplan search. A relaxed plan is one in which the actions are simplified by ignoring their negative effects. This eliminates harmful interactions between actions and means that the estimate of the work required to reach the goal takes into account only the positive work required to meet goals. The relaxed plan can be built very quickly, which is crucial to this strategy since it involves evaluating many states using the relaxed plan construction.

The strategy is remarkably effective in many problems, perhaps suggesting that complex interactions between actions is an artificial condition arising mainly in puzzle-like problems, rather than in more realistic domains. Nevertheless, some realistic interactions represent particular difficulties for relaxed plan estimates. Research continues into how best to handle these.

3.3.2 International Planning Competitions

A series of biennial international planning competitions (IPCs), initiated in 1998 by Drew McDermott of Yale University, has helped to encourage the shift of attention in research from an emphasis on theoretical results towards the development and empirical evaluation of planning systems. The competitions have so far consisted of a structured comparison of performance of competing systems on a collection of benchmark problems. These benchmark problems have become increasingly realistic and complex. It was an explicitly stated aim for the fourth competition, held in 2004, to use a collection of problems constructed in collaboration with potential users.

One stated goal of the third IPC was to encourage commitment to more sophisticated planning capabilities. The expressive power of the standard planning domain description language has been extended with the explicit intention of breaching some of the traditional restrictions of classical planning. In particular, domains used in the third IPC employ numerically measured resources and other numeric values. They also model the temporal structure of actions in the domains, including the duration of actions that can be executed.

The inclusion of actions with duration implicitly introduces the need for planners that can manage and exploit concurrency. This includes recognizing harmful interactions between concurrent activities even if they simply overlap, rather than synchronize their start or end points. Furthermore, plan metrics have been added to the language, so that it is possible to identify how to evaluate plans. This extension offers the power to harness planning for practical use in a way that is simply impossible if the only measure of a plan's quality is the number of steps it contains.

In most real planning applications the cost of resource consumption – including the time over which the plan is executed, possibly offset by the profit the plan generates – is an essential measure of plan quality, while the number of steps is of limited interest.

These extensions in expressive power made it possible to introduce into the third IPC several domains that make an interesting and convincing step towards real applications. The benchmarks now include a plausible model of logistics planning, a problem modelled on satellite observations planning and a further problem inspired by the need to develop autonomous planetary rovers for upcoming missions to Mars.

In the competition, planning systems competed with considerable success in these problems. Three or four fully automated systems performed convincingly on problems that included both resources and temporal complexity, including the scheduling of interacting concurrent activity. A subset of planning systems, using control knowledge constructed by hand for each planning domain separately to guide the planning process, showed remarkable performance on all problem domains.

It was particularly exciting to observe that (LPG), one of the best fully automated planning systems, produced better plans than those created using hand-coded control knowledge for a proportion of plans. To appreciate the significance of this, consider that the fully automated planners are not tailored to operate on any specific problem domain, but receive each new problem completely unprepared. The domain description contains no guidance on how to solve planning problems, but simply the fundamental description of possible actions and their effects within the domain.

The benefit of planning systems that use these problem descriptions is that constructing a domain description relies, in principle, only on a knowledge of the domain itself. This makes it possible to imagine providing planner software as a package for non-specialists, supported by appropriate tools, to allow them to explore the use of planners in a variety of contexts.

We can also construct plans containing hundreds of steps for complex domains. These plans can be produced on generic computing hardware in milliseconds or, for complex problems, a few minutes. Planners can attempt to optimize plans against metrics that are specified on a problem-by-problem basis. Fully automated planners can compete convincingly against planners exploiting carefully crafted control knowledge, built by experts, in planning, if not necessarily problem-solving in the specific domains.

3.3.3 Planning in a Wider Context

Planning research has also made inroads into handling uncertainty and in relaxing the assumption that a planner has a complete picture of the problem domain. Research has also explored the problem of linking planners to physical executive systems. This linkage has been an important barrier to application of planning systems in several interesting contexts. For a planner to be used within an intelligent autonomous system, it is necessary to solve infrastructural problems including the interpretation of sensor data such as visual data and the problem of executing abstract actions as concrete sequences of primitive instructions.

Developments in intelligent sensor interpretation, robotics and execution monitoring and failure diagnosis have all contributed to making the development of autonomous intelligent systems a realistic near-term goal.

Planning technology has been applied successfully to a wide variety of problems, as we have already seen. High profile success has been in the application of planning technology in space following the deployment of the RAX system on Deep-Space One, a technology test bed. An on-board planning system successfully controlled this autonomously for 24 mission hours in 1999.

NASA's planning research includes constructing plans for mobile robotic landers placed on Mars in 2003–2004. A more ambitious program is also being considered for 2009.

The European Space Agency (ESA) recently funded one of its biggest space ventures, the Aurora project, with goals that include a possible manned mission to Mars by 2030. This programme included early stages that involve robotic landers. Such an ambition can only be achieved with autonomous systems that can operate intelligently without human intervention over periods of hours or even days in space.

More earth-bound contexts in which intelligent autonomy could play a crucial role include deep sea exploration, disaster sites and other hostile environments. These situations are all examples of problems in which communication can be delayed and rapid responses could be called for on the part of an autonomous system. Planned responses, taking into account the objectives and longer-term activities of the executive, can allow coherent goal-directed behaviour to triumph over short term distractions.

A very different area of planning research is its application to support humans suffering from degeneration of their own planning capabilities, including the elderly and victims of Alzheimer's disease. Research funding on Alzheimer's in the USA recently injected a substantial sum into planning research for this purpose.

IBM recently announced an intention to invest in planning research as part of the Autonomic Computing initiative. NASA and JPL have significant research teams and a longstanding investment in planning and scheduling research. Planning has also attracted funding from the Defense Advanced Research Projects Agency (DARPA) in the United States, with a range of applications, from planning intelligent mission support at the squad level, through to scheduling naval repair and construction work for efficient use of resources.

Current applied planning systems are knowledge-intensive. They require significant technical input from domain experts and, perhaps primarily, from planning experts. To make planning technology more accessible and widely used, domain-independent systems offer a route past the bottleneck of planning system expertise. Domain-independent planning research has also made important contributions to the technology that is fielded in knowledge-intensive systems.

Robots as Executive Machines

The objective in constructing a plan is to find a sequence of steps that some executive can execute. One or several human or machine executives could manage the execution process. One target executive that has been considered is robotic systems.

We can consider robots to be executive machines that are capable of a reasonable repertoire of actions that can be described to a planning system and used as the basis of planning the activities of the robots. Several researchers have explored the integration of planners with robotic executives. Beetz and Simmons (Carnegie Mellon) are particularly significant researchers in this field. Nebel at the University of Freiberg has also explored the use of planning techniques in the context of the RoboCup soccer league.

There have been important developments in the construction of languages that form an intermediate layer between the plans, typically built from abstracted actions, and the execution layer, a real-time physical control system with closely linked feedback loops from sensors to actuators. Many questions remain to be resolved, particularly in understanding how to decide when an abstract action has failed in execution at the physical control level, and how such failure might affect on the plan from which the abstract action was drawn.

3.4 Planning: the Future

As we have seen, planning has moved on a long way from its early roots. Planners can now handle problems with time and numbers, allowing the expression of complex mixed scheduling and planning problems, demanding concurrency and resource management. In addition, the modelling language allows expression of plan metrics, so that planners can seek to optimize the plans they produce against a more useful measure of value than simple plan length. Use of this metric in the 2002 IPC confirmed that planners can tailor their performance towards production of plans of higher quality. Nevertheless, greater responsiveness to different plan metrics remains a high priority goal for research in fully automated planning systems.

Existing modelling standards do not support the modelling of exogenous events. Classical planning assumes that there can be no change that is not under the direct control of the planner. However, in many realistic situations it is necessary to plan around uncontrollable changes. For example, in planning full satellite observation, it is vital to represent the fact that opportunities for making observations, and for down-linking data, both arise in time windows that are not under the control of the executive and therefore cannot be planned by the planner. These opportunities arise as a consequence of orbiting the Earth. Some planners in the application-oriented tradition can plan with foreknowledge of such events.

When it is possible to anticipate such events, they can be encoded as constraints that can be mixed with all the other constraints that describe a problem. Any solution that emerges will then respect restrictions that these constraints impose. However, planning strategies that are not based on formulation of the problem as constraint sets have not yet successfully tackled planning with predictable external events.

External events that are not entirely predictable give rise to uncertainty in planning. Although many researchers have considered planning under uncertainty in various forms, there are still many questions. There is no clear consensus even on the form of a plan when managing uncertainty.

Uncertainty seems to arise in several different forms. Unpredictable external events give rise to a form that can be difficult to plan with. If unexpected changes occur frequently, the resulting uncertainty can undermine any planning effort. There are also more benign forms of uncertainty. For example, there is uncertainty about the precise duration of actions, and their resulting consumption of resources, such as fuel. This uncertainty typically can be described by continuous probability distributions, often normal or close to normal. This form of uncertainty might be considered benign in that allowing more resources for their execution can make plans increasingly robust to the uncertainty.

Uncertainty about the successful outcome of the execution of an action might be best described by a discrete probability distribution. It is harder to manage this form of uncertainty because it leads to a significant branching in the possible states of the problem world after execution. This uncertainty makes it hard to produce a robust plan without introducing contingent actions.

In addition to uncertainty, many real domains demand that a planner should manage complete ignorance. In this case, the executive will typically have access to some form of information-gathering actions. The planner can then plan to acquire information and react to the results. In this situation, and also in the case of handling uncertainty, it is often either impossible or a poor investment of effort to plan for long sequences of activity.

A more useful approach to problem-solving in this case is to interleave planning and execution, using execution monitoring, failure diagnosis and plan repair to resolve problems that arise during this process. Continuous planning, in which new goals can arise as a consequence of discoveries made at execution time, is an important development of planning, taking it further in the direction of autonomous behaviour.

These are all areas of active research. However, there is no commonly accepted empirical framework for evaluating such systems, or even for describing problems that the community could share. Putting problems such as these onto the agenda for the whole planning community is an important role for combinations of the competition series, the reporting of high-profile application areas demanding these kinds of functionality and the continual striving of the community as a whole to extend and develop the technology at its core.

3.5 Planning and Cognitive Science

As planning has become a less amorphous objective and has focused on more specific technical goals a number of projects have set out to integrate planning into a broader framework of problem-solving in a cognitively plausible framework. For example, the Soar project is intended to be a general cognitive architecture for developing systems that exhibit intelligent behaviour. Since its first use in 1983, Soar has involved a number of research groups and has evolved into what is best seen as an AI programming environment. It has supplied a collection of tools inspired by cognitive science that provide a framework for further development of cognitive systems. (http://ai.eecs.umich.edu/soar/). The TRAINS project, which dates from the late 1980s, is another example of planning integrated into a wider AI framework. Under the direction of James Allen, TRAINS has developed to explore the use of natural language dialogue in interaction with a planning system (http://www.cs.rochester.edu/research/trains/).

Development of this system required a sophisticated model of the structure of a plan in cognitive terms. This is necessary to interact usefully with a human agent in spoken and written natural language. The development prompted wide reaching investigations into the role of belief-desires-and-intentions within planning and in the understanding of human interactions about plans. Mixed-initiative planning – planning in which humans and planning systems cooperate in the development of a plan – has a vital role to play in many mission critical applications of planning, where we cannot expect human operators to trust the technology enough to make the transition to automated planning in a single step.

Projects such as these are less relevant to the main thrust of current planning research, with its emphasis on practical and concrete objectives. Nevertheless, research in planning and in cognitive and neurological sciences can usefully converge in some areas. For example, the work by Martha Pollack, of the University of Michigan, and others on the role of planning as a cognitive orthotic aid draws inspiration from the cognitive and neurosciences in understanding the role of planning in human interaction with the world. Key questions include the extent to which humans plan, rather than react or execute scripted behaviour, and the level of abstraction in human perception of actions when planning. These questions influence the degree to which AI planners can substitute for failed human capacity.

3.5.1 Learning

A historically important area of planning research, although one that has received less attention more recently, is the role of learning. The Prodigy system in particular is associated with this direction of research. Learning can play a useful role in several stages of planning. Researchers have considered several subjects here, including acquisition of operators and domain models, learning intelligent behaviour scripts to reduce future planning demands, learning intelligent planning strategies and learning strategies for plan repair. Cognitive science can play a valuable role in helping researchers to understand the role of learning in human problem solving.

Research in planning and cognitive sciences have been on diverging paths over recent years. However, developments in planning appear to offer the possibility of significant new insights into relationships between human and artificial planning and for these insights to contribute to the future of both disciplines.

3.5.2 *Cognitive Robotics*

An important area in which cognitive science and planning might interact significantly in the near future is in cognitive robotics, first developed by Ray Reiter at the University of Toronto. Much of the work in robotics has emphasized basic-level tasks such as sensory processing, path planning, manipulator design and control, and reactive agents. This is because many of the basic engineering problems involved in managing sensors, actuators and the tasks of navigation, self-location and environment interpretation were research issues that are only now finding stable solutions.

In contrast, cognitive robotics is concerned with the theory and the implementation of robots that reason, act and perceive in changing, incompletely known and unpredictable environments. Such robots must have higher level cognitive functions that involve reasoning about goals, actions, when to perceive and what to look for, the cognitive states of other agents, time and collaborative task execution; in other words, many of the problems that are directly or indirectly the province of planning research.

Cognitive robotics raises the level of research, in robotics, from the engineering and hardware levels (although these remain vitally important), to the level at which a robotic entity reasons about and interacts with a model of its environment in order to coordinate its activities into meaningful high-level plans.

The issues that arise in cognitive robotics cross many areas of AI research, with cognitive neuroscience offering input into understanding the nature of possible solutions at many levels. For example, cognitive models can provide a deeper understanding of the interfaces between abstracted plan-level processing and the reactive levels of motor control.

The gulf between low-level processing and supervisory-level strategic reasoning is still very wide, despite a number of efforts to define a common modelling paradigm. Shanahan's early work on the development of an event calculus-based planner proposed a way of traversing this gulf, but the difficulties involved in translating between symbolic concepts and raw sensor data still remain. These issues impact greatly on the ways in which planning problems should be modelled, how sensor readings should be interpreted and how plans should be translated into actions.

3.6 Open Questions

Major initiatives in deep space exploration, both in Europe and the United States, coupled with increasing interest in robotic systems, make it apparent that there will be a demand for intelligent autonomous systems over the next 5–10 years. Despite the inherent conservatism of space industries towards new software control, the difficulties of deep space communications make it inevitable that systems will have to be self-reliant for long periods.

To obtain the most from expensive missions with limited life, the systems must be capable of more interesting self-directed behaviour than simply retreating into a safe mode and awaiting further human instructions. Thus, a crucial challenge is to achieve full integration of planning, execution, execution monitoring and failure diagnosis. Resolving this issue does not depend on planning capability alone, of course. The integration of planning technologies with a wide range of products of other AI research will be an exciting challenge to the wider AI community.

Successful deployment of planning systems relies on an ability to model the problem domain. As the domains in which planning could apply become more complex, modelling becomes correspondingly challenging. As with most knowledge-intensive systems, encoding can be seen as analogous to software engineering, although the encoded information is less algorithmic and more declarative in style. Therefore, the planning community must confront the question of how best to support this process and to make it more accessible to non-specialists.

What tools are necessary to support domain modelling and to make planning technology more accessible to potential users? How much of the encoding process can be automated and what is the best way to achieve domain verification?

A separate question, concerning human–planner interaction, arises in the development of mixed-initiative systems. As planning technology becomes more practical for deployment into application, there could be a period during which planners must communicate with human mediators in constructing plans. This process demands a clear understanding of the nature of plans as humans understand them as well as the way that they are constructed and used by machines.

Developments of techniques for natural language interaction considered by Allen and the sophisticated interfaces explored by Tate will be important here. The process can involve a dialogue that will also require automatic recognition and understanding of plans, to allow humans to communicate plans to planners.

The greater expressive power offered by planning systems over recent years means that there is scope for far wider application. The planning community has already addressed the issue, but it continues to search for answers to issues of central concern, such as how to integrate the twin demands of planning and scheduling, particularly in resource-intensive contexts such as planning manufacturing processes, logistical operations and workflow management.

Another area of active research is a further extension of the expressive power of planning systems, to consider actions with non-deterministic or uncertain effects. Many domains introduce uncertainty about the effects of actions, perhaps because of the actions of other agents in the world, or because of inherent fragility of the actions of the executives themselves.

There is still disagreement about the best approaches to modelling and planning with uncertainty, including even the question of what constitutes a plan in the face of uncertainty. For example, one could seek a conformant plan, that will guarantee success regardless of the way that the world reacts, or a contingent plan, that provides opportunities to react to specific predicted failure points, or even a policy, a complete behaviour repertoire accounting for any possible observed state.

Planning has been limited, in the past, by its relatively narrow focus. The recent broadening of the definition of the planning problem has made it a relevant and core technology in many applications of cognitive systems.

3.7 Conclusion

Planning is a sub-field of AI research with a vital role to play in the development of intelligent autonomous systems and in many application areas of cognitive systems. The field has experienced a surge of energetic activity in which planning researchers in the UK play an important role. The United States sees planning research as a strategically important element in several high-profile research programmes, such as space systems, healthcare support and autonomic computing. European research is well-placed to pursue similar avenues, with recent funding in ESA's Aurora project, the collaborative efforts fostered under the EU funded Planning Network of Excellence (PLANET) and the placement of many key planning research teams in Europe.

4 AGENT INTERACTIONS

Interconnected computer systems are increasingly the norm. In such systems, people and machines interact with one another to achieve individual objectives and overarching ones. In this review, we have classified interactions into three types. The first is machine–machine interactions, including networking and information understanding. The second is people–machine interaction, often referred to as Human Computer Interaction (HCI). Finally, we consider how technology impacts on person-to-person communication.

4.1 Machine–Machine Interactions

An advanced area of computing where interactions are central is that of pervasive or ubiquitous computing. These terms are often used synonymously to describe the trend in computing towards smaller and more widespread devices (Weiser, 1993). The challenges range from hardware problems, such as miniaturization and power consumption, to software issues such as naming and communication with millions of devices.

The market drivers for this technology are obvious. People want access to resources – such as personal e-mail, company intranets, favourite websites – wherever they are, be it on the train, in the middle of a wood or halfway up a mountain. The assumption is that people want to operate as they would at a computer on their desk, even though they are mobile.

Traditional computer networks are static. Simple network protocols can find resources and devices, but these techniques don't operate well in a dynamic, mobile environment.

Recent developments in consumer products mean that a wider range of electronic devices can communicate with each other. Wireless networking is typically used for computer to computer communicating whilst technologies such as Bluetooth enable short-range communication between different devices such as mobile phones, digital cameras and printers.

Wireless networking and Bluetooth technologies require software layers to enable a user's equipment to 'discover' local devices and resources. To take a simple example, someone in an airport starts their laptop which negotiates high-speed access to the Internet. Once the device has established a connection, e-mail downloads as the user uploads images from their Bluetooth-enabled camera. The user's computer flags an important incoming e-mail message, discovers a local printer in the executive lounge and, using Bluetooth, prints out the attached document. All this can occur with little user intervention and no manual configuration – the user doesn't even have leave their seat.

One approach to device discovery is the JINI application technology. This enables resources, be they hardware or software, to connect in an ad hoc fashion without human intervention (http://www.jini.org). Based on Sun's Java language, JINI provides an infrastructure and programming model that is independent of any networking technology. It lets creators of devices and resources focus on the services they wish to provide and not be concerned with low-level issues such as resilience to network failure. Other activities such as Universal Plug-and-Play, uPnP, are aimed mainly at consumer devices, where the requirements are that there be no user configuration and low cost (http://upnp.org/focus).

Device discovery simply enables devices to establish who or what they wish to communicate with. It does not deal with how machines communicate or the data they use. With the increasing penetration of the Internet, its communication protocols (i.e. TCP/IP and UDP/IP) and infrastructures (i.e. DNS and routeing mechanisms) are becoming the de facto standard. Trends in pervasive and ubiquitous computing will massively increase the numbers of mobile networked devices. Initiatives such as Mobile Internet Protocol (Perkins, 1998) and IPv6 (Miller, 1998) are tackling the issues of network connectivity.

Research challenges for networking in such an environment focus on the ad hoc nature of devices forming opportunistic networks and how to establish correct (Royer and Toh, 1999) and efficient (Broch *et al.*, 1998) routeing. The activities of the World Wide Web Consortium, W3C (http://www.w3.org/) include creating standards for data communication between machines. This is centred on the Extended Mark-up Language (XML), a standard that defines how to annotate and format data, enabling machines to communicate complex structured data (http://www.w3.org/XML/). XML relies on the facilities of data encoding and data addressing, using Uniform Resource Identifiers (URIs). Common understanding for this data builds on Resource Definition Format (RDF) (http://www.w3.org/RDF/) and ontological work. Ontologies provide a conceptualization of a domain and facilitate knowledge sharing and reuse. They enable reasoning facilities to uncover information and translate between different information structures (Frank, 2002; Mitra and Wiederhold, 2001).

RDF allows us to specify relationships between entities such as a person belonging to an institution. These statements require a common understanding or ontology vocabulary and allow documents to be self-describing. ‘The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation’ (Berners-Lee, Hendler and Lassila, 2001). It builds on the ontology framework with layers to handle ‘Logic, Proof and Trust’ (see Fig. 8. 2).

The ‘logic’ layer enables the writing of rules. The ‘proof’ layer executes the rules and evaluates them together with the ‘trust’ layer. This provides a mechanism for applications to know whether to trust the given proof. For example, a rule in the logic layer may define that two researchers are colleagues if they belong to the same institution and are co-authors of a publication.

A proof would involve obtaining the resources necessary to establish whether two people are colleagues, for example a publications database and personnel database for the institution. The trust layer would establish how trustworthy those two sources of information are and what level of confidence

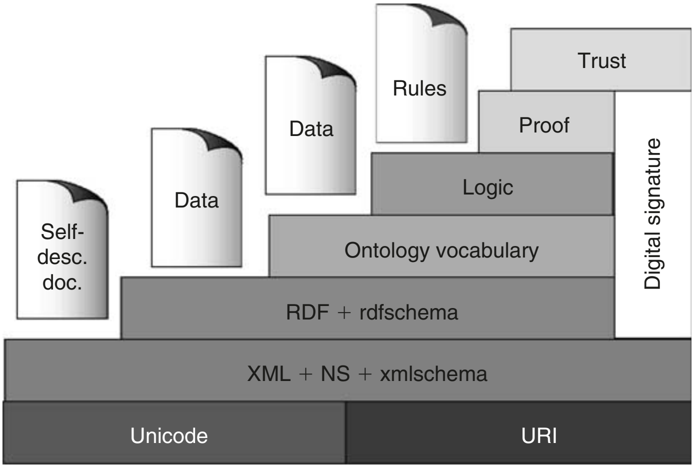


FIGURE 8.2 The Semantic Web layers. (This figure appears in the colour plate section)

to place on the resulting deduction. It relies on digital signatures to provide facilities to maintain the integrity and security of data. The addition of the level of inference engines and the layered approach form the cornerstone of the Semantic Web. The Semantic Web encourages machine–machine interaction in a complex system combining distributed processing and disparate information resources scattered across the Internet.

While the Semantic Web provides mechanisms for structuring knowledge, sharing this knowledge requires a communications infrastructure. Standards such as Simple Object Access Protocol (SOAP) provide lightweight protocols to support message communication in a structured manner (http://www.w3.org/TR/SOAP/). This allows services to be provided on the Internet by defining mechanisms for interaction and the language for communication. Services could range from direct information access, such as stock market quotes, to services that are more process-driven, such as negotiating the purchase of a product.

Efforts to tackle the discovery of these services focus on Universal Description, Discovery and Integration, UDDI (http://www.uddi.org/). This aims to provide a platform-independent, open framework for describing services, discovering businesses and integrating business services using the Internet. This resembles traditional business directories such as The Yellow Pages, often quite static and highly structured.

This activity is especially compelling in a mobile scenario, with a greatly reduced ability to access information. Devices and information resources are often available sporadically. The research challenges focus on more ad hoc and incidental interactions. For example, research on knowledgeable devices seeks to give users more utility, with small-scale wearable devices allowing knowledge to be obtained, processed and presented. This information capture is driven by the co-location of similar devices, in a meeting for example, where common interests between participants could provide cues to appropriate social interactions.

4.2 People-Machine Interactions

There are many aspects to the people–machine interface and human computer interaction (HCI). People typically interact directly with machines through input and output devices such as keyboards, mice and touch screens. Research in HCI tackles a range of issues. One trend is to reduce the cognitive overhead of interacting with machines, making best use of the devices. This might include speech recognition software, an interface for recognizing handwriting, and integrated smart boards. In all of these the focus is on more natural modes of interaction.

A second trend is towards novel interfaces that allow user experiences beyond what might normally be possible. This could be by immersing them in a digitally fabricated environment, through annotating reality systems, for example, that are virtual (Rheingold, 1991) and augmented (Rose *et al.*, 1995).

4.2.1 Speech Recognition

Speech recognition has long been a target for research. It is a natural form of communication and can be more convenient in a pervasive environment. To be useful, however, systems must be robust, with low error rates. They must also be able to cope with difficult acoustic situations and adapt to speech patterns that, even for a single individual, may vary according to anything from having a cold to their emotional state. Voice XML is one industry activity to standardize speech interfaces (http://www.voicexml.org).

Speech and handwriting recognition are direct forms of interaction. Human recognition often occurs in a more passive manner. Biometric recognition technology is becoming increasingly sophisticated and is currently a hot topic in the light of recent terrorist attacks. Biometric recognition systems attempt to automate processes that come naturally to humans, face or gesture recognition for example. Other biometric testing techniques with greater accuracy include fingerprint and retina scanning although these are not infallible. Alternative recognition techniques continue to be developed, with gait recognition proving successful (Yoo *et al.*, 2002).

The two problems associated with recognition systems are false-positives and false-negatives. False-positives are where the system identifies a person incorrectly, false-negatives where the person is not identified even though they are present. Much research aims to reduce false-positives and negatives. For example, many face recognition systems incorporate algorithms to prevent spoofing of the system using a still photograph of a face. Similarly, software for retina scanning often looks for evidence of blood flow, again to prevent the use of still images. These issues become important as we come to rely on recognition systems in high-risk areas such as airport security or identification in banking transactions.

4.2.2 Virtual Reality

Virtual reality (VR) has been around since the 1960s when Ivan Sutherland pioneered the use of head-mounted displays for visualizing information (Sutherland, 1965). Desktop VR is now commonplace with standards such as OpenGL and VRML (http://www.vrml.org/). Fully immersive VR include systems such as the CAVE, a room whose walls, ceiling and floor surround a viewer with projected images (Cruz-Neira *et al.*, 1992). As well as technological drivers, there is a growing body of work around the psychological effects of immersive virtual environments and alternative applications such as novel input devices for people with disabilities (Browning *et al.*, 1994).

The current trend in VR systems is towards large-scale, multi-user experiences. These range from Computer-Supported Collaborative Work (CSCW) (Dourish, 1996), to on-line gaming (Smed *et al.*, 2001). These systems require sophisticated caching strategies, detailed world modelling, increasingly high resolution graphics, including light modelling and swarming algorithms for simulating large numbers of independent entities.

Augmented Reality (AR) systems combine real world scenes and virtual scenes, augmenting the real world with additional information. This can be achieved by using tracked see-through head-mounted displays (HMD) and earphones. Rather than looking at a desktop or hand-held screen, hardware overlays visual information on objects in the real world. In a typical AR application, users wear a see-through HMD with a camera mounted onto it. The HMD shows video from the camera so that the real world is visible, with virtual graphics overlaid on it (Mitra and Wiederhold, 2001).

Tangible Augmented Reality is based on applying physical user interface techniques, where real objects are used as input and output devices, to augmented reality; for example, using augmented marker cards to pour virtual labels onto objects (Sinclair *et al.*, 2002) (see Fig. 8.3). This has resulted in systems that combine the intuitiveness of physical input devices with the enhanced display possibilities provided by virtual image overlays.

As we move toward a more pervasive computing environment, traditional computer interfaces, such as those based around desktop computing, become less appropriate. Researchers are looking at alternative interface metaphors to make interaction more natural. However, to do this they must overcome the ambiguities of free-form communication styles.

4.3 Person to Person Interactions

Aside from face-to-face contact, most person-to-person communication is mediated by technology. More advanced services are becoming available as mobile phones spread traditional technologies such as telephones. E-mail and text messages, such as the simple messaging service (SMS), allow users to communicate in a more asynchronous manner while on the move. The mobile phone is rapidly becoming the single point of contact.

Increased personal access brings new issues to personal communication. For example, people increasingly expect to be able to access e-mail and receive messages anywhere. They also expect to be able to contact others even if they are 'out of the office'.

Moving communication away from more recognized access points also causes problems

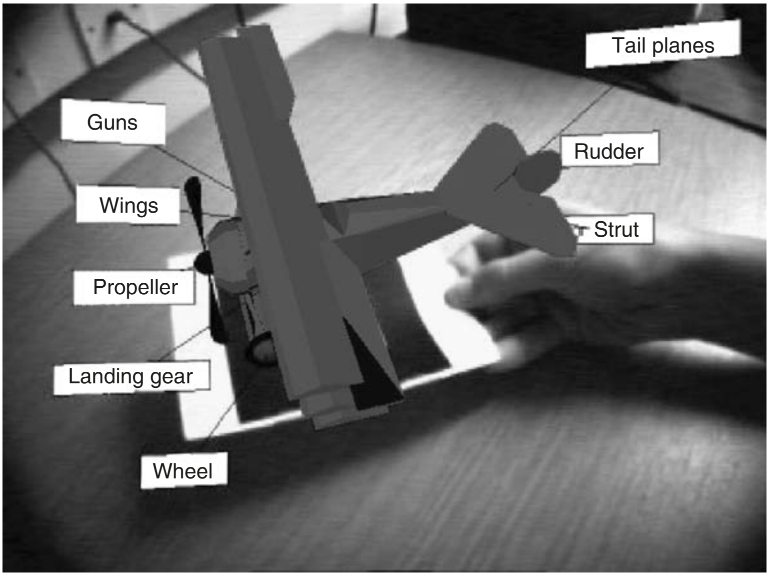


FIGURE 8.3 Exploring a tri-plane using a Tangible Augmented Reality interface. (This figure appears in the colour plate section)

of intrusiveness, where mobile phones can be seen to ring at inappropriate times, for example. Research is addressing this problem and is enabling devices to negotiate an appropriate level of intrusiveness with surrounding devices (Ramchurn and Jennings, 2002). For example, the cinema might request mobile phones to switch to silent alert, or meeting rooms could route messages to an appropriate, less intrusive, display mechanism.

Video conferencing technologies are increasingly deployed where people in different locations want to communicate. With 70% of human communication non-verbal, video conferencing provides advantages over simple telephony. Appropriate communication of non-verbal indicators such as body-language and eyeline are active research areas with the aim of improving the effectiveness of participant interaction (Vertegaal *et al.*, 2002). This is particularly important where there are multiple participants and additional information is needed to establish the social cues.

An example of a large-scale virtual community that exists over video conferencing is the 'access grid' used by researchers around the world (Foster and Kesselman, 1999). Each node on the access grid typically contains a number of cameras, microphones and large displays or projectors (see Fig. 8.4). During multi-site meetings, participants can view all camera streams. This often leads to visual confusion with no clearly presented focus.

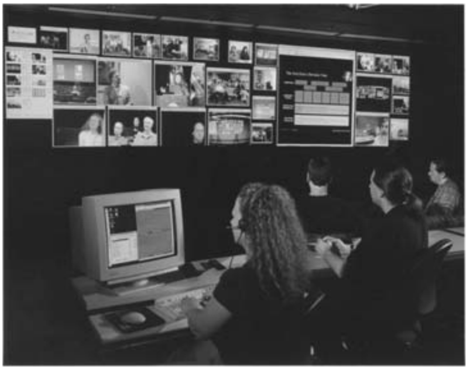


FIGURE 8.4 Researchers around the world video conferencing on the Access Grid. (Photo/logo courtesy of Argonne National Laboratory). (This figure appears in the colour plate section)

Coupled with streams of data, video conferencing can also provide an enriched meeting environment. Technologies such as digital whiteboards and shared applications allow Computer-Supported Collaborative Work (CSCW). Research in this area stretches beyond computer science into fields such as sociology and psychology.

4.4 Key Issues for Neuroscience

Technologies deployed on the Internet can involve many devices and entities. This poses problems in managing the scalability in an effective manner, a challenge with parallels in nature. Some research takes a more organic approach to failure and heterogeneity.

A model of how the brain manages the flows of information it receives would be beneficial in two ways. First, it would inform the design of computer systems that have to manage similar flows of information. Secondly, it could help in the design of interfaces to communicate information more effectively to humans. Both aspects are important when dealing with human attention, and for informing the designers of human computer interfaces on how best to limit intrusion, or capture attention.

Many problems encountered in HCI focus on issues of information overload or appropriate information modelling to make best use of people's ability to 'absorb' and 'retain' information. Humans are very good at filtering information, be it hearing their name mentioned in a noisy room or rapidly absorbing key information from many sources. In computer science terms, having captured information it then needs to be efficiently stored in a structured way that allows easy access and retrieval.

Functionally, there are clear parallels with the human brain. However, a number of properties make an interesting contrast to the traditional techniques of computer science. The organization of highly structured and interwoven knowledge is a seemingly innate function of the human brain. A key property of the brain that is missing from computer systems is the ability to function effectively with limited information and decaying capacity.

While there has been much research on modelling neural networks, higher-level semantic structures are more pertinent to issues of interaction and pervasive computing. Finally, humans are very effective at interacting with one another, even though they may never have met previously and may share very little in the way of common language. The processes involved in these human activities may shed light on how computer systems could achieve these tasks more flexibly, efficiently or robustly.

References

Andre, E. and Rist, T. (2000) Presenting through performing: on the use of multiple lifelike characters in knowlege-based presentation systems. *Proc. 2nd Int. Conf. Intelligent User Interfaces*, pp. 1–8.

Aylett, R.S. and Coddington, A.M. (2000) Agent-based continuous planning. *Proc. 19th Workshop UK Planning & Scheduling Special Interest Group.* See http://mcs.open.ac.uk/plansig 2000/

Balkenius, C. (1993) The roots of motivation. In H.L. Roitblat and S.W. Wilson (eds), *From Animals to Animats 2: Proc. 2nd Int. Conf. on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press/Bradford Books, p. 513.

Beaudoin, L.P. and Sloman, A. (1993) A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge and A. Ramsay (eds), *Prospects for Artificial Intelligence*, Amsterdam: IOS Press, pp. 229–238.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Sci. Am.*, 284(5), 34–43.

Billinghurst, M., Kato, H. and Poupyrev, I. (2001) *The MagicBook: Moving Seamlessly between Reality and Virtuality.* In IEEE Computer Graphics and Applications, May/June, pp. 2–4.

Blum, A. and Furst, M. (1997) Fast planning through planning graph analysis. *Artif. Intell.*, 90: 281–300.

Broch, J., Maltz, D.A., Johnson, D.B., Hu, Yih-Chun, and Jetcheva, J. (1998) A performance comparison of multihop wireless ad hoc network routing protocols. *Mobile Computing and Networking*, pp. 85–97. http://citeseer.nj.nec.com/broch98performance.html.

Browning, D., Cruz-Neira, C., Sandin, D.J., DeFanti, T.A. and Edel, J.G. (1994) Input interfacing to the CAVE by persons with disabilities. *Proc. 2nd Ann. Int. Conf. Virtual Reality and People with Disabilities*.

Coddington, A.M. (1998) Self-motivated planning in autonomous agents. *Proc. 17th Workshop UK Planning and Scheduling Special Interest Group*, pp. 49–59.

Coddington, A.M. (2001) Self-motivated planning in autonomous agents. PhD thesis, University of London.

Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V. and Hart, J.C. (1992) The cave: audio visual experience automatic virtual environment. *Commun. ACM*, 35 (6): 65–72.

Damasio, A.R. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Grosset/Putnam.

Dourish, P. (1996) Open implementation and flexibility in CSCW toolkits, PhD Thesis, University College, London.

Elliott, C. and Brzezinski, J. (1998) Autonomous agents as synthetic characters. *AI Magazine*, 19 (2): 13–30.

Foster, I. and Kesselman, C. (eds) (1999) *The Grid: Blueprint for a New Computing Infrastructure.* San Francisco, CA: Morgan Kaufmann.

Frank, M., Szekely, P., Neches, R., Yan, B. and Lopez, J. (2002) WebScripter: World-wide grassroots ontology translation via implicit end-user alignment. Eleventh International World Wide Web Conference (WWW2002).

Halliday, T. (1983) Motivation. In T.R. Halliday and P.J.B. Slater (eds), *Causes and Effects*. Oxford: Blackwell Scientific.

Halperin, J.R.P. (1991) Machine motivation. In J.A. Meyer and S.W. Wilson (eds), *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behaviour.* Cambridge, MA: MIT Press/Bradford Books.

Hinde, R.A. (1998) *Ethology: Its Nature and Relations with Other Sciences.* London: Fontana Press.

Jennings, N.R. (2001) An agent-based approach for building complex software systems. *Comms of the ACM*, 44 (4) 35–41.

Kunda, Z. (1990) The case for motivated reasoning. *Psychol. Bull.*, 108 (3): 486–498.

MIT Media Lab (1993) Tangible media group. http://tangible.www.media.mit.edu/groups/tangible.

Luck, M. (1993) Motivated inductive discovery. PhD thesis, University of London.

Luck, M. and d'Inverno, M. (1998) Motivated behaviour for goal adoption. In C. Zhang and D. Lukose (eds), *Multi-Agent Systems: Theories, Languages and Applications: Proc. 4th Australian Workshop on Distributed Artificial Intelligence*, volume 1544 of Lecture Notes in Artificial Intelligence. New York: Springer.

Maes, P. (1989a) The dynamics of action selection. In *Proc. 11th Int. Joint Conf. on Artificial Intelligence.* (IJCAI 89), vol. 2, Morgan Kaufmann, San Francisco, CA, pp. 991–998.

Maes, P. (1989b) How to do the right thing. *Connection Sci.*, 1 (3): 291–323.

Maes, P. (1991) A bottom-up mechanism for behaviour selection in an artificial creature. In J.A. Meyer and S.W. Wilson (eds), *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behaviour.* Cambridge, MA: MIT Press/Bradford Books, pp. 238–246.

Mayfield, J., Labrou, Y. and Finin, T. (1995) Intelligent Agents II. In *Evaluating KQML as an Agent Communication Language*. New York: Springer, 1995.

Miller, S. (1998) *IP vol. 6: The Next Generation Protocol.* Digital Press (Elsevier).

Mitra, P. and Wiederhold, G. (2001) An algebra for semantic interoperability of information sources. In *2nd Ann. IEEE Int. Symposium on Bioinformatics and Bioengineering*, pp. 174–182.

Moffat, D. and Frijda, N.H. (1995) Where there's a will there's an agent, in M. Wooldridge and N.R. Jennings (eds), *Intelligent Agents: Theories, Architectures, and Languages*, volume 890 of Lecture Notes in Artificial Intelligence. New York: Springer.

Moffat, D., Frijda, N.H. and Phaf, R.H. (1993) Analysis of a computer model of emotions. In A. Sloman, D. Hogg, G. Humphreys and A. Ramsay (eds), *Prospects for Artificial Intelligence*, Amsterdam: IOS Press, pp. 219–228.

Newell, A. (1982) The knowledge level. *Artif. Intell.*, 18: 87–127.

Norman, T.J. and Long, D. (1995a) Alarms: an implementation of motivated agency. In M.J. Wooldridge and N.R. Jennings (eds), *Intelligent Agents: Theories, Architectures, and Languages*, volume 890 of Lecture Notes in Artificial Intelligence. New York: Springer.

Norman, T.J. and Long, D. (1995b) Goal creation in motivated agents. In M.J. Wooldridge and N.R. Jennings (eds), *Intelligent Agents: Theories, Architectures, and Languages*, volume 890 of Lecture Notes in Artificial Intelligence. New York: Springer.

Perkins, C.E. (1998) *Mobile IP: Design Principles and Practices*. Reading, MA: Addison–Wesley.

Ramchurn, S.D. and Jennings, N.R. (2002) Multi-agent architecture for non-intrusive mechanisms. Technical Report Deliverables 3.1 and 3.2 of the IST FEEL Project, Department of Electronics and Computer Science, University of Southampton, June 2002.

Rheingold, H. (1991) *Virtual Reality.* New York: Touchstone.

Rose, E., Breen, D., Ahlers, K.H., Crampton, C., *et al*. (1995) Annotating real-world objects using augmented vision. Computer Graphics International '95, 1995. URL http://www.ecrc.de/research/uiandv/publications.html.

Royer, E. and Toh, C. (1999) A review of current routing protocols for ad-hoc mobile wireless networks. http://citeseer.nj.nec.com/royer99 review.html.

Schnepf, U. (1991) Robot ethology: A proposal for the research into intelligent autonomous systems, in J.A. Meyer and S.W. Wilson (eds), *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behaviour.* Cambridge, MA: MIT Press/Bradford Books.

Simon, H.A. (1979) Motivational and emotional controls of cognition. In *Models of Thought*. Harvard, CT: Yale University Press, pp. 29–38.

Sinclair, P.A.S., Martinez, K., Millard, D.E., Weal, M.J., (2002) Links in the palm of your hand: tangible hypermedia using augmented reality. *Proc. 13th ACM Conference on Hypertext and Hypermedia*, pp. 127–136.

Sloman, A. (1987) Motives, mechanisms, and emotions. *Cognition Emotion*, 1 (3): 217–233.

Sloman, A. and Croucher, M. (1981) Why robots will have emotions. *Proc. 7th Int. Joint Conf. on Artificial Intelligence*, Vancouver, 197–202.

Smed, J., Kaukoranta, T. and Hakonen, H. (2001) Aspects of networking in multiplayer computer games. In L. Sing, WanHak Man and W. Wai (eds), *Proc. Int. Conf. on Application and Development of Computer Games in the 21st Century*, Hong Kong SAR, China, November 2001, pp. 74–81.

Sutherland, I. (1965) The ultimate display. *Proc. IFIP Congr.*, pp. 506–508.

Vertegaal, R., Weevers, I. and Sohn, C. (2002) Gaze-2: an attentive video conferencing system. Conference Extended Abstracts on Human Factors in Computer Systems, pp. 736–737. New York: ACM Press.

Waltz, D.L. (1981) Eight principles for building an intelligent robot. In J.A. Meyer and S.W. Wilson (eds), *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behaviour.* Cambridge, MA: MIT Press/Bradford Books.

Weiser, M. (1993) Some computer science issues in ubiquitous computing. *Commun. ACM*, 36: 75–85.

Wooldridge, M. and Jennings, N.R. (1995) Intelligent agents: theory and practice. *Knowledge Eng. Rev.*, 10 (2): 115–152.

Yoo, J-H, Nixon, M.S. and Harris, C.J. (2002) Extracting gait signatures based on anatomical knowledge. In *BMVA Symposium on Advancing Biometric Technologies*. BMVA, The British Machine Vision Association and Society for Pattern Recognition, March 2002.

Source Material for Planning

The following websites are useful sources of information on planning research.

The European Planning Network of Excellence (PLANET): http://www.planet-noe.org/

The UK Planning and Scheduling Special Interest Group: http://www.research.salford.ac.uk/plansig/

Planning at JPL: http://www-aig.jpl.nasa.gov/public/planning/

Planning at NASA Ames: http://ic.arc.nasa.gov/ar.html

Information about the Remote Agent Experiment on Deep Space 1: http://ic.arc.nasa.gov/projects/remote-agent/

A planning resources page maintained by the American Association of AI: http://www.aaai.org/AITopics/html/planning.html

The third International Planning Competition: http://www.dur.ac.uk/d.p.long/competition.html

The Durham University Planning Group: http://www.dur.ac.uk/computer.science/research/stanstuff/

The Assisted Cognition research project at Washington University, applying planning technology to support victims of Alzheimer's and other cognitive-degenerative conditions: http://www.cs.washington.edu/assistcog/

The Soar Project: http://ai.eecs.umich.edu/soar/

The TRAINS project homepage: http://www.cs.rochester.edu/research/cisd/projects/trains/

SECTION

4

Memory

Natural and artificial cognitive systems can encode information, store it, reason with it and retrieve it later to guide behaviour.

Section Contents

This page intentionally left blank

Cognitive Systems Remember Experience: Introduction

Richard Morris and Lionel Tarassenko

A life without memory is unthinkable. Memory enables us to learn from experience, to 'travel in time' and to recall events from the past. It is, therefore, a continuing subject of research. Thanks to non-invasive neuroimaging techniques, we can now observe what goes on where in the brain during the processes of learning and memory. This and other research have established that there are different types of 'memory systems' within the brain. At the same time, we all benefit from and now even expect computers to take on more and more tasks of growing complexity. These include the ability of software tasks to learn from experience.

The next two chapters juxtapose life-science and physical-science approaches to learning and memory. Richard Morris, Graham Hitch, Kim Graham and Tim Bussey, in their contribution on Learning and Memory (Chapter 9), summarize the current state of the art in experimental psychology and neuroscience with respect to memory systems. They point out that instead of memory being a unitary faculty of mind and brain, it is a constellation of interacting, interdependent systems located in different brain areas and networks, and mediated by overlapping but distinctive neurobiological mechanisms. For example, the memory function that most people call 'memory' – the ability to recall events that happened in the past – is known as *episodic memory*. The brain's vast storehouse of factual knowledge about the world – concepts, the meanings of words and about people – is the *semantic memory* system. This knowledge may include the fact that pianos are musical instruments, that Paris is the capital of France, that DNA encodes genetic information in a sequence of base pairs, and so on.

Our capacity to acquire *cognitive and motor skills* is yet another system. Each of these is capable of holding onto information for long periods of time, a lifetime even, and so should be contrasted with the *working memory* system, which is optimized to hold a limited amount of information, with great fidelity, for short periods of time. Our working memory did not evolve to remember telephone numbers for the few seconds it takes to make a call, but it does the job well.

While there are clear points of contact, these distinctions do not obviously map in any direct way onto the ideas that have proved influential in the development of machine learning. Kieron O'Hara, Wendy Hall, Keith van Rijsbergen and Nigel Shadbolt, in Memory, Reasoning and Learning (Chapter 10) describe how computer scientists have benefited from Moore's Law and the ever-cheaper costs of physical memory devices to develop software systems that can very quickly process vast amounts of data, and both store it and reason with it for later use.

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

Clearly, neither the ever-faster speed of silicon processing chips nor the costs of memory devices are constraints that influence the direction of life-sciences research, yet these are precisely what has made it possible to have impressive hand-held devices that we all now take for granted. However, there is clear conceptual overlap in ideas such as those of *associative memory* and *content addressability* in neural networks, ideas that have been influential in both domains. Computer scientists also have a distinctive focus on how memory and reasoning can provide a basis for the development of new software.

In recent years, *brain imaging* has added an anatomical dimension to the information on memory that life scientists have arrived at through psychological studies. By measuring local changes in blood flow in the brain, functional brain imaging allows us to look at the operation of different areas of the brain as subjects perform particular tasks. In parallel with this, another major source of information on memory comes through studying people who have suffered some sort of brain damage, giving insight into what functions of the brain are compromised and what unaffected by such damage. Observations on experimental animals provide more direct measures of the activity of brain cells, their synaptic connections and their signal transduction pathways that are involved in information processing.

We are now in a position to supplement the early neural network models developed by physical scientists with more detailed knowledge of biological systems at the network, local circuit, cell, synapse and molecular-genetic levels.

Interestingly, both brain imaging itself and *neural network models* are benefiting enormously from sophisticated mathematical computing tools, and creating intriguing neuroinformatics issues such as those articulated by the Human Brain Project (http://www.nimh.nih.gov/neuroinformatics/index.cfm) and the OECD Neuroinformatics project (http://www.neuroinf.de/)

One intriguing debate to emerge in the course of discussion about these different approaches to learning and memory concerned *forgetting*. How do we cast off, deliberately or by accident, previously remembered information? And should we try to compensate for the problems that this sometimes creates?

Life scientists assert that forgetting is more than just a failure of memory. It allows people to 'clean out' obsolete and useless data such as memories of what they ate for breakfast several days ago. Forgetting allows us to focus our attention and mental effort, and so be able to search memories only for the most relevant information. If memory is for living, then forgetting is part of what makes that manageable.

In contrast, physical scientists are close to having at their disposal digital recording and computing systems that may make it possible to record virtually all the events of a person's life. They could also have such records – memories for life – accessible at a reasonable cost. For them, forgetting is an irritation, a failure, and something that could and should be superseded by new technology. It will be intriguing to see how the dialogue continues from these very different premises and what emerges to help people to cope with the increasing cognitive demands of the information age.

CHAPTER

9

Learning and Memory

Richard Morris, Graham Hitch, Kim Graham and Tim Bussey

1 INTRODUCTORY SUMMARY

Memory is central to human individuality. Our ability to change our behaviour in response to experience, to acquire knowledge and to travel mentally in time to recall events from the past brings joy, laughter and tears. All depend on memory. A life without memory is unthinkable.

Our central theme is that there are many types of memory. We consider a major distinction between short-term and long-term

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

memory in relation to research that has identified distinct components of working-memory and of long-term memory.

We outline the characteristics, organization and further subdivision of these different forms of memory from both a psychological and a neuroanatomical perspective. For example, some types of long-term memory are explicit, with memories later retrieved into conscious awareness: other types are implicit, such that the initial encoding of information and the processes by which it is retrieved are unavailable to consciousness.

Insight into the processes and mechanisms of learning and memory has emerged from the simplest of psychological protocols through to the most modern brain imaging techniques. One way of securing insights about function is to study dysfunction and there is a fine tradition in the UK of studying neurological patients with highly specific patterns of brain damage. These have allowed fractionations to be identified between ostensibly independent brain systems of memory.

Human brain imaging provides new information that, in part, justifies the anatomically discrete nature of these brain systems. However, comparing data obtained from this technique with that from more traditional methods has yielded surprises.

There has been longstanding concern among researchers studying patients about the validity of arguments for localization of function. This is now being echoed in the brain-imaging community as it maps out networks of interconnecting brain areas rather than ostensibly localized functional centres for memory.

Computational modelling provides insightful theoretical links between these experimental approaches. The importance of guiding research on learning and memory by theory increases as empirical techniques diversify.

Work on human subjects is complemented by studies on animals that enable the investigation of neurobiological mechanisms at levels of analysis that are not (yet) possible with humans. Certain invasive procedures, such as recording from single-cells in the brain or applying drugs to discrete brain areas, are only feasible with animals.

Emerging opportunities to record from or optically monitor large numbers of cells simultaneously are bringing new challenges, some of them computational. Neurobiological approaches also lend themselves to realistic neural network modelling.

Certain neurological disorders are degenerative. Several of these disorders, such as Alzheimer's disease, attack memory function as one of their first targets. Neuropsychologists are contributing to the development of tests for accurate diagnosis, characterizing the cognitive changes that take place, and are making inroads into the task of providing remedial assistance. Aspects of this work are pragmatic, while others are building upon the foundation of the concept of multiple memory systems.

2 DEFINITIONS, CONCEPTS AND TECHNIQUES

2.1 Definitions and Concepts

We use the words *learning* and *memory* routinely in ordinary discourse but they are also scientific concepts, defined formally by psychologists and neuroscientists. Lay usage of the term 'learning' is generally restricted to situations where there is some element of deliberation or intent – such as in learning a language or learning to drive. One would not, for example, ordinarily learn what one had for breakfast. In contrast, memory tends to be used most frequently in reference to the recall of events that, at the time they happen, we do not deliberately memorize – as in remembering what happened last Christmas.

In contrast, formal psychological definitions of these terms do not entail any reference to intent. Learning is generally defined as 'the act of acquiring information or skill such that knowledge and/or behaviour change'. It may occur in a variety of different ways.

Memory is defined in at least two ways. It is used to refer to a presumed 'mental storage device' in which information may be held, as in the concept of a phonological store. Additionally, it is used to refer to a putative 'capacity of mind', as in the concept of episodic memory. Psychologists recognize different types of memory, distinguished in relation to the types of information they process (e.g. words *vs* pictorial information), their capacity or persistence (e.g. short-term *vs* long-term), and their operating characteristics (e.g. the mental codes in which information is held).

Definitions of learning and memory in neurobiology bring in such factors as the neuroanatomical localization of a putative system, or the physiological and cell-biological mechanisms involved. In stepping along this path, neuroscientists recognize that learning is an act of acquiring information, as psychologists define it, but also assert that it is a process that is thought to engage specific areas of the brain, to depend on specific patterns of neural activity and, importantly, to result in biological changes in brain cells that outlast the learning experience.

Similarly, the term memory is also widely used alongside specific networks in the brain, such as a group of structures or set of neuronal connections that is thought to carry out memory functions. The 'medial temporal lobe memory system' is such a concept (Squire, 1992), as is the idea of amygdala-dependent memory (LeDoux, 2000).

In his pioneering book on computational aspects of vision, Marr (1982) distinguished what he referred to as computational, algorithmic and implementation levels of analysis in information processing science. This tripartite distinction has been useful, but it does not map very directly onto the numerous levels of analysis at which individual neuroscientists operate.

Contemporary approaches to learning and memory are concerned with linking these levels of analysis, but this is far from easy, largely because most neuroscientists find themselves at the limit of their understanding when they stray outside the disciplines in which they were trained. Few seem to realize the complexity of developing a 'general theory of memory' that would link the many levels at which it can be analysed.

Whereas neuropsychologists are interested in understanding the mapping of psychological process onto neuroanatomical structures and networks, using patients and functional magnetic resonance imaging, fMRI (see Chapter 4 'Advanced neuroscience technologies'), their eyes tend to glaze over when attention turns to the chemical pathways that mediate biological changes at the cell level within neurons. Conversely, whereas the 'autophosphorylation of the alpha sub-unit of calcium-calmodulin kinase within the postsynaptic density of glutamatergic neurons' is the stuff of coffee-room debate in hard-core neuroscience departments, the relevance of this and other biochemical mechanisms to explicit or implicit memory might not grab the same level of their attention (Fig. 9.1).

We have begun by contrasting the approach adopted by experimental psychologists with that of neurobiologists. One meeting point of these cultures has to do with the fundamental property of memory. This is that memory is of necessity a change in the brain that outlasts the stimuli that trigger it. The change may be entry into an active state of reverberation amongst a network of neurons or a physical change in neurons thought to mediate more lasting memories. This brings us to the concept of a 'memory trace' – the physical 'substrate of storage' (Hebb, 1949).

Changes at synapses – the connection points between neurons – are currently the favoured locations for storage of long-term memory traces. Much current research focuses on how synapses change in strength (Martin *et al.*, 2000). Indeed, synaptic neurophysiological researchers often describe plasticity as a model of memory, with its neural mechanisms the focus of interest, including the autophosphorylation of αCAMKII. Study of these physiological mechanisms, together with computational

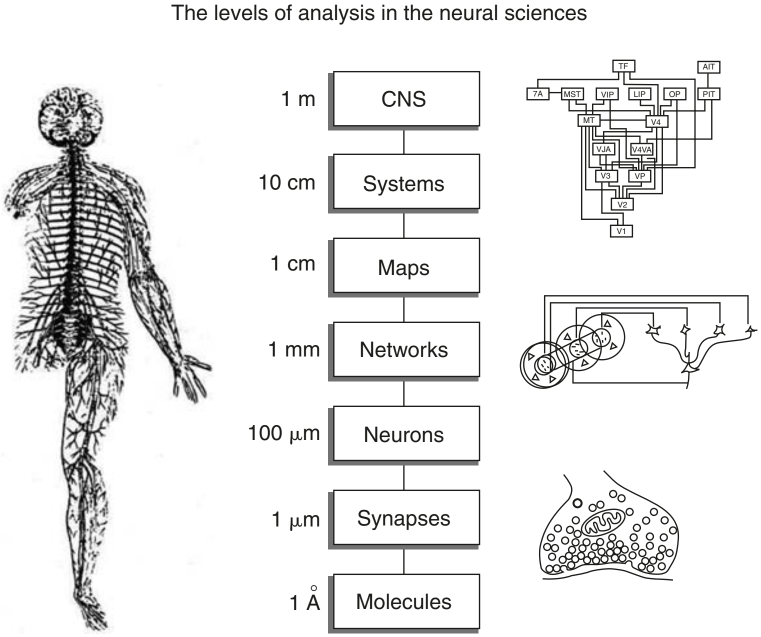


FIGURE 9.1 Issues can be tackled at the level of the whole person, the anatomical brain area, local circuits, cells, synapses, or yet at the level of molecules and genes. A neurobiological theory of memory would be one that successfully integrated information across levels – where this would be fruitful and illuminating (not always). (Reproduced with permission from Churchland and Sejnowski, 1992)

modelling, has revealed the possible existence of many different learning rules that could determine whether a trace is stored and how it represents information within various kinds of neural network (Rumelhart and McClelland, 1986; O'Reilly, 1998). Patterns of neural activity serve as memory cues and reactivate traces later. The resulting output is what psychologists and neuroscientists agree as being memory.

Another meeting point for experimental psychologists and neurobiologists is that we all recognize that we can subdivide learning and memory into distinct temporal phases or processes – encoding, storage, consolidation and retrieval. Encoding has to do with the formation of memories – what must happen for a memory to form in the first place. Storage has to do with what lasts in the mind or brain, with different kinds of storage device mediating short-term and long-term memory. Retrieval refers to the process of memory reactivation. The concept of consolidation refers to something that happens to memory traces after they have been stored and that alters their persistence or sensitivity to brain damage (McGaugh, 2000). This 'something' is not the same as memory retrieval *per se*, although one view of consolidation is that it entails repeated acts of retrieval and re-storage that may even happen during sleep.

2.2 Techniques

Researchers use a wide variety of techniques to study the workings of the brain. While major discoveries have often emerged from the simplest 'paper and pencil' techniques, these are gradually giving way to experimental protocols using PCs and even virtual reality.

Typically, researchers engage human subjects in an experiment. They explain some but not all aspects of its purpose to the subjects – deception is sometimes essential, with participants always de-briefed later. The researchers then conduct the test. Most experiments involve a limited number of subjects, as required for the purposes of statistical analysis. Questionnaire studies, of flashbulb memories for example (e.g. Conway *et al.*, 1994), can involve hundreds of participants; while work with neurological patients can be through carefully conducted single-case studies.

The neuropsychological study of patients has been the classical route for understanding the neurological basis of memory, but logical pitfalls strew the path.

Following a stroke, brain tumours or viral infections, such as herpes encephalitis, patients often show remarkably circumscribed deficits in cognitive function, including memory. When the brain damage is very circumscribed, it is natural to suppose that a site has been localized where the lost function used to be. However, following an amusing critique, several disquiets about this logical deduction have appeared (Gregory, 1961) The favoured view now is that studies of lesions establish only that the integrity of a particular structure is essential for the normal operation of an identified dysfunction. Double-dissociations are particularly instructive because they reveal that damage in area X of the brain affects memory process A but not B, whereas damage in brain area Y affects memory process B but not A. Experiments of this kind are the gold-standard, but even their apparently straightforward interpretation is not without pitfalls (Shallice, 1988).

> 'It is not so much the injury that captures our attention as how, through injury or disease, normal function is laid bare.' (Sir Henry Head, twentieth-century neurologist)

Non-invasive techniques of human brain imaging, such as PET and fMRI, provide a way of visualizing differential patterns of activation of the nervous system by monitoring haemodynamic signals – such as the BOLD signal in fMRI (Frackowiak *et al.*, 1997). Extensive experience of human brain imaging has revealed a pattern of statistically consistent effects in a range of memory tasks.

The analysis protocol developed in London by Friston and colleagues, statistical parametric mapping (SPM), is now used worldwide. The brain imaging approach comes with the built-in advantage that the subjects are generally normal rather than brain-damaged. However, the interpretation of functional brain imaging studies is also not easy, indeed something of an academic industry in its own right. There is concern about signal size, a wide range of susceptibility artefacts, repeatability and other issues.

The marriage of findings between studies of patients and brain imaging studies has been particularly problematic in work on long-term memory. However, divorce has been avoided. The rejuvenation of cohabitation has come about by virtue of the fact that brain imaging is opening up new experimental protocols. These include online event-related monitoring, and new and subtle distinctions between different types of memory process that were not dreamed of in the era of 'patients only' neuropsychology – for example, that between retrieval effort and retrieval success.

Work with laboratory animals plays a key role in studies of the neurobiology of memory. While there remains a division of opinion, most scientists accept that invasive experiments on animals are justifiable. They enable recording of neuronal activity, placing of experimental lesions, giving novel drugs or, more recently, engineering transgenic alterations in their genetic make-up (to examine the molecular basis of brain mechanisms). For example, Riedel *et al.* (1999) have developed a technique to switch off the hippocampus of a rat for a week and then turn it on again.

The typical infrastructure of an 'animal laboratory' is very different from that using human subjects – with facilities for keeping the animals to the high standards now rightly expected by the Home Office, and a wide range of high-technology equipment for electrophysiology, drug monitoring and microscopy. Not all work with brain tissue involves work on living animals. Experiments on *in vitro* brain slices and organotypic cell cultures use tissue derived from animals, but the procedures and level of analysis are far removed from that of the typical memory experiment in a human or *in vivo* animal experiment.

The way forward to integrate these different levels of analysis is via computational modelling. Constructing a formal computational model can be useful in forcing hidden assumptions to be made explicit, making novel predictions that empirical researchers can test and generating new insights. It is particularly useful when there is plenty of data with which to constrain model construction, and where it would otherwise be difficult to use current knowledge to make testable predictions, for example in predicting complex interactions.

Connectionist techniques, beginning with work on parallel distribution processing (PDP, Rumelhart and McClelland, 1986), have the particular appeal of relatively high biological plausibility when compared with other methods. They simulate processes of learning, storage and retrieval through changes in connection strengths in artificial neural networks and are proving particularly fruitful in memory research.

Such models have the potential for generating 'emergent features'. For example, a model might produce interesting behaviour that was not explicitly built-in by the modeller, such as the effects of simulating neural damage by making artificial lesions. Alternatively, a model might generate a novel insight into a well-known phenomenon.

An intriguing example of the latter is the so-called age of acquisition effect, whereby words acquired early in life are typically processed more efficiently than words learned later on. Examining relatively simple models has shown how early inputs to a neural network have a disproportionately large effect on the organization of its connections and make it clear why these effects are not overcome by learning subsequent inputs (Ellis and Lambon-Ralph, 2000).

2.3 Open Questions

- *Are learning and memory really distinct from perception, attention or motivation?*

Traditional definitions of learning and memory distinguish these sharply from other cognitive concepts such as perception, attention or motivation. However, while one would not expect matters of definition to be open questions in a topic that has been investigated for over a century (James, 1890), recent neuropsychological research has begun to call into question the sharp distinction between perception and memory.

There is a fundamental logical distinction between perception and memory. One may perceive some stimulus and act upon it – such as seeing a red stoplight at a road-junction – but acting appropriately does not *require* a memory of it to be formed. However, perception and memory cannot be independent.

First, the possession of memory about what a stoplight means is essential for acting appropriately. Second, the moments when people merely perceive cannot be separated from those when they are encoding new information. Third, the relevant brain areas overlap, particularly in the dorsal and ventral processing streams of the posterior neocortex.

Many acts of perception and action do result in some 'automatic recording of experience' (Morris and Frey, 1997). Asked a short while after the road junction, our driver should be able to tell you whether the lights just before had been red or green. In the absence of any need to do so, a memory has been formed, though it generally does not last.

Attention is particularly important in this context. It serves as a gate, determining which of the many aspects of sensory processing should command cognitive resources. Novelty and emotional significance can also be important. Both trigger neuromodulatory transmitter systems in the brain that play a critical role in determining whether the trivia of everyday life are remembered persistently (e.g. 'flashbulb memories' such as remembering exactly where you were and what you were doing on 9/11).

Motivation is irrelevant to automatic recording, but effort is vital for the more deliberate acts of learning and memory – such as what we learn at school, learning a musical instrument, or acquiring a motor skill. Learning difficult concepts involves retrieving what we already know and then interleaving new information.

- *Memory in the digital age?*

In the pre-Gutenberg age, the possession of a good memory was highly valued (Yates, 1966). The past 500 years has been the analogue information era in which books, newspapers, radio and television have provided us with all manner of information to learn about, to remember and to share with others.

As we enter the digital information age, we are not only confronted with an information explosion, in which forgetting may soon become as valued a cognitive capacity as remembering, but also a rich variety of software tools for finding things out. Will the need for an endogenous memory become a thing of the past?

We live in a world with more information at our fingertips (literally) than most of us can deal with. A world of laptop computers, PDAs, picture messaging mobile phones and MP3 players. If every home were to have an always-on broadband connection with ever-faster speeds of access to the World Wide Web, what would be the point of remembering anything? Why bother to remember if the Google search engine or its successors can find it within seconds? The answer to this puzzle begins in the next section, where we introduce the concept of multiple memory systems. To anticipate, a search engine is useless for certain memory tasks. There are also biological memory processes that might lead to better search engines.

3 THE ORGANIZATION OF MEMORY SYSTEMS IN THE BRAIN

There is no one memory system of the mind or brain, no single device or brain area to which all the information we ever learn is shuttled for storage. Instead, the brain has a variety of memory systems that serve different purposes. This is the key concept of 'multiple memory systems'. Some systems hold limited amounts of information for a short time in an active, conscious state. These systems are contained within 'working-memory'. The much larger storehouse of information is called 'long-term memory'. It is also sub-divided into distinct systems.

Various taxonomies of memory have been published. These divide memory first with respect to persistence – short- *vs* long-term memory – and then with respect to the character of mental experience. Within long-term memory, it is common to speak of a distinction between 'explicit' and 'implicit' memory, where the former is held to require certain aspects of consciousness that the latter does not.

We remember and can recover conscious awareness of the people with whom we spent last Christmas – a type of explicit memory – but may be quite unaware of how we make a particular stroke in a game of tennis – it just happens. For this reason, event memory is said to be explicit while motor skills are often classified as implicit. Psychologists make other increasingly subtle distinctions to distinguish one form of memory from another – autobiographical versus public memories, prospective versus retrospective and so on. Whether these binary distinctions are merely descriptive, or really tapping into theoretically fundamental distinctions, is a matter of debate.

3.1 Working Memory

Like a pad on a desk for jotting down names or telephone numbers that we briefly need to remember, the brain has a 'working-memory' for accurately holding onto and manipulating small amounts of temporary information (Baddeley, 1992). We use working memory in a wide variety of contexts.

To give just some idea of the range, these include doing two things at once (e.g. talking while driving), performing a complex cognitive skill (e.g. mental arithmetic), planning and problem-solving (e.g. organizing a shopping trip), keeping track of novel information (e.g. learning how to say a new foreign word or remembering where one put one's keys). Fidelity is central to the system – a feature that is probably bought at the cost of both limited capacity and limited persistence over time.

3.1.1 Psychological Characteristics

Working memory is widely thought to consist of a set of limited capacity subsystems, as in current conceptions of the original Baddeley and Hitch (1974) model (see Fig. 9.2). The central executive is assumed to control the overall operation of working memory and its interactions with long-term memory and systems for perception and action. The slave stores are more restricted and are principally involved in maintaining small amounts of either verbal or visuospatial information over short intervals.

The multi-component view explains empirical dissociations between different types of immediate recall task that tap the various subsystems selectively. Thus, verbal short-term memory span is the longest sequence of random digits, letters or words that can be recalled from a single presentation, and is taken to reflect the capacity of the 'phonological loop'.

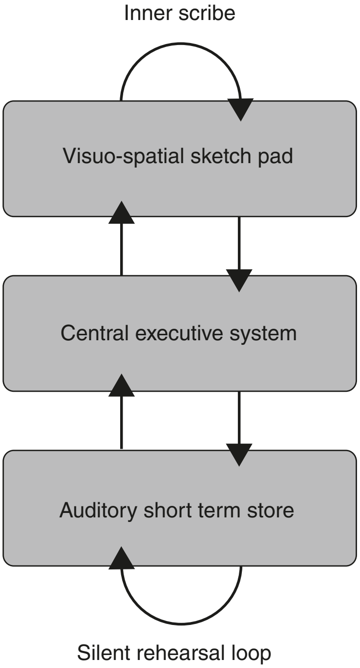


FIGURE 9.2 An attentional central executive interacts with two slave stores, one specialized for verbal-spoken information – the phonological loop (consisting of an auditory short-term store and a silent rehearsal loop). The other slave store is for visuo-spatial information (the visuo-spatial sketch-pad). (Based on Baddeley, 1992).

The multi-component Corsi Blocks Span is an analogous measure for spatio-temporal sequences and taken to reflect the operation of the 'visuo-spatial sketchpad' and its inner scribe. Working memory span measures the ability to manipulate as well as store temporary information and is assumed to require the central executive. For example, Reading Span assesses how many sentences a person can read and understand whilst keeping track of the last word in each sentence.

The evidence for dissociations among these three types of task comes from a diverse range of sources. One line comes from correlational analyses of they the way they cluster in studies of individual differences (e.g., Engle *et al.*, 1999). Another source is tasks requiring a person to do two things at once, 'dual-task studies', where activities that tap the same component of working memory show more mutual interference than activities that tap different ones (e.g., Smyth *et al.*, 1988).

Yet another type of evidence is gathered from special populations. For example, the genetic disorders of Down and Williams syndromes show a double dissociation with impaired digit span and relative sparing of Corsi span in Down and the converse in Williams (Jarrold *et al.*, 1999).

The same tasks are also doubly dissociated in adult neurological patients. The syndrome of conduction aphasia is associated with reduced auditory-verbal digit span combined with normal visuo-spatial working memory (Vallar and Shallice, 1990), whereas other patients show selective impairment of Corsi Blocks span with preserved auditory-verbal digit span (Hanley *et al.*, 1991).

There is also a large group of neurological patients with frontal lesions that typically display various problems of attentional control, such as difficulty in planning, switching and dividing attention, or inhibiting inappropriate behaviour (Shallice, 1988). These difficulties can occur even when there are no deficits in short-term storage and have been interpreted as reflecting impaired executive function. Convergence with the multi-component model is not precise, however, as tasks sensitive to frontal impairment tend not to intercorrelate very highly.

It is important to note that the multi-component model of working memory is deliberately simplistic. It serves primarily as a shorthand for linking diverse phenomena, for making predictions and asking new questions. While the model continues to be widely used, there are alternative theoretical approaches (see, e.g., Miyake *et al.*, 2000). Moreover, the model itself is progressively being elaborated as new findings emerge and points of weakness are tackled (Baddeley, 2000).

3.1.2 Cerebral Localization

To a first approximation, evidence from neurological patients with specific lesion sites and from areas of activation in functional neuroimaging studies of healthy individuals is consistent with the multi-component model of working memory. Thus, patients with selective deficits of auditory-verbal short-term memory typically show damage in left inferior parietal cortex, whereas patients with selective impairment on visuo-spatial working memory tasks show right hemisphere damage that includes posterior regions. This lateralization reflects the left/right division of the human brain with respect to language and visuo-spatial functions.

Neuroimaging studies of normal individuals reveal a similar pattern of cerebral organization (Henson, 2001). Thus, maintenance of verbal information is typically associated with activation in left inferior parietal cortex in conjunction with other regions. In contrast, activation associated with the maintenance of visuo-spatial information is seen in corresponding right hemisphere regions. Neuroimaging studies of dual-tasks and tasks that require monitoring and updating current information implicate dorsolateral prefrontal cortex (D'Esposito *et al.*, 1995), broadly consistent with patient data showing an association between frontal lesions and executive impairments.

The most detailed evidence on cerebral organization concerns the phonological loop. Behavioural studies suggest that the loop has two components – a transient phonological store and a process of subvocal rehearsal that can refresh its contents.

Neuropsychological evidence is consistent with separate localization of the phonological store and the rehearsal mechanism. Thus, damage to the left inferior parietal region of the brain is associated with impaired auditory-verbal short-term memory span but normal speech output, whereas lesions to left inferior frontal cortex disrupt speech production with much less effect on span. Data from neuroimaging of healthy individuals converge with the localization suggested by neuropsychological evidence (see Henson, 2001; Fig. 9.3).

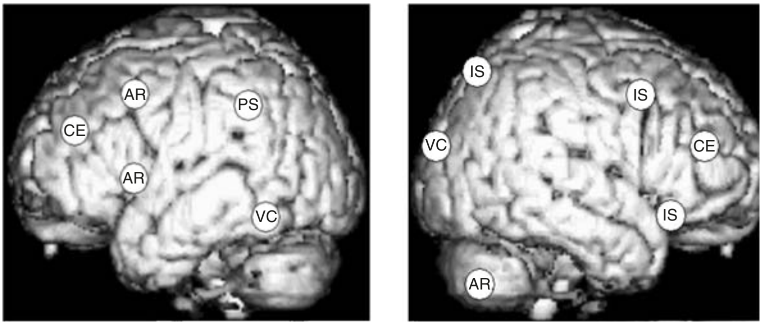


FIGURE 9.3 The multi-component model of working memory can be tentatively mapped on to a wide number of brain regions in both the left and right-hemisphere. CE, PS etc. refer to the components identified in the text. (Reproduced with permission from Fletcher and Henson, 2001)

Connectionist modelling of the way the phonological loop fulfils the important function of preserving information about serial order has suggested that it includes yet a third component that corresponds to some sort of timing signal (Burgess and Hitch, 1999). Neuroimaging during memory for rhythmically grouped lists has tentatively identified this third component with left premotor cortex. The example of the phonological loop illustrates the value of theoretical models for integrating evidence from diverse sources – patients, adults, children, behaviour, fMRI etc. Like others, we note the particular promise of biologically plausible connectionist models for linking neural and behavioural data, for example by simulating lesion effects.

3.1.3 Neural Mechanisms

Our understanding of the physiological mechanisms of working memory remains fragmentary. That it is an active mental state, readily subject to interference, implies that it probably relies on reverberating neural loops in which the activity of one set of neurons depends on the continued activity of others. Experiments with non-human primates have identified neurons in the frontal lobe that fire continuously for short-periods after the triggering stimulus has been turned off and have shown that this activity is essential for accurate choice behaviour in memory tests conducted a few seconds later (Goldman-Rakic, 1992, 1995). However, the local circuits that mediate this have not yet been mapped out. Short-lasting synaptic changes such as post-tetanic potentiation (PTP) that may underlie reverbatory neural activity occur in many cortical neurons. However, they have not been definitively linked to working-memory.

Another potentially important neuronal mechanism is the synchronization of oscillatory firing of neurons at distant sites. This has been proposed as a mechanism for the binding process in visual perception (Singer, 1999) and similar processes may underlie the binding of information in different subsystems of working memory. For example, there is evidence for synchrony in prefrontal and posterior EEG rhythms during short-term retention of verbal and visuo-spatial information (Sarnthein *et al.*, 1998). Raffone and Walters (2001) show how a model involving synchronization and desynchronization of cell assemblies in prefrontal and inferotemporal cortex can account for temporary binding in both visual perception and visual working memory, producing limited storage capacity as an emergent feature. Many other models also exist.

3.1.4 Evolution of Working Memory

How did working memory evolve? Animals, even most mammals, do not have quite the same working memory system as ourselves. We can only speculate how humans come to have it in its present form. It clearly did not evolve to help early hominids to remember telephone numbers and they were spared the burden of remembering whether they had replied to an email.

Unfortunately, we know little about the evolution of working memory: it would be enlightening to know more. There is a high correlation between working memory span and intelligence in humans. If the capacity to store and manipulate temporary information is necessary for activities such as planning, reasoning, monitoring and problem-solving, then it seems likely to confer many evolutionary advantages.

The visuo-spatial component of working memory seems likely to be the most primitive. The verbal component is presumably unique to humans, having co-evolved with speech and language. Early suggestions of its purpose arose from the sequential property of language, the idea being that temporary storage might be necessary during comprehension or speaking.

In one case memory for the words in a sentence is required in order to arrive at its meaning. In the other a similar form of temporary storage is needed to translate a meaning into a sequence of words. Somewhat curiously, the phonological loop is not crucial for comprehending relatively simple sentences. However, studies with young children, adults and neuropsychological patients point to a critical role for the phonological loop in repeating novel words and acquiring new vocabulary items, through its function of preserving information about the serial order of phonemes (Baddeley *et al.*, 1998).

3.1.5 Open Questions

- *Do we need to invoke a 'central executive' in working memory?*

Several questions surround the central executive, the most controversial aspect of the multi-component model. Although widely used in psychology and neuroscience, the concept of an executive processor is often criticized as too close to assuming a homunculus, a 'mind within a mind'. However, this criticism misses the important point that the long-term goal of attempting to specify executive processes is to reduce the need for a homunculus.

The central executive was initially described as a workspace for combining attention and temporary storage, later as a purely attentional system, and more recently as a fractionated attentional system with separate resources for several different types of attention. Among the difficulties for assuming a unitary executive are low intercorrelations between different measures of executive function and empirical dissociations between such measures.

One solution to these difficulties is to assume that the functions of control and coordination are emergent features of a distributed system, and that there is no central executive. However, this approach faces the problem of explaining how such properties emerge and how such a system might operate.

The assumption that executive processes are fractionated is another solution, but it too has problems. For example, all theories have to explain how the coherence over time of mental operations and behaviour is achieved. Norman and Shallice (1986) showed how a unitary attentional system could ensure coherence in a general model of action. A challenge for any fractionated account is to explain how multiple attentional systems interact in an orderly way to realize the orderly control of cognition.

Another set of issues concern the binding problem that we referred to earlier, of how to keep track of related information spread over multiple subsystems. Baddeley (2000) proposed an episodic buffer that binds information from different modalities into integrated episodic records. In this account the episodic buffer is seen as part of the central executive that serves as an interface between working memory and long-term memory. However, the details of how the system operates are still largely unknown.

Yet another aspect of our open question is whether we can equate the central executive with 'conscious awareness'. Dual-task studies have established an empirical link between working memory and consciousness by showing that reported experience alters when working memory subsystems are loaded in different ways (Baddeley and Andrade, 2000). Executive processes may control what representations in working memory become conscious, rather than being identifiable with conscious experience itself (Baars, 1997; Andrade, 2001). However, given the elusive nature of consciousness, rapid progress in understanding how it relates to working memory is unlikely.

In summary, the central executive can be seen as closely linked to a range of key issues in cognition concerning attention and control, the coherence of mental operations, the binding problem, consciousness and intelligence. That this is by no means an exhaustive list merely serves to emphasize the need for further progress in our understanding.

3.2 Long-term Memory

Long-term memory is also sub-divided into different systems. Like working memory, these are located in widely dispersed parts of the brain. The reason for distinct systems is because long-term storage can have very different functions (Sherry and Schacter, 1987; Nadel, 1992).

There is still no universal agreement about the organization of long-term memory, but many psychologists and neuroscientists consider that it consists of at least the five main components that we have space to discuss – semantic, episodic (including recollection), recognition memory (e.g. that based on familiarity), skill and value memory. However, the boundary between these subsystems is, at best, fuzzy and it is best to consider them as inter-dependent rather than independent memory systems. We shall consider their distinction in psychological and anatomical terms, beginning with episodic memory.

3.2.1 *Episodic Memory*

Memory for events is the type of memory that laymen regard as memory – the bringing to the mind's eye of a picture of something from the past. This is technically called 'episodic' memory and it is used to keep track of events (Tulving, 1972, 1983).

What is special about events is that they are unique – they happen only once. It is a snapshot system. A neural network that requires hundreds of trials to learn would not work for episodic memory. Having encoded attended information, the episodic memory system holds onto most of it for a while. It cannot hold onto all attended information forever and we gradually lose access to most of it over time.

Neuroscientists think this happens to the trivial and unimportant information that we automatically pick up during the course of a day. Some of this forgetting is true-forgetting – the memory traces are simply no longer there any more (trace decay). Some forgetting, however, is accompanied by the phenomenological sense that we do really know, if only we could remember. In such circumstances, others may remind us with a clue and the information will then come flooding back to mind. This kind of forgetting is called 'retrieval failure' because the memory traces are still there but there are problems of access. Schacter (2001) describes this as one of what he endearingly calls the 'seven sins of memory'.

An intriguing and important feature of episodic memory is the 'mental time travel' component of recollection. When we remember a past event, we travel back in time in our mind's eye to the place and time where the event happened. The ability to travel back in time carries with it an implicit sense of personal identity. That is, you remember the event not only happened at a particular time and place (what, where and when), but also that it happened to you.

Tulving claims that true episodic memory requires 'autonoetic consciousness', an awareness of the self. He argues that no living thing that lacks this form of consciousness could be truly said to possess this form of memory.

However, some scientists investigating rapidly formed event memories in animals argue that other vertebrate species definitively possess an 'episodic-like' system because they can be shown to remember what, where and when (Clayton and Dickinson, 1998). It will, however, be difficult to establish whether animals have a sense of their own personal identity (Griffiths *et al.*, 1999).

Episodic memory is impaired in a condition known as 'global amnesia', which results from damage to structures including the hippocampus, neocortical structures that are interconnected with the hippocampus such as the entorhinal, perirhinal and parahippocampal cortex, and mid-brain structures such as the mamillary bodies and dorsomedial thalamus. The fibre connections interlinking these brain structures, such as the fornix, are also important.

The people affected by global amnesia cannot remember whether they have recently eaten a meal, or that they ought to have one, and they forget where things have been put down around the house. Shown a complex drawing, amnesic patients can copy it accurately but cannot draw it from memory after 30 minutes.

An amnesic's life lacks all structure in time and place, and has been described by one extensively studied amnesic patient 'H.M.' as like continually 'waking from a dream' (Corkin, 2002). Yet, in general, these same amnesic patients retain their appreciation of the world around them, their command of language and the meaning of words (intact semantic memory – see below), and enough short-term memory to do crossword puzzles or to carry on a sensible conversation (intact working memory). It is not until one has exactly the same conversation with the patient a few minutes later that the devastating isolation of their existence becomes clear.

Often, but not always, amnesic patients cannot remember things that happened before they became ill. This is called 'retrograde amnesia' – where retro refers to time past – to distinguish it from the memory loss seen for material presented after they become ill, anterograde amnesia. The presence of extensive retrograde amnesia contributed to Warrington and Weiskrantz's argument that some aspects of amnesia must be due to retrieval failure (Warrington and Weiskrantz, 1968; Warrington, 1982).

A possible problem is that some amnesic patients do retain memories from the remote past, losing only those from a period immediately prior to the onset of amnesia – temporally graded retrograde memory impairment. Given this pattern, it is not clear whether we can regard amnesia as a deficit in the capacity for mental time travel: if an amnesic patient has any intact remote memory, then clearly he or she can mentally time travel back to re-experience the past.

Some see temporally graded loss of episodic memories as evidence for the existence of a consolidation process, whereby initially labile memories are strengthened over time and become independent of the structures involved in their acquisition (Squire, 1992; Graham and Hodges, 1997). However, the finding of temporally graded retrograde amnesia is contentious, and there are some grounds for suggesting it arises as a consequence of the formation of multiple memory traces rather than an explicit consolidation process (Nadel and Moscovitch, 1997).

Functional brain imaging has provided a new window on episodic memory. One might expect, given the consistency of findings with neurological patients, that imaging studies would also reveal the diencephalon and medial-temporal lobe to be activated in subtractive experimental designs focusing on memory processing. To everyone's surprise, work throughout the 1990s revealed this not to be the case.

Study after study showed differential activation in the prefrontal lobe during memory encoding and retrieval, and in an area of the parietal lobe called the precuneous during the retrieval of verbal and visual long-term memory (reviewed in Frackowiak *et al.*, 1997). Activations in the medial temporal lobe were rare.

The situation has changed with new studies revealing differential activation in association with parameters such as novelty (medial temporal lobe), retrieval effort (frontal lobe activation) and retrieval success (medial temporal lobe). Moreover, the advent of event-related fMRI studies has made it possible to narrow the window of time over which the BOLD signal can be identified to the presentation of individual words or pictures. Memory for these items can be tested later, outside the scanner, making it possible to identify the imaging signature of items that later proved to have been memorable and those that were not. In this way, brain imaging scientists are starting to dissect the cerebral localization of distinct memory processes such as encoding by identifying successful versus unsuccessful encoding.

3.2.2 Familiarity-based Recognition Memory

'Recognition memory' is defined as a judgement of prior occurrence. An experimental paradigm might involve studying a list of 50 words, after which the subject should indicate which of a new series of words – some novel and some seen previously – were presented earlier. This and a visual version are components of the internationally used Warrington Recognition Memory Test, developed in the UK.

While recognition memory may seem to be tapping into the ability to travel back in time to the study episode, cognitive research suggests that two processes – familiarity and recollection – may contribute in different ways. This psychological dissociation appears to have been re-discovered after being first proposed over 20 years ago (Mandler, 1980).

Familiarity is restricted to the phenomenological experience of merely having seen something before. Recollection, on the other hand, involves the conscious remembrance of one or more past episodes. Familiarity and recollection are studied in the laboratory by asking subjects to state, alongside their memory judgement, the additional information of whether they 'remember' seeing an item in a list (recollection) or that they merely 'know' the item was on the list (familiarity). This distinction and several others are discussed thoroughly in an impressive series of papers that accompanied a recent conference on episodic memory at the Royal Society (Baddeley *et al.*, 2002).

Linking their ideas about the neuroanatomical basis of spatial and episodic memory with the data on object recognition memory, Aggleton and Brown (1999) have suggested that a structure in the medial temporal lobe, the perirhinal cortex, mediates the simple sense of familiarity, whereas the hippocampus is important for episodic recollection. Aggleton and Brown's suggestion that the perirhinal cortex is important for familiarity is supported by much laboratory work.

Rats, monkeys and humans with damage to this brain area have difficulty in object recognition memory. Patient H.M., with damage extending across several medial temporal lobe structures, has severely impaired object recognition memory. However, no one had linked this deficit to his perirhinal cortex damage before the advent of animal studies.

The first such evidence came from single-unit recording studies, revealing decreased firing rates the second time a stimulus was presented (Brown *et al.*, 1987). This repetition suppression effect has been proposed as part of a neural mechanism for object recognition that Brown and his colleagues have built into a neural network model. Second, perirhinal cortex lesions disrupt object recognition memory in both the visual and tactile modalities. Third, studies examining the expression of immediate early genes, such as c-fos, have corroborated the electrophysiological and lesion studies (Wan *et al.*, 1998). The supposition is that these genes are activated strongly on the first occasion that a stimulus is processed by a group of neurons, but to a lesser extent thereafter.

The perirhinal cortex is not the only region of neocortex to show repetition suppression effects in neuronal firing. Other temporal cortical regions, such as area TE, also show these effects. Lesions in area TE, like those in perirhinal cortex, disrupt object recognition.

However, whereas perirhinal cortex lesions disrupt object recognition at long but not short delays (suggesting an impairment in memory rather than perception), TE lesions can impair recognition even at very short delays. This has been taken as evidence for a role for TE and other regions of inferotemporal cortex in visual perception rather than memory (Buffalo *et al.*, 2000). This is consistent with the long-established finding that damage in these inferotemporal cortical regions can disrupt the perceptual discrimination of visual stimuli (Dean, 1976).

These inferotemporal cortical regions have traditionally been thought of as cerebral components of a ventral visual stream (VVS) for object perception (the 'what' pathway) that can be contrasted with a dorsal visual stream (or 'where' pathway) for the processing of visuo-spatial information. The latter extends from the visual cortex into the parietal lobe. The characterization of the information processing streams remains contentious, with various research groups (e.g. Goodale and Milner, 1992) arguing that dorsal pathway is more concerned with the processing of visuo-spatial actions than the identification of location. These streams contain many anatomically distinct regions, and are thought by some to extend through numerous subdivisions as far as the frontal cortex (Fig. 9.4).

The distinction between anatomically separate systems for memory and perception remains very much an open question. For example, perirhinal cortex lesions, like

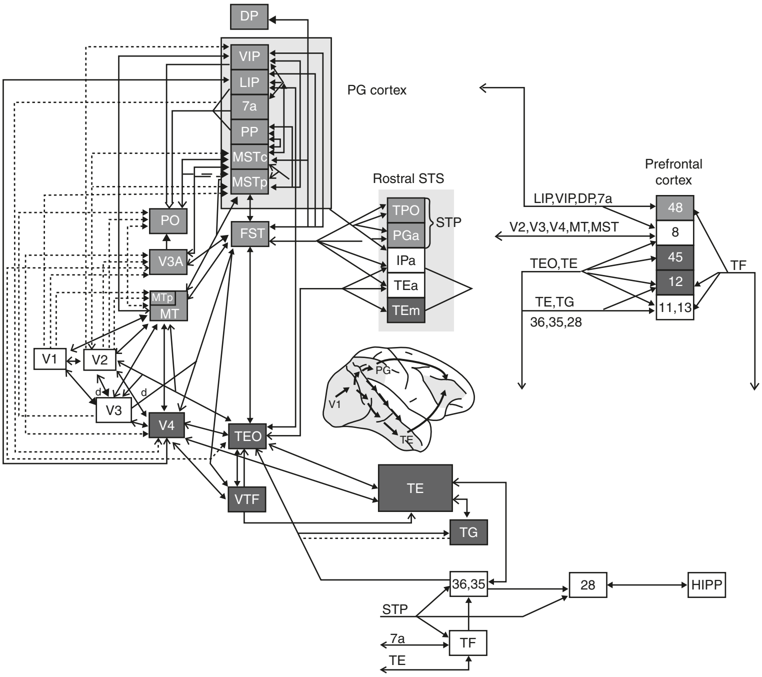


FIGURE 9.4 The presumed dorsal (light grey) and ventral (dark grey) visual pathways of the macaque brain. Note presence of both forward and backward neocortical projections, and pathways on to the frontal lobes. (Based on the proposals of Ungerleider and Mishkin and reprinted by permission of the authors. This figure appears in the colour plate section.)

lesions in other areas of the VVS, can produce deficits in visual discrimination (Buckley and Gaffan, 1998; Bussey *et al.*, 2002). It may, therefore, be important for both memory *and* perception. If so, it may be preferable to describe functional aspects of the brain organization of the VVS in terms of perceptual representations rather than with reference to a sharp, binary distinction between perception and memory.

If we could use these representations for both perception and memory, it becomes of interest to ask what types of representations exist and where are they stored? This approach is also now being applied to the dorsal visual stream. For example, although the parietal cortex is usually thought of as important for spatial perception and attention, Gaffan and Hornak (1997) have pointed out that retrieval of spatially organized memory is also impaired in patients with parietal cortex damage. They argue that parietal cortex contains a retinotopically organized representation of space that is important for spatial perception, attention and memory.

3.2.3 Semantic Memory

The semantic memory system of the brain contains a vast storehouse of factual knowledge about the world, i.e. concepts, the meanings of words and people (Tulving, 1972; Warrington, 1975). Knowledge may include the fact that pianos are musical instruments, that Paris is the capital of France, that DNA encodes genetic information in a sequence of base pairs, and so on. Semantic memory is quite different from the mental experience of episodic memory in that the contextual tags encoded at the time of learning are long since lost. We know that canaries are yellow, but do not remember where or when we learned this fact. In episodic memory, by contrast, it would be difficult for most people to remember what Nelson Mandela said to us during a meeting without also remembering where and when this conversation happened.

Much of our current knowledge about semantic memory, and its relationship to other cognitive systems, has come from the study of neuropsychological patients, in particular those who suffer damage to anterior temporal cortex. Patients with a semantic memory deficit cannot remember the names of familiar objects around the house or the names of friends whom they see only once/twice a year, and show difficulties on tests that require knowledge about semantic concepts. For example, patients may be unable to select the picture that matches a spoken word (e.g. cow) or sound (e.g. moo) from an array of animal photographs, and even on tests that are not linguistically demanding, such as selecting the appropriate colour (yellow) for a black and white line drawing (banana), they can be strikingly impaired (Patterson and Hodges, 2000).

Studies of these patients reveal fascinating patterns in the way in which semantic memory can break down. For example, patients often retain broad, general knowledge about concepts, such as knowing that a 'tiger' is an animal, but may not know that a 'tiger' is foreign, dangerous and has stripes. Consequently, they may happily endorse suggestions that a tiger would make a suitable household pet.

The influential model of semantic memory of (Collins and Quillian, 1969) consists of a hierarchically arranged network of links between concepts, with properties that apply to a set of concepts stored at the highest level to which they were generally applicable (see Fig. 9.5). Specific nodes in the hierarchy were understood to inherit the properties of the more general nodes to which they were linked, providing an efficient means of knowledge storage, retrieval, and generalisation. For example, since most birds fly, this property is best attached to the general concept of 'bird' rather than to every example of a bird, while the fact that penguins 'swim' is best attached to the more specific 'penguin' node.

The theory predicted that simple propositions should take longer to verify when their terms were far apart in the hierarchy

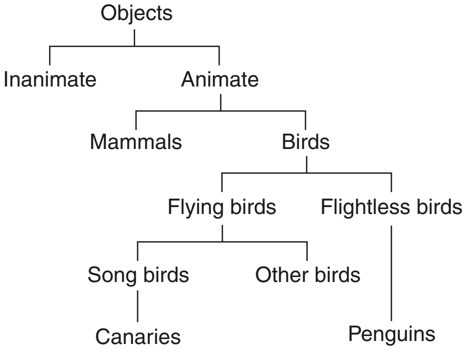


FIGURE 9.5 A semantic-memory tree diagram (based on framework due to Collins and Quillian, 1969)

(e.g. a canary can move) than when they were close together (e.g. a canary can sing). Early experiments seemed to bear this out, and at first glance, the model also provides a useful framework in which to interpret data from patients with loss of semantic memory, which is suggestive of a pruning back of a hierarchically organized semantic tree.

Later studies in normal subjects, however, found that reaction times in verification experiments were influenced predominantly by prototypicality rather than by distance in a semantic hierarchy (Ripps *et al.*, 1973). More recently, typicality has also been found to be a powerful predictor of performance in neuropsychological patients with semantic impairments. For example, individuals with the neurodegenerative condition semantic dementia who show a progressive loss of semantic knowledge as the disease progresses, will be more likely to select as real a picture of a camel without a hump than one with or to pick out a word that follows lexical rules (e.g. dollop) as opposed to one that does not (e.g. polyp).

One challenge of understanding the organization of semantic memory is how to explain why factors such as typicality influence performance. The hierarchical model of Collins and Quillian (1969), and modifications that came after it, do not clearly account for this factor.

An alternative view assumes a distributed network in which factual knowledge is an emergent property of interactive activation between modality-specific perceptual representations. Unlike the Collins and Quillian framework, semantic knowledge is not stored within a hierarchy of explicit propositions, but instead takes the form of patterns of activity in the brain, that capture hierarchical structure and prototypicality implicitly (Rogers and Plaut, 2002).

In this distributed account, the semantic system is arranged within a neuroanatomical processing hierarchy, with different modality-specific perceptual regions of cortex converging upon an amodal and relatively homogeneous region of association cortex in the anterior temporal lobes. This region serves to store gradually accumulating associations among percepts in different modalities (e.g. the face of Tony Blair and the word 'politician'); and in so doing, it discovers representations which capture the general semantic similarity among different concepts in distributed form.

Effects of familiarity and typicality emerge as a matter of course in such networks; hence the framework offers a means of explaining both hierarchical and typicality effects in a single model, which itself is consistent with the anatomically hierarchical organization of cortex. The distributed account further suggests that semantic and perceptual information are not stored independently, but are mutually interdependent, a suggestion that accords well with recent findings documenting an influence of semantic impairment on various kinds of high-level perceptual tasks such as object recognition and discrimination.

Category-specific deficits constitute another fascinating pattern seen in patients with semantic memory loss. Warrington and Shallice (1984) reported four encephalitic patients who had problems understanding words that referred to living things, but performed much better on inanimate objects. Although rarer, other cases have been described who show the opposite pattern (living → inanimate).

Warrington and Shallice suggested that these types of pattern could reflect the profile of sensory dimensions that contribute to each domain, with living things more dependent upon perceptual characteristics and inanimate upon their functional attributes. A number of patients have also been reported who show poor knowledge of famous people in the context of good retrieval of factual information about objects and animals (Ellis *et al.*, 1989).

The opposite dissociation has not often been reported, and it is currently unclear whether these two categories of semantic knowledge are psychologically and/or neuroanatomically distinct. It is an open question, therefore, as to whether semantic memory is organized categorically or whether category-specific differences actually reflect other factors, such as specificity or uniqueness.

Distributed models can account for such patterns: if different regions of the brain are involved in encoding sensory properties, it is possible that a semantic system might become loosely structured around these regions and damage could affect one system more than another. In the distributed network described by Rogers *et al.* (2004), a similar principle can explain category-specific deficits: for example, damage to connections between representations and semantics may affect one category more than another (e.g. difficulties discriminating animals may arise after a loss of connections between vision and abstract semantic representations).

3.2.4 Semantic and Episodic Memory

The study of developmental amnesia has provided new information about the possible relationship between episodic and semantic memory, an open question that is much debated in the literature. A particular controversy is whether the learning of new semantic facts is impaired in amnesia.

For many years, researchers believed that there was disruption to both the formation of episodic and semantic memories, with the latter affected because the acquisition of new semantic facts was dependent upon episodic memory. Startling new findings in cases of developmental amnesia, with damage to medial temporal regions sustained very early in life, challenge this assumption.

These individuals, while having severely impaired episodic memory, often perform well at school and acquire new vocabulary and factual knowledge about the world, a pattern which suggests episodic and semantic memory can be learnt independently of each other (Vargha-Khadem *et al.*, 1997). Intriguingly, MRI examination of the brains of these individuals has found that damage to the hippocampus but not other medial temporal lobe regions (e.g. entorhinal cortex), a finding consistent with animal lesions studies. It is an open question whether the pattern in developmental amnesia can tell us something fundamental about the organization of human memory systems or instead reflects developmental plasticity in the young brain.

The opposite dissociation, good episodic memory in the context of poor semantic knowledge (e.g. knowing you saw an elephant at the zoo yesterday but not knowing what an elephant is), has also been reported. Patients with semantic dementia perform well on some tests of episodic memory, in particular those tapping recognition memory, even for stimuli for which they can no longer name or point to accurately from the spoken word, or colour in appropriately.

Impairments in memory for 'unknown' but not 'known' stimuli are seen, however, when a perceptual change is made between study and test. Such findings challenge an influential model of long-term memory (Tulving, 2001), in which semantic knowledge is a prerequisite for episodic memory (Fig. 9.6a), and suggest instead that perceptual information about studied items also contributes to episodic memory (Graham *et al.*, 2000; Fig. 9.6b). It is not clear, however, given that recognition memory can be supported by familiarity independently of

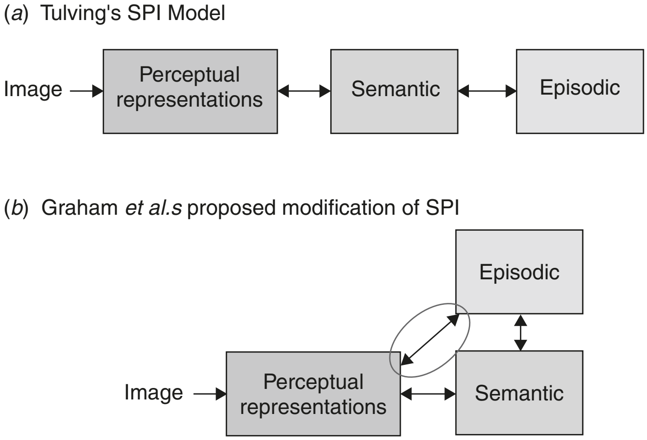

(a) Tulving's SPI Model
Image
Perceptual representations
Semantic
Episodic
(b) Graham et al.s proposed modification of SPI
Episodic
Image
Perceptual representations
Semantic

FIGURE 9.6 (*a*) Tulving's model, termed SPI (serial, parallel and independent) asserts that perceptual information feeds into semantic memory, and that from semantic memory into episodic memory (serial processing), while the processing within these systems operates independently in parallel. (*b*) Data from semantic dementia suggests an alternative account, whereby the output from perceptual representations can also support some forms of episodic memory.

recollection, whether patients with semantic dementia show true episodic memory (i.e. mental time-travel).

These findings, when considered alongside those from developmental amnesia, imply a psychological and probable neuroanatomical dissociation between episodic and semantic systems, at least in the acquisition of new memories. Such a theoretical view seems oversimplistic, however, and the challenge for the future is to ask under what circumstances are episodic and semantic memory dependent upon each other, and to create more accurate theoretical models that allow us to capture the nature of their interactions, and those they have with other critical long-term memory systems, such as perceptual representations. We also need to grapple with the mapping of these psychological models onto the brain, which may not respect the same principles (e.g. hierarchy or category) that seem so critical from patient investigations.

3.2.5 *Sensorimotor Skills*

Knowing that a piano is a piano is one thing, being able to play it is another. This and other sensorimotor skills, such as sports skills, are learned through deliberate and extensive practice. Practice is necessary to realize the accuracy and precision timing of movements. Skill learning is also thought to be an associative process, but one between a person's own actions and that of reward.

The reward may take the form of the person's own appraisal of their acquisition of skill (feedback with results) or the receipt of a separate reward for the action correctly emitted (e.g. a pellet of food to a hungry rat). Once someone has acquired a motor or cognitive skill, it generally becomes habitual and relatively insensitive to reward. People do gradually forego unrewarded habits, but their execution is not goal-motivated in quite the way that actions are. Interestingly, the skills that underlie habits are rarely forgotten. An adult who has not been on a bicycle since childhood can usually ride one again immediately or, at least, pretty quickly.

Amazingly, amnesic patients can learn skills but then cannot remember having done so. This extraordinary dissociation between explicit and implicit memory never ceases to fascinate. Amnesics can learn motor skills or the cognitive skill of reading backwards.

Learning to read backwards quickly takes a while. This is true for amnesic patients no less than for anyone, but whereas most would remember the painful process of learning to do this, amnesics do not. This dissociation in their conscious awareness is part of our reason for believing that the distinctions between different types of memory described above may be honoured by the nervous system. Consistent with the double-dissociation concept, damage elsewhere in the brain affects skill learning (basal ganglia and cerebellum), but neither episodic nor semantic memory. These and other putative roles for the basal ganglia are reflected in the motor and cognitive deficits seen in patients with disorders of the basal ganglia, such as Parkinson's and Huntington's disease.

3.2.6 Value and Emotional Memory

Our understanding of conditioning processes has gone through something of a revolution in the past 20 years with the discovery of several experimental phenomena that collectively point to value learning or emotional memory being something that happens when a person or animal's expectations about reward are violated.

Knowledge acquired during value learning is about what items in the world are 'good', because they have been associated with *positive* events, and what items are 'bad', having been associated with *negative* outcomes. This form of learning has been most often studied in animals through associative conditioning, which involves the pairing of initially neutral stimuli with biologically significant stimuli such as food (e.g. that are valued by a hungry animal), or potentially harmful stimulation (e.g. the sight of a predator). Discovered by Pavlov at the start of the twentieth century, associative conditioning is now thought to reflect a process through which stimuli come to reflect the causal texture of the world (Dickinson, 1980).

Certain stimuli acquire value by virtue of predicting positive or negative stimuli, whereas others acquire inhibitory value by virtue of predicting that these stimuli will not occur. Thus, in a regular environment in which various kinds of events happen with a certain degree of predictability, stimuli will either become good predictors of these events (both positive and negative events) or predictors of the absence of these events. Different learning rules have been proposed for this process with one of these, the Rescorla/Wagner rule (Rescorla and Wagner, 1972), being strikingly similar to the so-called 'delta rule' often used in artificial neural networks.

The brain region that has dominated our studies on value learning and emotional memory is the amygdala (so called because of its resemblance to an almond). It is located in the medial temporal lobe, anterior to the hippocampus, consists of a number of separate nuclei, and has numerous interconnections with cortical and subcortical structures.

The first realization that the amygdala was important for emotion emerged in a classical study by Kluver and Bucy (1939), in which monkeys sustained surgical removal of the temporal lobe including the amygdala. The monkeys were described as exhibiting 'psychic blindness', evidenced by blunted emotional responses to normally fearful stimuli (more recent studies are described by Meunier *et al*., 1999). Humans with temporal lobe damage exhibit similar emotional changes.

Earlier, we discussed patient H.M., whose temporal lobe lesions led to severe impairments in declarative memory. H.M.'s lesion also included the amygdala. As a result, although he can detect and respond to painful stimuli, he does not describe them as being painful – he appears emotionally indifferent to this type of pain. Recent studies have shown that more selective damage to the amygdala results in other emotional changes, such as a decreased ability to recognize emotion in human faces (e.g. Adolphs *et al.*, 1995).

Studies in rodents have also examined the contribution of the amygdala to learning. Work by McGaugh and his colleagues has long emphasized a key role for the amygdala in the emotional modulation of other forms of memory (McGaugh, 2000). According to this view, emotional stimuli activate hormonal systems that, in turn, interact indirectly with the amygdala. The resulting amygdala activation influences, through a network of cerebral interconnections, consolidation within other memory systems, such as episodic memory. This process has long been studied in experimental paradigms in which rodents are required to learn to inhibit a response to avoid a painful stimulus. Such memories are acquired rapidly, but may or may not persist. By the careful post-training application of drugs that interact with the successive consolidation phases of memory, neurobiological mechanisms mediating this process have been identified at both a network and cellular level.

Research by LeDoux, (2000) focused on models of emotional learning. A light or a tone is presented to a rat, followed by the aversive stimulus of a mild footshock. This is severely impaired in rats with lesions of the amygdala, leading him to argue, in contrast with McGaugh, that memory traces are formed and stored *within* the amygdala itself. Other studies have revealed that the amygdala is also engaged in positive emotional conditioning (Hall *et al.*, 2001), although other brain regions are also important such as the cingulate cortex and ventral striatum.

It has recently become clear that the different subnuclei of the amygdala contribute to different aspects of emotional behaviour. Specifically, the basolateral nucleus is thought to mediate stimulus-value associations, knowledge which can be used to modify voluntary behaviour; the central nucleus, in contrast, is thought to be part of a stimulus-response system that automatically runs off autonomic emotional responses, changes in heart rate, sweating, in response to emotional stimuli. It is interesting that these two systems within the amygdala parallel the flexible declarative and inflexible habit systems described earlier.

New studies that most directly illustrate the amygdala's role in value learning are called 'reinforcer devaluation experiments'. In these, an animal such as a monkey learns stimulus-value or response-value associations, for example that choosing one of two objects leads to the reward of a 'fruit-snack' while choosing another object leads to a peanut. One of these two rewards is then devalued by allowing the monkey to become sated on it. While normal monkeys display the very human phenomenon of stimulus-specific satiety, those with amygdala damage do not show this reinforcer devaluation effect (Malkova *et al.*, 1997). Disconnection of the amygdala from the orbital prefrontal cortex has a similar effect. The interpretation is that the amygdala provides information about stimulus-value relationships, and the orbital prefrontal cortex mediates the alterations in voluntary behaviour that reflect these changed relationships.

The role of the amygdala in value learning and emotional memory points to a potentially important role in social behaviour. Indeed, some authors have recently raised the possibility of an important role for the amygdala in social interactions (Emery and Amaral, 2000). An open question is whether amygdala dysfunction contributes to human conditions characterized by emotional and social changes, such as autism in which it has recently been implicated (Baron-Cohen, 1995).

It is important to appreciate that multiple memory systems are not isolated modules that cannot interact. In fact, they can, do and must interact to realize the seamless control of behaviour. Although we have discussed episodic and emotional memory separately, it is clear that the content of emotional memories can reach conscious awareness and, in this sense, can be thought of as declarative. That said, an open question about this system relates to its accessibility to consciousness. Certainly we know and are consciously aware that certain things make us feel good. However, we may be unaware of the conditioning processes that gave rise to that state of affairs. Emotional memory has this curious dissociative feature – an awareness of final emotional value coupled so often with limited knowledge of provenance.

3.2.7 Open Questions

A key open question relates to the central concept around which this summary is structured: the concept of multiple memory systems. This widely accepted idea is coming under review. What is the relationship between memory systems such as working, episodic, perception and semantic memory? This open question subdivides into further issues.

- *The concept of multiple memory systems, while apparently widely accepted, is coming under review (Gaffan, 2002; Bussey, 2004). What is the relationship between different memory systems, for example episodic, perception and semantic memory? What are the inputs to working memory from semantic or episodic memory?*

A key issue is the relationship between working memory and long-term memory. One theoretical approach assumes entirely separate memory systems and focuses on questions such as whether working memory is involved in encoding and retrieving information in long-term memory. Experimental evidence suggested that loading working memory with irrelevant material impairs encoding in episodic memory, and has a smaller effect on retrieval, while having no effect on implicit learning.

An alternative theoretical approach is the idea that working memory corresponds to an activated part of long-term memory rather than an entirely separate store (Cowan, 1999). One difficulty for this kind of view is how to account for dissociations between performance on short-term and long-term memory tasks. For example, auditory-verbal short-term memory can be severely impaired while long-term episodic memory is unaffected, whereas the con-verse holds for the classical amnesic syndrome. However, impairment of phonological short-term memory is now known to be associated with poor long-term memory and learning for phonological information even though long-term memory in other domains is normal.

This neuropsychological evidence suggests the possibility of a separate phonological memory system in which mechanisms for short and long-term storage are closely related, rather than separate short-term and long-term phonological stores. The computational model developed by Burgess and Hitch (1999) illustrates how this might be realized in a neural network in which the weights of modifiable connections have two components, one with short-term stability, the other long-term.

Other researchers have suggested a close connection between working memory and long-term memory. Ericsson and Kintsch (1995) discussed the finding that chess experts have much higher working memory capacity in chess-based tasks than novices, principally because they have learned to organize chess information into larger chunks in long-term memory. Similar observations have been made within other domains of expertise. To account for them, Ericsson and Kintsch introduced the concept of long-term working memory, a mechanism whereby working memory has rapid access to knowledge structures in long-term memory.

Baddeley's concept of an episodic buffer is another solution to the same problem. While it is premature to draw firm conclusions, it seems highly unlikely that the totality of working memory is no more than activated long-term memory, for if this were so we would surely not be able to escape the constraints of our prior habits and learning. This is not to deny the importance of our past, but to acknowledge that perhaps the most important function of working memory is to allow us to go beyond prior learning to reason, plan and problem-solve intelligently.

- *Characterizing amnesia and its anatomical substrates*

A longstanding if somewhat tired issue is how best to characterize human amnesia. Is it a selective impairment in episodic memory? Or is semantic memory, especially the acquisition of new semantic facts, also affected?

A newer focus concerns the nature of mental time travel, and whether humans the only species that have autonoetic consciousness – i.e. awareness that the self that is experiencing *now* is the same self that experienced in the *past*? There is also continuing debate about the critical pathology that leads to human amnesia. Gaffan and his colleagues, for example, have argued that damage to the hippocampus is insufficient, and that damage to pathways leading into the temporal lobe, such as the fornix and the temporal stem, contributes. This claim needs continued testing in animal models, particularly primate models, and to be taken seriously by researchers working with that rare subset of amnesic patients with circumscribed brain damage.

- *Perception and familiarity-based recognition memory*

We touched upon the idea that the neural substrates of object recognition memory may be coextensive with those of perception. Researchers in this area have generated computational models of perirhinal cortex, a brain area thought to mediate both functions. Although models have been built to simulate recognition memory and visual discrimination learning (Bussey and Saksida, 2002), no models integrate the two. How can we reconcile the fact that damage to the perirhinal cortex can paradoxically lead to both 'delay-dependent' impairments in memory and 'difficulty-dependent' deficits in perception?

- *Semantic memory, perceptual-representations and connectionist modelling*

Our current understanding of the organization of semantic memory is strongly influenced by connectionist, parallel-distributed processing models, rather than hierarchically organized networks – at least at the cognitive level. Burgeoning interest in neurodegenerative conditions that allow us to investigate the breakdown of conceptual knowledge systematically in a particular disease (such as semantic dementia) has created a unique opportunity to generate an experimental database that could be simulated by computational models.

We are also in the dark about the relationship between non-semantic cognitive systems, such as perceptual-representational memory and true referential semantic memory. Are these really distinct? Do the deficits seen in linguistic and visual tasks in semantic dementia reflect the dysfunction of a common mechanism? And what do we really mean by abstract, modality-independent semantic knowledge – is it a separate system or does it emerge in some way from the interaction of sensory representations? The broader goal here is to map these increasingly sophisticated distributed-processing models onto the brain, in which there is thought to be a representational hierarchy (e.g. the ventral visual processing stream).

- *The binding problem and memory*

Some of the important types of memory we have discussed, for example episodic and semantic memory, are complex, and require the 'binding' together of multiple types of information. How is this binding achieved? Is the formation of conjunctive representations relevant, perhaps alongside spatial attentional processes (Treisman, 1996)? If so, is there a mechanistic role for associative plasticity (such as LTP)? Alternatively, or additionally, what is the role of the synchronous firing of neurons that code the bound representations (Singer, 1993, 1999)? Recent research has identified synchronization of gamma and theta oscillations as a possible neuronal substrate of memory formation (Fell *et al.*, 2003).

- *Computational modelling and animal studies will help us to better understand the mechanisms of memory consolidation*

As briefly discussed in section 3, memories appear to remain temporarily in a labile state until they are 'consolidated'. Early work on memory consolidation was conducted in the absence of careful thinking about retrieval mechanisms and the possibility that memory failure could be due to impaired retrieval rather than interrupted consolidation. However, while a great deal of attention has been focused on memory consolidation in human memory, we are at an impasse in understanding this process – especially at a cognitive level. Patients sometimes (but not always) display temporally graded loss of episodic memories. Why is the pattern not consistent across cases with similar pathology, and how do we move forward and test models that are currently underspecified?

Computational models are being developed that simulate the process of systems-level consolidation. These capture the rapid associative learning properties of the hippocampus (appropriate for snap-shot encoding of events and episodes) and intersect with the slower interleaving properties of neocortical learning. A next step might be to use these models to strengthen our theoretical views and to create predictions that can be tested in amnesic patients and relevant animal models.

Computational models to date have generally addressed only systems-level consolidation (at the level of anatomical structures in the brain). However, an arguably better understood concept is that of *cellular-level* consolidation (thought to be due to protein-synthesis in the same cells and circuits in which individual synaptic weight changes are encoded). The latter can only be studied invasively (using animals).

To take this endeavour forward, and to understand how systems-level and cellular-level consolidation are related, will require the collaboration of researchers at a variety of levels, including behavioural, pharmacological, transgenic and other gene-targeting techniques, alongside computational models that are informed by neuropsychological findings. Can new computational models be developed to capture the multiplicity and interactive nature of memory?

In summary, an obstacle as we see it is that computational models of learning and memory of which we are aware, ranging from machine-learning approaches that use symbolic computational techniques through to distributed neural networks using Hebbian learning or back-propagation, do not yet build in the concept of multiple memory systems. We have argued that this concept is a major advance, not one to be accepted without a continuing critical appraisal, but an advance nonetheless. There are different types of memory. It is vital that neuropsychologists and computational/neural network modellers work more closely together to understand how different cortical areas implement, together and in conjunction, the different forms of memory of which the human brain is capable.

4 THE NEUROBIOLOGY OF LEARNING AND MEMORY

Our understanding of the neural mechanisms of learning is developing at a rapid rate, based on careful research using laboratory animals, *in vitro* brain slice and cell culture techniques. Developments in molecular neurobiology have led to the identification of genes and gene-products that may have a primary role in experience-dependent alterations of the nervous system. Following the Human Genome Project, this discovery process will only continue. We highlight three major discoveries made by UK-based neuroscientists that are a continuing focus of attention: (1) the discovery of long-term potentiation; (2) place-cells in the hippocampus, and (3) similarities between experience-dependent self-organization of the developing nervous system and adult learning mechanisms.

4.1 Long-term Potentiation and Memory

In keeping with predictions made over 50 years ago (Hebb, 1949), long-lasting forms of synaptic plasticity have been identified in hippocampal and cortical neurons. The discovery of long-term potentiation (LTP) by (Bliss and Lomo, 1973) and the subsequent worldwide efforts to identify its mechanisms and functions is one of the great scientific discovery stories of modern neuroscience (Martin *et al.*, 2000). LTP was discussed at a Discussion Meeting of the Royal Society in May 2003 (Bliss *et al.*, 2004).

LTP is a physiological phenomenon in which specific patterns of electrical activity in neurons cause a long-lasting increase in synaptic efficacy. LTP has the basic properties of persistence, input-specificity and associativity – and a number of other newly discovered properties that are all desirable properties of a memory system.

Neural network models of memory formation, beginning with Marr (1971) and continuing to a range of parallel distributed systems first described by (Rumelhart and McClelland, 1986), have led to widely discussed connectionist accounts of memory encoding and storage processes. Networks with a variety of different learning rules and network architectures have been shown to possess intriguing properties such as content-addressability, pattern completion and graceful degradation in response to injury.

In parallel with research aimed at understanding how synaptic plasticity embedded into networks can realize memory is a tough-minded reductionist effort to pin down the precise physiological, pharmacological and molecular mechanisms of synaptic change. Key findings include major discoveries about the neurotransmitter glutamate and its associated receptors. One of these, the NMDA receptor, is critical for the induction of LTP (Collingridge *et al.*, 1983). Activation of this receptor triggers a cascade of synaptically localized mechanisms that result in the expression of synaptic potentiation (i.e. an increase in synaptic weight). This change can be on the pre-synaptic (sending) side of the synapse and the post-synaptic (receiving) side. In the case of the latter, it is now clear that AMPA receptors, a different type of glutamate receptor, cycle between the cytosol and the membrane in a highly active and dynamic manner. Some cycling is constitutive while other aspects are activity-dependent. In this way additional AMPA receptors can be inserted into the membrane to express a stronger synapse (Fig. 9.7). As the AMPA receptor is ionotropic (i.e. contains an ion-channel through which charged ions travel), alterations in synaptic weight can also be expressed as changes in the amount of current that can be passed through the channel.

Developing knowledge about the intricate mechanisms involved in synaptic plasticity is likely to play a key role in developing novel drugs to boost memory function. This assertion is based on several findings. First, blocking the NMDA receptor either pharmacologically (Morris *et al.*, 1986) or using molecular-genetic techniques (Tsien *et al.*, 1996) causes learning impairments in animals. Second, transgenic expression of NMDA receptor sub-units that boost the function of the receptor in adult animals can improve learning and memory (Tang *et al.*, 1999). Third, recent behavioural research indicates that the types of learning and memory affected depend critically on the brain region in which these pharmacological or molecular changes are expressed. Changes in the hippocampus seem to affect types of memory in animals that are at least similar to episodic memory in humans; changes in the amydgala affect value and emotional memory.

The consolidation of changes in synaptic weight has not escaped scrutiny; research

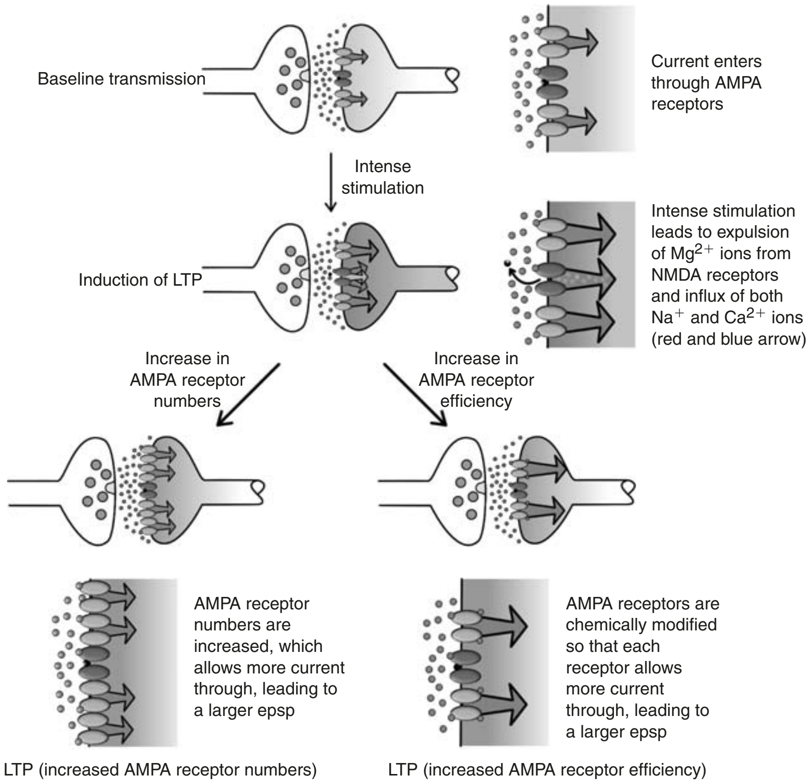

Baseline transmission
Current enters through AMPA receptors
Intense stimulation
Induction of LTP
Intense stimulation leads to expulsion of Mg2+ ions from NMDA receptors and influx of both Na+ and Ca2+ ions (red and blue arrow)
Increase in AMPA receptor numbers
Increase in AMPA receptor efficiency
AMPA receptor numbers are increased, which allows more current through, leading to a larger epsp
AMPA receptors are chemically modified so that each receptor allows more current through, leading to a larger epsp
LTP (increased AMPA receptor numbers)
LTP (increased AMPA receptor efficiency)

FIGURE 9.7 Cartoon model of how long-term potentiation (LTP) might induce changes in the number (*bottom left*) or effectiveness (*bottom right*) of receptors at excitatory synapses. In each cartoon, the different receptors and their associated ion-channels are shown as small ovals. Broadly speaking, NMDA receptors (dark grey) are the engines of change, AMPA receptors (light grey) the engines of expression of change. (Figure courtesy of A. Doherty and G.L. Collingridge (Bristol Neuroscience), 2003; reproduced with permission, this figure appears in the colour plate section)

has identified second-messenger pathways, transcription factors and genes whose protein products are implicated in making such synaptic changes last. Key molecular players include MAPkinase (mitogen activating protein kinase) and CREB (cyclic AMP response element binding protein).

We have known for some time that protein synthesis inhibitors selectively impair long-term memory. The race is now on to identify which proteins are critical, when and where they are synthesized, and how they realize their effects. This research matters because disruption of these biochemical mechanisms may be at the heart of certain kinds of age-related memory loss. Several start-up companies, such as those founded by Nobel Laureates Leon Cooper (Sention) and Eric Kandel (Memory Pharmaceuticals), are built around intellectual property related to compounds that affect synaptic plasticity and consolidation.

4.2 Place-cells, Spatial Memory and Event Memory

A second important discovery is that of place cells (O'Keefe, 1976). These are neurons in the hippocampus that fire

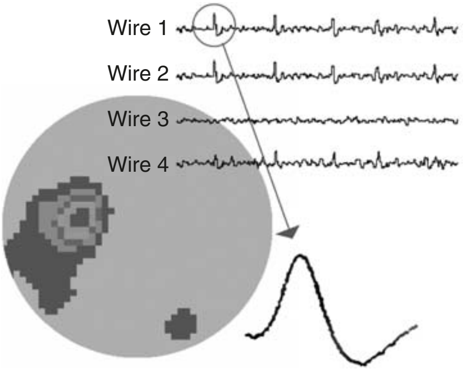


FIGURE 9.8 Place cells in the hippocampus of the awake, freely moving rat. The large grey circle shows an area of space through which the animal can move. The 'hot spot' is where the cell fires most strongly (which shows here as white). Extracellular recordings are typically made with a UK invention, the tetrode, which enables the neural activity signature of individual cells to be picked up and distinguished from that of neighbouring cells. (Based on work of O'Keefe and colleagues; drawn by Livia de Hoz and used with permission, this figure appears in the colour plate section)

action-potential spikes only when an animal, such as a laboratory rat, explores a specific part of a familiar environment (Fig. 9.8). Different cells code for different parts of the environment such that a population of cells is involved in mapping a whole area. Other cells in a nearby brain area code for the direction the animal is moving in (head-direction cells). The two areas working together – somehow providing a map of space and a sense of direction – help animals learn to find their way around the world. This is clearly very important, as finding food and water and then the way back to the burrow or nest, is vital for survival and a navigational mechanism for which there is very likely to have been considerable evolutionary pressure.

The initial observations on place and head-direction cells have been widely replicated in many laboratories, but opinion is divided about whether these neurons have a purely spatial function. Some hold that the evolution of the hippocampus in vertebrates, and particularly mammals, is intimately linked to spatial awareness and memory (O'Keefe and Nadel, 1978). Others argue that space is a special case of a more general memory function (Eichenbaum and Cohen, 2001). One contentious way of looking at this is to raise the possibility that spatial memory, no less than other forms of propositional memory, can be subdivided into its semantic and episodic subcomponents.

Animals form a stable representation of where landmarks are in their territory – just like the associative framework of other factual knowledge that humans acquire about our world (semantic memory). However, this map of space also provides a memory framework in which to anchor and so remember events – such as where a predator was last seen (episodic-like memory). If so, the continuing argument between a strictly spatial and a more episodic view of hippocampal function might be resolved. New behavioural tasks are being developed to help to address these issues, such as tasks in which animals are required to remember the what, where and when of events (Clayton and Dickinson, 1998; Morris, 2001).

4.3 Neuronal Development and Learning

We acquire much of what we know and many of our skills early in life. Neuroscientists now believe that many aspects of the fine tuning of neural connections in the developing brain, a process that depends on neural activity, are also used during early learning (activity-dependent self-organization of the brain). The developing brain is particularly plastic and its self-organizing mechanisms are used to fine-tune neural circuits vital for survival. The Fore-sight Research Review on Representation summarizes observations made on the way in which sensory afferents can determine cortical fate (Sur and Leamey, 2001).

The attachment that develops between an infant and its mother has been studied in young chicks that imprint on their mother.

First studied systematically by ethologists, elegant neurobiological experiments have shown where this learning process takes place in the young chick's brain and the cascade of chemical transmitters that are released to act on receptors involved in storing some kind of an 'image' of the mother.

In a systematic programme of research over 25 years, Horn and his colleagues (Horn *et al.*, 1973; Horn, 2000) have identified a particular brain region, the intermediate and medial part of the hyperstriatum ventrale (IMHV), as a site of storage for this information, and so made it possible to study the neural basis of memory. The changes that take place in the right and left IMHV are being studied in parallel using molecular, anatomical, pharmacological and electrophysiological methods. An important feature of this research is that quantitative measures of learning and memory are related to quantitative measures of the brain changes. This work includes studies of the neural basis of 'predispositions' to learn and computational modelling.

Similarly, young animals need to learn very quickly what foods are safe to eat through a process of one-trial learning. 'Taste-aversion learning' means that they rapidly develop an aversion in their first encounter with noxious foods (Rose, 1992). Young animals are genetically predisposed to be cautious about what they eat: learning ensures that they make a mistake about food only once. They taste only small amounts of food at a time, and register any foods that taste bad or are nauseous. They avoid these thereafter by associating the sight of an inappropriate foodstuff with its nauseous taste.

Neurobiological studies of both forms of early learning have revealed that, in addition to the role of glutamate receptors (such as the NMDA receptor), a cascade of downstream second-messengers transmit signals to the nuclei of brain cells where genes are activated to make the plasticity-proteins that consolidate memory. Cell-adhesion molecules are involved in shaping the connections that are being formed in the developing nervous system as the genetic program unfolds alongside experience.

What happens is not unlike an old-fashioned photographic process with an image first being captured by the initial pattern of neural activity, using NMDA and AMPA receptors, and then fixed (consolidated) by these later biochemical signals. The similarity between these mechanisms activated during early learning and those identified in the physiological phenomenon of long-term potentiation is striking. Both synaptic plasticity and early learning can go through clearly defined temporal stages of induction, expression and consolidation, though what determines whether they pass through these stages on any individual learning occasion is still not understood.

4.4 Open Questions

- *Neural network models of memory processes would benefit from greater biological realism at the level of representation, local circuits and mechanisms of plasticity*

Much current research on long-term potentiation is highly reductionist. Many neuroscientists embrace the view that only a molecular understanding at the level of identified proteins will do. Optical imaging techniques are coming to the fore as researchers endeavour to observe these changes in living tissues in real-time, partly because 'seeing is believing' and partly because optical techniques offer certain advantages over electrophysiological methods (as well as disadvantages).

However, this reductionist focus comes at a price. There is a gaping conceptual hole between the current preoccupation with molecular events at individual synapses and the complexities of understanding how the expression of synaptic plasticity within different circuits realizes different forms of memory. A full account of the neural mechanisms of memory would incorporate the astonishing molecular advances, with an equally sophisticated understanding of local circuits, of the physiological events happening within these circuits, and the connectivity between larger neuronal networks. Closer contact with computational scientists could foster a shift of emphasis.

- *There are both technical and computational problems associated with simultaneous recording from large numbers of cells.*

Much of the electrophysiological recording from neurons in the brain of awake animals to date used techniques that record one or, at most, only a few cells at one time. It has been apparent for some time that population coding is important for many functions (e.g. motor function) and that the secrets of how information is represented in specific brain areas and information transferred between brain areas will require the simultaneous recording of hundreds or even thousands of cells.

There are formidable technical and computational problems in doing this. The first steps have been taken, including the invention in the UK by O'Keefe and his colleagues of the 'stereotrode' and 'tetrode' that enable multiple single-cell recordings, and 'proof-of-principle' experiments in which recordings have been made from up to about 100 cells simultaneously (Wilson and McNaughton, 1993). Various electronics companies and other SMEs are positioned to take part in this research, including the UK company Axona, although the market may prove small given the specialized nature of the work. The Foresight Research Review of Advanced Neuroscience Techniques offers greater detail on this topic.

- *Are lessons to be learned about neurobiological mechanisms of neuronal plasticity that are applicable to wider issues such as drug addiction?*

It is striking that spatial learning in the hippocampus, emotional learning involving the amygdala and many aspects of early learning studied in the avian brain display overlapping mechanisms involving glutamate receptors and downstream second-messenger pathways. It would be a beautiful simplification (perhaps an over-simplification?) if common mechanisms of cell biology mediate different forms of learning and memory largely by virtue of the different neuronal circuits in which they are embedded. If this contention has any validity, we may be able to apply the lessons we learn from the three cases we have described to widely different forms of learning.

A practical case in point is drug-addiction. Here behaviour changes in well-described and predictable ways. Dopamine is involved, as well as glutamate receptors, playing a critical role in 'rewarding' drug-seeking behaviours that are initially flexible, goal-directed actions but gradually become inflexible, automatic habits (Everitt *et al.*, 2001; Hutcheson *et al.*, 2001).

5 NEURODEGENERATIVE DISEASES

Memory dysfunction is exceptionally prevalent in our society. It has major socio-economic implications. The annual cost of caring for patients with dementia in England and Wales has been estimated as high as £10 billion (Schneider *et al.*, 1993). Not only does memory ability decline as we age, but a number of different neurological conditions (e.g. stroke, viral infections, head injury, prolonged alcohol abuse, epilepsy and dementia) can result in memory impairment. Memory-impaired people often cannot cope with the demands of work, find it virtually impossible to live independently, and can become reclusive. The embarrassment of not recognizing someone we have met previously is familiar to us all, so it is easy to empathize with the difficulties facing the amnesic patient, Jack, reported by Wilson (1999), who writes:

> I can't perform basic sociable tasks, such as taking orders to buy a round of drinks or noting the names and faces of new

> acquaintances. In fact, I am sure that on many occasions I have met people who are not aware of my condition, and then upon not recognizing them on a second meeting, will have appeared rude and impolite.

It can be harder to imagine the more generalized impact of having a memory-impaired person in the family: the constant repetition of stories and questions, and the continual need for help with even the simplest of tasks (e.g. finding a wallet or noting down a telephone call), frequently strains relationships with family members and friends, who themselves feel intense frustration and mourn the loss of their own independence.

Helping to improve the quality of life for patient and carer alike is a major long-term goal of those interested in human memory. Centres such as the new Iris Murdoch Dementia Centre in Stirling are helping to focus on the needs of carers.

5.1 Memory in the Elderly

As they get older, virtually everyone reports that their minds are not quite what they used to be, with one of the most common complaints being difficulty with memory. As these types of memory slips are also common in Alzheimer's disease, the increasing presence of them in old age can cause tremendous worry.

Older adults show marked deficits in their ability to retain information for events over short time periods (e.g. hours and days), especially on tests of recollection compared to familiarity-based recognition. A possible explanation for this dissociation is reduced processing resources as opposed to memory failure *per se*. Recollection demands more attention than mere recognition. If older subjects are resource-depleted, they are likely to perform poorly on tasks that are more demanding and/or provide few cues to the correct answer. When recollection and familiarity-based recognition are equated for difficulty, however, age-related impairments can often be documented on both processes (Baddeley, 1996).

The memory problems evident in old age are much more widespread than simple episodic slips, also affecting semantic and working memory. Older adults are more likely to suffer from tip-of-the-tongue states, whereby they are unable to recall a sought-after word but have a strong feeling that they know what the word is. While young and old subjects all suffer from this annoying affliction, older subjects seem more susceptible, especially when searching for a proper name. This difficulty may reflect a failure to access phonological information about vocabulary, rather than a problem within semantic memory, as the elderly perform well on tests of knowledge of vocabulary and semantic concepts. It remains to be seen, however, whether older adults show normal semantic knowledge for categories that might be more vulnerable to age-related decline, such as those typically affected early in certain neurodegenerative conditions (e.g. knowledge of famous people).

It is generally accepted in the psychological community that performance on working memory, especially tasks that require the manipulation of information, is affected by increasing age (e.g. recalling a set of digits backwards). One contributing factor may well be greater sensitivity to interference from irrelevant and distracting information, and task manipulations that reduce the likelihood of interference can improve memory (e.g. a break between trials), albeit not to the same level as seen in younger adults.

5.2 Neurodegenerative Disease

One of the most frequent causes of memory loss is dementia, which is thought to affect some 5–8% of all people above the age of 65 years. This figure hides the dramatic increase with advancing age: the prevalence doubles every five years over the age of 65, reaching over 20% in 80-year-olds.

Although Alzheimer's disease is the most common cause of dementia, and most familiar to the general public, a number of other types of dementia also affect memory, sometimes in ways quite different than seen in Alzheimer's disease. The main causes of early onset dementia (prior to age 65) are frontotemporal dementia, vascular dementia and dementia as part of neurological conditions such as Huntington's and Creutzfeldt–Jakob disease (CJD). After age 65, three neurodegenerative conditions – Alzheimer's disease, vascular disease and dementia with Lewy bodies – account for virtually all cases. Definitive diagnosis is still possible only after death, Alzheimer's disease being identified by the presence of amyloid plaques and neurofibrillary tangles. The preliminary diagnosis, in life, includes dysfunction of at least two cognitive functions (i.e. not just memory) and gradual decline over time. The progressive loss of cognition that is the hallmark of dementia provides a unique opportunity to track the breakdown of memory systems longitudinally.

The profile of early memory loss in dementia typically reflects the brain regions that bear the brunt of early pathology. In Alzheimer's disease, the most consistent early neuropsychological symptom is increasing problems with the acquisition of new memories. Patients become repetitive (e.g. frequently asking what day it is or when someone is visiting), forget appointments and telephone calls, and may not remember to turn off the gas after cooking. As proposed previously, structures in the medial temporal lobe, in particular the hippocampus, perirhinal and entorhinal cortex, seem critical for new learning.

Consistent with this, patients with Alzheimer's disease have atrophy in these regions, in particular in the entorhinal cortex and hippocampus (Braak and Braak, 1991). While episodic memory is the trademark cognitive impairment in Alzheimer's disease, other types of memory such as working and semantic memory are also affected later in the disease, a finding that presumably reflects the spread of pathology into more lateral temporal and frontal regions. Notably, however, the degree of structural loss seen in medial temporal lobe regions is somewhat out of step with the profound impairment to episodic memory evident early in the disease (Galton *et al.*, 2001).

New findings using positron emission tomography (PET), in which it is possible to measure which brain regions are underfunctioning in patients with dementia, have helped address this issue. In these investigations, the earliest site of poor functioning (as measured by glucose metabolism) was the posterior cingulate cortex, which is part of a brain network, including the hippocampus, mamillary bodies and thalamus, that supports episodic memory (Nestor *et al.*, 2002). This result highlights the need for cognitive neuroscientists to think beyond individual brain areas and to consider how networks interact in the acquisition and storage of human memory. It also implies that drug therapies for Alzheimer's disease may be more effective if they target the initial site of pathology, as opposed to areas that, while clearly involved in episodic memory, may be hypometabolic because they are anatomically downstream.

Not all dementia results in episodic memory impairment. In the temporal variant of frontotemporal dementia, also called semantic dementia, the predominant deficit is to semantic memory. Patients with semantic dementia typically present with difficulties in word finding and comprehension. A spouse might report: I asked John to put the kettle on the other day, and he responded, 'Kettle, kettle, what is a kettle?' By contrast, individuals with semantic dementia can remember where they left the car in the supermarket car park, rarely get lost in new environments, or forget appointments. Not only is episodic memory good in semantic dementia, but also basic visuo-perceptual skills and working memory remain preserved even at late in the disease. These kinds of dissociation are logically puzzling: how can someone not know what a kettle is, yet find their car in a car park. But they exist.

The locus of pathology in semantic dementia is within anterior and inferior temporal lobe regions that are unaffected early in Alzheimer's disease. While there is involvement of medial temporal lobe structures, this is typically asymmetrical (left to right) and involves perirhinal cortex more than entorhinal cortex. Findings from semantic dementia, and other diseases that involve anterior temporal regions (e.g. encephalitis), suggest that retrieval of semantic memory, or at least of knowledge that has already been consolidated via repeated exposure, does not depend upon the hippocampus. There is some question about the role of non-hippocampal medial temporal lobe structures in semantic memory, such as the perirhinal cortex. If patients with semantic dementia show damage to perirhinal cortex, as seems likely, it is not clear why recognition memory and aspects of perceptual functioning are so good in the disease. Studies in these cases, especially those that utilize the same paradigms adopted by animal lesion researchers, are necessary to address this issue.

Semantic dementia is probably caused by the same underlying pathology as the frontal variant of frontotemporal dementia, in which patients present with marked behavioural and personality changes. While case J.W. reported no particular problems with cognition or behaviour when he presented in 1992, his wife reported personality changes over several years, including increased rigidity, disinhibition, reduced concern for others, and poor financial judgement.

The main locus of brain pathology in the frontal variant of frontotemporal dementia is thought to be in the orbitofrontal cortex. Orbitofrontal cortex is part of the brain network involved in emotion and social behaviour (which also includes the amygdala). Memory is not entirely spared in frontotemporal dementia: Simons *et al*. (2002) found that a group of such patients were unable to remember in which of two sets a picture was initially presented (so called source memory), despite normal memory for these items compared to novel pictures (familiarity). This type of contextual recollection depends upon interactions between frontal and medial temporal lobe regions in healthy controls, and is also susceptible to normal aging (Schacter *et al.*, 1991).

Further studies in frontal variant frontotemporal dementia, therefore, may allow us to tease apart how brain networks interact to support different components of memory, and to determine whether psychological subcomponents of episodic memory, such as familiarity and recollection, depend upon different regions of the brain.

We know less about the memory deficits seen in other dementias, such as vascular dementia and dementia with Lewy bodies (common causes of late onset dementia). At its broadest, the term vascular dementia corresponds to individuals who show cognitive decline in association with cerebrovascular disease, as measured by evidence of infarcts on brain imaging.

Although findings have been inconsistent across studies – predominantly due to poor matching of patient groups and difficulties in diagnostic criteria – episodic memory is impaired in vascular dementia, although probably not as severely as seen in Alzheimer's disease. By contrast, executive problems (e.g. planning and shifting attention) are quite prominent in vascular dementia, perhaps more so than in Alzheimer's disease. The difficulties patients with vascular dementia have on tests of episodic memory may be due to problems with retrieving and manipulating information from memory, rather than a particular deficit in the acquisition and maintenance of memory representations. If so, patients with vascular dementia will also be impaired on other types of memory, such as semantic memory, in particular when the task involves a strategic component – e.g. category fluency, in which subjects are asked to produce as many animal exemplars as possible in a minute.

Dementia with Lewy bodies has also been little investigated. The predominant deficit in this condition is thought to be in early visual processing, supposedly in the context of good episodic and semantic memory. PET studies measuring glucose metabolism have found underfunctioning of occipital and primary visual cortex. Notably, however, studies of the distribution of Lewy body pathology in patients with dementia with Lewy bodies have shown involvement of a range of regions, including subcortical areas (e.g. substantia nigra), medial temporal lobe regions and temporal cortex.

The aspects of memory (e.g. verbal episodic memory) that are preserved in the disease remain to be investigated more systematically, as presumably impairments to visual processing are likely to preclude good visual memory.

5.3 Open Questions

- *A theoretical framework for understanding age-related memory loss.*

The cause of memory deficits in ageing remains little understood. It is unclear whether the profiles seen in elderly individuals are due to a general disruption of processes necessary for memory (e.g. attention) or of memory mechanisms *per se*. Impairments to both episodic and working memory seem most consistent with a loss of executive control over cognitive processes. The literature suffers, however, from a need for greater theoretical specification. What do we mean when we talk about executive functioning? How does this system oversee retrieval and organization of material across different types of memory? Broadly, what impact does frontal damage have on profiles of memory loss across dementing conditions?

- *Fractionating the dementias and the implications for carers and for treatment*

It has long been assumed that patients with dementia were too impaired globally to provide unique insights about memory. There has been a major shift in this view over the past decade which continues to have profound implications for the way we think about clinical disorders of memory.

We now know that not all dementia is the same, just as not all memory is identical. Knowing this does not equate to understanding, however, and we need to learn how to recognize the unique cognitive signature of different dementias, and to think about how to distinguish these. Dementia is also unique in allowing us to watch memory breakdown and, as it does so, to glimpse sections of the underlying framework. Typicality seems critical for semantic memory, but what other types of glue bind everything together?

- *What impact will the study of transgenic animals targeting Alzheimer's disease have on understanding the neurobiology of disease and the development of new therapeutics?*

A particularly important development is the engineering of transgenic animals, usually mice, in which one or more of the human mutations that have been identified as occurring in familial forms of Alzheimer's or other neurodegenerative diseases is artificially overexpressed. Animals have now been engineered with alteration in amyloid precursor protein (APP) and the presenilin family of genes (PS1 and PS2), and found to develop aspects of the characteristic pathology, although a 'complete' animal model remains to be developed.

Securing a valid animal model opens the possibility of exploring the neurobiology of the disease process, using longitudinal studies, and coupling behavioural observations with end-state pathology, *in vivo* structural MRI, and electrophysiological measurements. Work with these mice has already revealed that they show progressive cognitive dysfunction. This opens the possibility of treating these mice with novel therapeutics, such as vaccination or drugs that target the secretase sites at which APP is cleaved to make the toxic 1–42 moiety of amyloid beta peptide.

6 CAN WE IMPROVE MEMORY?

There is much to learn about learning. Neuroscientists hope that by doing so, they will find a way to make learning easier and to make memories last longer. It is natural to think that it would be good to improve rate of learning or persistence of memory. Older people often complain about memory, and busy executives. Young people facing examinations might also find enticing the prospect of a memory pill or 'smart-drug'. The likely market for such drugs has not gone unnoticed by the pharmaceutical industry.

There are benefits, to be sure, but we need to take into account a number of problems when attempting to improve memory in the young, to alleviate stress-induced or even benign forgetfulness in the normal, or to remediate memory problems in the elderly or those with brain damage. Rose, (2002) provides a thoughtful summary of the issues.

6.1 Pharmaceuticals, Prospects and Perils

Improving memory would be valuable in certain situations, but would almost certainly come at a price. We are not thinking here of the usual side-effects of pharmaceuticals. Rather, the price we have in mind is that a good memory is necessarily a balance between remembering and forgetting. Were we to improve memory, we might then have difficulty forgetting trivial things that happened during the day that there is no need to remember. The 'yin and yang' of a good memory is one that remembers and organizes the right things in the brain, but forgets the less important things of life. One should never forget the tragic life of Luria's 'mnemonist' (Luria, 1969). The magic bullet of a pill that can improve memory, at least in normal people, probably does not exist because evolution has ensured that the system is well adapted to the environment in which we, or at least our ancestors, live. That said, the environment in which we live is very different from the African savannah and an argument can be made for pharmaceutical fine-tuning to tackle the vicissitudes of modern life.

What are the prospects? Drugs that act as agonists at the NMDA receptor or otherwise boost the cascade of downstream second-messenger signals in neurons may be useful in alleviating the benign forgetfulness seen in neurologically normal people such as those in stressful occupations (Lynch, 2002). We have also seen that certain neurodegenerative diseases, such as Alzheimer's disease, cause damage to brain tissue where memories get encoded and stored.

Clearly it would help to find some way of, first, stemming the course of these diseases, and second, restoring normal brain function for longer. With the population demography of virtually all developed countries veering towards a greater preponderance of older people, treatments that could help those at risk lead independent lives for longer would be greatly valued (and economically beneficial).

There has been some progress in identifying drugs that boost acetylcholine neurotransmission and so-called nootropic drugs that seem to target AMPA receptors. Researchers are actively pursuing other promising leads based on better understanding of the neurobiology of disease. The availability of pharmaceutical agents is likely to increase over the next 20 years. Despite this, benefits from these types of drug may be limited. For example, if we accept that the mild memory lapses seen in older adults are not a direct consequence of weaker memory representations, but instead relate to a reduction in the amount of executive resources available, a drug that enhanced memory encoding might be only minimally effective. Instead, the pharmaceutical approach should aim to boost attention or the functioning of the central executive. This is feasible, but it will require the pharmaceutical industry to rethink its approach.

Targeting the right cognitive impairment is a key problem in attempting to improve the memory deficits seen in dementia. Clearly we would wish to give drug therapies to patients as early as possible, yet we are currently unable to identify the different types of dementia in their presymptomatic stages with any consistency. Certain neuropsychological tests, such as memory for the object-locations, look promising (Swainson *et al.*, 2001), particularly in distinguishing individuals at risk of developing Alzheimer's disease from people with mild memory difficulties caused by depression.

The diagnostic obstacle is partly because we need large, costly, community-based studies to identify early cognitive markers of dementia, and also because much of the work to date on early diagnosis has concentrated on discriminating between Alzheimer's disease and normal aging. As we have seen, however, there are distinctions to be made in dementia. Drugs aimed specifically to target the memory problems in Alzheimer's disease are unlikely help patients with semantic dementia who do not suffer the episodic memory loss characteristic of Alzheimer's disease.

6.2 Cognitive Engineering

Given these are not insignificant hurdles, many cognitively oriented neuroscientists believe that we will need cognitive engineering alongside pharmaceutical engineering. You do not read so much about cognitive engineering in the newspapers as about new drugs, but ill-informed journalism does not make it any less important.

The idea is to take advantage of what has been learned about how information is encoded, stored, consolidated and retrieved – as described in this summary. The application of cognitive engineering takes willpower to put its principles into action; it's harder than just taking a pill – but there are grounds to think that it may work better.

Experiments have shown that better encoding of information at the time of learning improves memory – by thinking about things carefully in an attentive manner and establishing connections with established knowledge. Psychologists call this the encoding-specificity principle.

Spacing of learning sessions helps long-term memory; neuroscientists link this to the conditions of relevant gene-activation and the synthesis of plasticity-proteins. Frequent reminders ensure that information is retrieved at a time when memories are fragile, and then associatively interconnected in semantic memory. This engages the consolidation process: neural-network engineers will readily link this to the process of effective interleaving of information into distributed networks.

Recent rehabilitative work using errorless learning in patients with memory loss is a good example of how it can be help to understand the mechanisms by which we acquire and retain new memories. This approach stemmed from studies of learning in animals. It was pioneered by Baddeley and Wilson (1994), who explored whether memory-impaired patients would learn better if they were prevented from making mistakes during the learning process. The scientific rationale behind this hypothesis was that amnesic patients are much more likely to rely upon implicit memory – a system that readily acquires information but does not discriminate between what is right or wrong (all responses being equally familiar). The initial study by Baddeley and Wilson revealed better learning in amnesic patients who were encouraged to pursue an errorless learning strategy.

Researchers interested in memory rehabilitation were quick to take up this technique. It has also been shown to be successful, at least in the short-term, in patients with Alzheimer's disease.

Case V.J., who was profoundly amnesic, could only name on average 2–3 members of his social club. After training, however, V.J. could name all 11 members and astonishingly, maintained this level of performance for up to 9 months (Clare *et al.*, 1999).

Some patients with semantic dementia show a different profile. Case D.M. attempted to learn new vocabulary by practising with a children's dictionary at home. D.M. showed a remarkable improvement in his ability to produce words in response to a category label after practice (even outperforming controls). Disappointingly, he could not sustain his new level of performance without continued practice. Although D.M. was also using an errorless approach, he may have shown a different pattern to V.J. because he did not have a fully functioning semantic system in which to embed his new learning (Graham *et al.*, 1999). Interpreting the differences between these two patients would be impossible without knowing something about the relationship between episodic and semantic memory.

Recognizing the operating principles of the different types of memory is also essential – you will never learn a skill by merely hearing about it, even though this works fine for episodic memory. The use of external memory aids in amnesic patients has been relatively unsuccessful until recently, predominantly because their functions are complicated to learn. As Wilson (2003) aptly notes: 'the very people who need external memory aids are often the people who are most likely to have difficulty in learning how to use them'.

To circumvent this problem, Hersh and Treadgold (1994) developed a paging system (NeuroPage), in which a schedule of personalized reminders and cues is entered into a computer and subsequently, sent to the pager on a designated day and time. The device is simple and portable – the user has only to learn to operate one large button (to confirm receiving the message).

Two clinical studies of the efficacy of NeuroPage have shown that many amnesic patients, and their families, can benefit significantly from this approach. For example, (Evans *et al.*, 1998) report a patient, R.P., who had suffered a stroke seven years earlier, and had problems carrying out tasks. R.P. would spend too much time in the bath because she kept forgetting where she was in a sequence of washing. Using a checklist and NeuroPage, R.P. could break free from her stereotyped routines and bath in a reasonable time. This subsequently allowed her to attend a day centre, which had been impossible before, thereby releasing her husband from the daily monotony of constant reminders.

The use of technological aids such as pagers and computers in memory rehabilitation is likely to increase, in part due to developments in information technology but also as there are more memory-impaired people in the population. The main disadvantage of a scheme like NeuroPage is that it requires a central computing resource to organize, manage and transmit the messages, taking away some of the inherent flexibility in these types of systems.

6.3 Open Questions

- *New drugs or cognitive engineering or both?*

It remains to be seen how useful and cost-effective cognitive engineering strategies will be in the future. The successful modification of external aids, such as pagers, is heartening for families and individuals with memory impairment, but serious issues remain to be tackled if everyone is to have access to this type of system. What are the technical implications of increasing the scale of the project? What patients will best benefit from such a device and how do we measure improvement? How would we measure value for money? Should it be in terms of the benefit reported by the individual or via some monetary cost–benefit function calculated by the National Health Service?

New pharmaceuticals will play an ever-more important role in an ageing society. With respect to dementia, for which there are pharmaceutical agents available already, the pace of research is such that the drugs we have at present, which are at best only moderately successful, will be replaced by more selective compounds with improved efficacy. The situation is likely to change gradually but dramatically over the next two decades. The impact of the Human Genome Project, a developing understanding of proteins and protein-networks, should lead to a new pharmacology of cognition. However, we suspect that to utilize new drugs effectively, we will need to be more sophisticated in how we characterize memory impairment, taking account of the fact that different components of memory can be differentially impaired. The advent of smart drugs is also likely to raise moral questions about whether they should be available to healthy individuals.

7 COMPUTATIONAL MODELLING

An ambition of the Cognitive Systems Foresight Project is to bring together biologists, physical scientists and mathematicians. This raises a major communication problem, related to the very different semantic (i.e. professional) memories we each possess given our diverse training backgrounds. It will take time to overcome this communication problem, even in circumstances in which there is serious intent to do so by all parties. However, computational modelling may provide a common language with which the parties may communicate. Indeed, it has the potential to be an invaluable tool in understanding the relationship between brain and behaviour because it addresses not only what functions are performed by brain regions, but how they are performed.

Whilst most researchers have a theoretical framework that informs and motivates their experiments, often the particular mechanisms that could yield predicted results are not made explicit. Different kinds of computational modelling can be particularly useful in that they bring forward such implicit assumptions. Models can be important tools for the analysis of data and can help in the development of clearly motivated experiments with clear predictions. Questions that arise from this endeavour include: What might non-biologists provide biologists with respect to algorithms? What might neuroscientists provide with respect to knowledge of network architecture and cell-properties?

7.1 Small Networks with Biologically Realistic Parameters

Our central theme of multiple memory systems points to the possibility of building different kinds of models of memory function. Some might be guided by psychological rather than neurobiological constraints – as in models of working-memory. Others might be guided by explicit knowledge about networks, such as one based on the anatomy and physiology of the hippocampus, perirhinal cortex or cerebellum.

It will be valuable to explore the capabilities of such networks, and to use these networks to create specific predictions that can be tested in neurobiological and neuropsychological studies. What control mechanisms are required to maintain network stability? How well do they distinguish signal and noise? What would be the implications of different types of synaptic modification rules (Hebbian and non-Hebbian)? There is a great deal of existing work on this, including studies establishing that storage capacity can be enhanced by having both LTP and LTD (Willshaw and Dayan, 1990) or be sensitive, in familiarity discrimination networks, to the combination of learning rule and decision function (Bogacz and Brown, 2002).

7.2 Connecting Networks

An important theme in much current human brain imaging is connectivity between networks. In this and in other ways, the focus is shifting from an emphasis on individual brain centres towards how different brain regions cooperate and compete in controlling behaviour. This leads naturally to the issue, to be addressed in computational models, of whether and how small networks could be connected together into a working system. What feed-back/feed-forward is essential? Are there temporal constraints? What are the constraints on information-carrying capacity of having (or not having) synchronized firing and oscillation? Bogacz and Brown *et al.* (2001) present arguments based on computational modelling for why you would not want to perform familiarity discrimination (judgement of prior occurrence) and feature detection (categorization) in the same network – though it is good to have them in the same brain area. What does computational modelling have to say about Nadel and Moscovitch's (1997) model of episodic memory storage in which multiple traces are formed in an overlaid manner?

The neuropsychological literature is at an impasse with respect to understanding memory consolidation. There is a real need for greater theoretical specificity. And, linking this in to one of our neurobiology examples, is it possible to build a memory system for imprinting where the responsiveness of individual neurones is radically variable across time (Horn *et al.*, 2001)?

7.3 Neuroinformatics

Help across the life-sciences/physical sciences divide will also be important in neuroinformatics. The effectiveness of different research groups to share data is still very limited. Important steps are being taken in the brain-imaging domain, but it is striking how little data is currently shared.

8 OPEN QUESTIONS: A SUMMARY

We list here the 'open questions' that we have identified in our summary of the state of the art (the appropriate section number of the chapter is given in brackets). It is inevitably a personal list. We recognize that others would create different lists or put emphasis elsewhere. However, we hope that it will be provocative.

8.1 Definitions, Concepts, Techniques

- Are learning and memory really distinct from perception, attention or motivation? (2.3)
- Memory in the digital age? (2.3)

8.2 The Organization of Memory

- Do we need to invoke a 'central executive' in working-memory? (3.1.5)
- The concept of multiple memory systems, while apparently widely accepted, is coming under review. What is the relationship between different memory systems, for example episodic, perception and semantic memory? What are the inputs to working-memory from semantic or episodic-memory? (3.2.7)
- Characterizing amnesia and its anatomical substrate. (3.2.7)
- Perception and familiarity-based recognition memory. (3.2.7)
- Semantic memory and connectionist modelling. (3.2.7)
- The binding problem and memory (3.2.7)
- Computational modelling and animal studies will help us better understand the mechanisms of memory consolidation. (3.2.7)

8.3 The Neurobiology of Memory

- Neural network models of memory processes would benefit from greater biological realism at the level of representation, local circuits and mechanisms of plasticity. (4.4)
- There are both technical and computational problems associated with simultaneous recording from large numbers of cells. (4.4)

There may be lessons learned about neurobiological mechanisms of neuronal plasticity that are applicable to wider issues such as drug addiction. (4.4)

8.4 Neurodegenerative Diseases and Cognitive Dysfunction

- A theoretical framework for understanding age-related memory loss. (5.3)
- Fractionating the dementias and the implications for carers and for treatment (5.3)

8.5 Can We Improve Memory?

- New drugs or cognitive engineering or both? (6.3)

8.6 Computational Modelling

- How can we best sustain communication across boundaries created by different patterns of professional training? (7.2)

References

Adolphs, R., Tranel, D., Damasio, H. and Damasio, A.R. (1995) Fear and the human amygdala. *J. Neurosci.*, 15: 5879–5891.

Aggleton, J.P. and Brown, M.W. (1999) 'Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behav. Brain Sci.*, 22: 425–489.

Andrade, J. (2001) The contribution of working memory to conscious experience. In J Andrade (ed.), *Working Memory in Perspective*. Hove: Psychology Press, pp. 60–78.

Baars, B. (1997) Some essential differences between consciousness and attention, perception, and working memory. *Consciousness Cogn.*, 6: 363–371.

Baddeley, A. (1992) Working memory: the interface between memory and cognition. *J. Cogn. Neurosci.*, 4: 281–288.

Baddeley, A. (1996) Applying the pyschology of memory to clinical problems. In D.J. Herrmann, C.L. McEvoy, C. Hertzog, P. Hertel and M.K. Johnson (eds), *Basic and Applied Memory Research: Theory in Context.* Mahwah, NJ: Erlbaum, pp. 195–219.

Baddeley, A.D. (2000) The epsiodic buffer: a new component of working memory? *Trends Cogn. Sci.*, 4: 417–423.

Baddeley, A.D. and Andrade, J. (2000) Working memory and the vividness of imagery. *J. Exp. Psychol. Gen.*, 129: 126–145.

Baddeley, A.D. and Hitch, G.J. (1974) Working memory. In G. Bower (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory.* New York: Academic Press, pp. 47–89.

Baddeley, A.D. and Wilson, B.A. (1994) When implicit learning fails: amnesia and the problem of error elimination. *Neuropsychologia*, 32: 53–68.

Baddeley, A., Gathercole, S.E. and Papagno, C. (1998) The phonological loop as a language learning device. *Psychol. Rev.*, 105.

Baddeley, A., Conway, M. and Aggleton, J.P. (2002) *Episodic Memory: New Directions in Research*. Oxford: Oxford University Press.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press/Bradford Books.

Bliss, T.V. and Lomo, T. (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol.*, 232: 331–356.

Bliss, T.V.P., Collingridge, G.L. and Morris, R.G.M. (2004) Long term potentiation: enhancing neuroscience for 30 years. Oxford: Oxford University Press.

Braak, H. and Braak, E. (1991) Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol.*, 82: 239–259.

Brown, M.W., Wilson, F.A.W. and Riches, I.P. (1987) Neuronal evidence that inferomedial temporal cortex is more important than hippocampus in certain processes underlying recognition memory. *Brain Res.*, 409: 158–162.

Bogacz, R., Brown, M.W. and Giraud-Carrier, C. (2001) Model of familiarity discrimination in the perirhinal cortex. *J Comput Neurosci* 10: 5–23.

Buckley, M.J. and Gaffan, D. (1998) Perirhinal cortex ablation impairs visual object identification. *J. Neurosci.*, 18: 2268–2275.

Buffalo, E.A., Ramus, S.J., Squire, L.R. and Zola, S.M. (2000) Perception and recognition memory in monkeys following lesions of area TE and perirhinal cortex. *Learn. Mem.*, 7: 375–382.

Burgess, N. and Hitch, G.J. (1999) Memory for serial order: a network model of the phonological loop and its timing. *Psychol. Rev.*, 106.

Bussey, T.J. (2004) Multiple memory systems. *Q. J. Exp. Psychol.*, 57: 89–94.

Bussey, T.J. and Saksida, L.M. (2002) The organization of visual object representations: A connectionist model of effects of lesions in perirhinal cortex. *Eur. J. Neurosci.*, 15, 355–364.

Bussey, T.J., Saksida, L.M. and Murray, E.A. (2002) Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *Eur. J. Neurosci.*, 15: 365–374.

Bussey, T.J., Saksida, L.M. and Murray, E.A. (2003). Impairments in visual discrimination after perirhinal cortex lesions: Testing 'declarative' versus 'perceptual-mnemonic' views of perirhinal cortex function. *Eur. J. Neurosci.*, 17: 649–660.

Churchland, P.S. and Sejnowski, T.J. (1992) *The Computational Brain*. Cambridge, MA: MIT Press.

Clare, L., Wilson, B.A., Breen, K. and Hodges, J.R. (1999) Errorless learning of face-name associations in early Alzheimer's disease. *Neurocase*, 5: 37–46.

Clayton, N.S. and Dickinson, A. (1998) Episodic-like memory during cache recovery by scrub jays. *Nature*, 395: 272–274.

Collingridge, G.L., Kehl, S.J. and McLennan, H. (1983) Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus. *J. Physiol.*, 334.

Collins, A.M. and Quillian, M.R. (1969) Retrieval time from semantic memory. *J. Verbal Learning Verbal Behav.*, 8: 240–247.

Conway, M.A., Anderson, S.J., Larsen, S.F., Donnelly, C.M., McDaniel, M.A., McClelland, A.G.R., Rawles, R.E. and Logie, R.H. (1994) The formation of flashbulb memories. *Memory Cogn.*, 22: 326–343.

Conway, M.A., Pleydell-Pearce, C.W., Whitecross, S. and Sharpe, H. (2002) Brain imaging of autobiographical memory. *Psychol Learning Motivation: Adv. Res. Theory*, 41: 229–263.

Corkin, S. (2002) What's new with the amnesic patient H.M.? *Nature Rev. Neurosci.*, 3: 153–160.

Cowan, N. (1999) An embedded-process model of working memory. In A.M.P. Shah (ed.), *Models of Working Memory: Mechanisms of Active Maintenace and Executive Control.* Cambridge: MA: MIT Press.

D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S. and Grossman, M. (1995) The neural basis of the central executive system of working memory. *Nature*, 16: 279–281.

Dean, P. (1976) Effects of inferotemporal lesions on the behaviour of monkeys. *Psychol. Bull.*, 83: 41–71.

Dickinson, A. (1980) *Contemporary Animal Learning Theory*. Cambridge: Cambridge University Press.

Eichenbaum, H. and Cohen, N.J. (2001) *From Conditioning to Conscious Recollection.* New York: Oxford University Press.

Ellis, A.W. and Lambon-Ralph, M.A. (2000) Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *J. Exp. Psychol. Learning Memory Cogn.*, 26: 1103–1123.

Ellis, A.W., Young, A.W. and Critchley, E.M.R. (1989) Loss of memory for people following temporal lobe damage. *Brain*, 112: 1469–1483.

Emery, N.J.and Amaral, D.G. (2000) Role of the amygdala in primate social cognition. In R.D. Lane and L. Nadel (eds), *The Cognitive Neuroscience of Emotion*. Oxford: Oxford University Press, pp. 156–191.

Engle, R.W., Tuholski, S.W., Laughlin, J.E. and Conway, A.R.A. (1999) Working memory, short-term memory and general fluid intelligence: a latent variable approach. *J. Exp. Psychol. General*, 128: 309–331.

Ericsson, K.A. and Kintsch, W. (1995) Long-term working memory. *Psychol. Rev.*, 102: 211–245.

Evans, J.J., Emslie, H. and Wilson, B.A. (1998) External cueing systems in the rehabilitation of executive impairments of action. *J. Int. Neuropsychol. Soc.*, 4: 399–408.

Everitt, B.J., Dickinson, A. and Robbins, T.W. (2001) The neuropsychological basis of addictive behaviour. *Brain Res. Rev.*, 36: 129–138.

Fell, J., Klaver, P., Elfadil, H., Schaller, C., Elger, C.E. and Fernandez, G. (2003) Rhinal-hippocampal theta coherence during declarative memory formation: interaction with gamma synchronization? *Eur J Neurosci*, 17: 1082–1088.

Fletcher, P.C. and Henson, R.N. (2001) Frontal lobes and human memory: insights from functional neuroimaging. *Brain*, 124: 849–881.

Frackowiak, R.S.J., Friston, K.J., Frith, C.D., Dolan, R.J. and Mazziotta, J.C. (1997) *Human Brain Function*. London: Academic Press.

Gaffan, D. (2002) Against memory systems. *Phil. Trans. R. Soc. Lond. B*, 357: 111–1121.

Gaffan, D. (1994) Scene-specific memory for objects: a model of episodic memory impairment in monkeys with fornix transection. *J. Cogn. Neurosci.*, 6: 305–320.

Gaffan, D. and Hornak, J. (1997) Amnesia and neglect: beyond the Delay–Brion system and the Hebb synapse. *Phil. Trans. R. Soc. Biol. Sci.*, 352: 1481–1488.

Galton, C.J., Patterson, K., Graham, K.S., Lambon-Ralph, M., Williams, G., Antoun, N., Sahakian, B.J. and Hodges J.R. (2001) Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology*, 57: 216–225.

Goldman-Rakic, P.S. (1992) Working memory and the mind. *Sci. Am.*, 267: 110–117.

Goldman-Rakic, P.S. (1995) Cellular basis of working-memory. *Neuron*, 14: 477–485.

Goodale, M.A. and Milner, A.D. (1992) Separate visual pathways for perception and action. *Trends Neurosci*, 15: 20–25.

Graham, K.S. and Hodges, J.R. (1997) Differentiating the roles of the hippocampal complex and the neocortex in long-term memory storage: evidence from the study of semantic dementia and Alzheimer's disease. *Neuropsychology*, 11: 77–89.

Graham, K.S., Patterson, K., Pratt, K.H. and Hodges, J.R. (1999) Relearning and subsequent forgetting of semantic category exemplars in a case of semantic dementia. *Neuropsychology*, 13: 359–380.

Graham, K.S., Simons J.S., Pratt, K.H., Patterson, K. and Hodges, J.R. (2000) Insights from semantic dementia on the relationship between episodic and semantic memory. *Neuropsychol.*, 38: 314–324.

Gregory, R.L. (1961) The brain as an engineering problem. In W.H. Thorpe, O.L. Zangwill (eds), *Current Problems in Animal Behaviour*. Cambridge: Cambridge University Press.

Griffiths, D., Dickinson, A. and Clayton, N. (1999) Episodic memory: what can animals remember about their past? *Trends Cogn. Sci.*, 3: 74–80.

Hall, J., Parkinson, J.A., Connor, T.M., Dickinson, A. and Everitt, B.J. (2001) Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. *Eur. J. Neurosci.*, 13: 1984–1992.

Hanley, R., Young, A.W. and Pearson, N.A. (1991) Impairment of the visuo-spatial sketchpad. *Q. J. Exp. Psychol.*, 43.

Hebb, D.O. (1949) *The Organization of Behaviour.* New York: Wiley.

Henson, R. (2001) Neural working memory. In J. Andrade (ed.), *Working Memory in Perspective*. Hove: Psychology Press, pp. 151–173.

Hersh, N.A. and Treadgold, L.G. (1994) *NeuroPage: The Rehabilitation of Memory Dysfunction by Prosthetic Memory and Cueing*. San Jose, CA: Hersh and Treadgold.

Horn, G. (2000) Memory. In J.J. Bolhuis (ed.), *Brain, Perception, Memory: Advances in Cognitive Neuroscience.* Oxford: Oxford University Press, pp. 329–363.

Horn, G., Nicol, A.U. and Brown, M.W. (2001) Tracking memory's trace. *Proc Natl Acad Sci USA*, 98: 5282–5287.

Horn, G., Rose, S.P.R. and Bateson, P.P.G. (1973) Experience and plasticity in the nervous system. *Science*, 181: 506–514.

Hutcheson, D.M., Everitt, B.J., Robbins, T.W. and Dickinson, A (2001) The role of withdrawal in heroin addiction: enhances reward or promotes avoidance? *Nature Neurosci.*, 4: 944–947.

James, W. (1890) *Principles of Psychology*. New York: Dover Press.

Jarrold, C., Baddeley, A.D. and Hewes, A.K. (1999) Genetically dissociated components of working memory: evidence from Down's and Williams syndrome. *Neuropsychologia*, 37: 637–651.

Kluver, H. and Bucy, P.C. (1939) Preliminary analysis of functions of the temporal lobes in monkeys. *Archives of Neurology and Psychiatry*, 42: 979–1000.

LeDoux, J.E. (2000) Emotion circuits in the brain. *Ann. Rev. Neurosci.*, 23: 155–184.

Luria, A.R. (1969) *The Mind of a Mnemonist.* London: Cape.

Lynch, G. (2002) Memory enhancement: the search for mechanism-based drugs. *Nature Neurosci.* November 5 Supplement 1, 1035–1038.

Malkova, L., Gaffan, D. and Murray, E.A. (1997) Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *J. Neurosci.*, 17: 6011–6020.

Mandler, G. (1980) Recognizing: the judgement of previous occurrence. *Psychol. Rev.*, 87: 252–271.

Marr, D. (1971) Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond. B*, 262: 24–81.

Marr, D. (1982) *Vision.* San Francisco: W.H. Freeman.

Martin, S.J., Grimwood, P.D. and Morris, R.G.M. (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. *Ann. Rev. Neurosci.*, 23: 649–711.

McGaugh, J.L. (2000) Memory – a century of consolidation. *Science*, 287: 248–251.

Meunier, M., Bachevalier, J., Murray, E.A., Malkova, L. and Mishkin, M. (1999) Effects of aspiration versus neurotoxic lesions of the amygdala on emotional responses in monkeys. *Eur. J. Neurosci.*, 11: 4404–4418.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A. and Wager, T.D. (2000) The unity and diversity of executive functions and their contribution to complex 'frontal lobe' tasks: a latent variable analysis. *Cogn. Psychol.*, 41: 49–100.

Morris, R.G. (2001) Episodic-like memory in animals: psychological criteria, neural mechanisms and the value of episodic-like tasks to investigate animal models of neurodegenerative disease. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 356: 1453–1465.

Morris, R.G.M. and Frey, U. (1997) Hippocampal synaptic plasticity: role in spatial learning or the automatic recording of attended experience? *Phil. Trans. R. Soc. Lond. Ser. B: Biol. Sci.*, 352: 1489–1503.

Morris, R.G.M., Anderson, E., Lynch, G.S. and Baudry, M. (1986) Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature*, 319: 774–776.

Nadel, L. (1992) Multiple memory systems: what and why. *J. Cogn. Neurosci.*, 4: 179–188.

Nadel, L. and Moscovitch, M. (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.*, 7: 217–227.

Nestor, P.J., Fryer, T.D., Smielewski, P. and Hodges, J.R. (2002) Dysfunction of a common neural network in Alzheimer's disease and mild cognitive impairment: a partial volume corrected region of interest 18FDG-positron emission tomography study. *Neurobiol. Aging*, 23: S358.

Norman, D.A. and Shallice, T. (1986) Attention to action: willed and automatic control of behavior. In R.J. Davidson and D. Shapiro (ed.), *Consciousness and Self-regulation.* New York: Plenum, pp. 1–18.

O'Keefe, J. (1976) Place units in the hippocampus of the freely moving rat. *Exp. Neurol.*, 51: 78–109.

O'Keefe, J. and Nadel, L. (1978) *The Hippocampus as a Cognitive Map.* Oxford: The Clarendon Press.

O'Reilly, R.C. (1998) Six principles for biologically-based computational models of cortical cognition. *Trends Cogn. Sci.*, 2: 455–462.

Patterson, K. and Hodges, J.R. (2000) Semantic dementia: one window on the structure and organization of semantic memory. In L. Cermak (ed.), *Revised Handbook of Neuropsychology: Memory and Its Disorders.* Amsterdam: Elsevier Science, pp. 314–335.

Raffone, A. and Walters, G. (2001) A cortical mechanism for binding in visual working memory. *J. Cogn. Neurosci.*, 13: 766–785.

Rescorla, R.A. and Wagner, A.R. (1972) A theory of pavlovian conditioning: the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy (eds), *Classical Conditioning II: Current Research and Theory*. New York: Appleton-Century-Crofts.

Riedel, G., Micheau, J., Lam, A.G.M., Roloff, E.L., Martin, S.J., Bridge, H., de Hoz, L., Poeschel, B., McCulloch, J. and Morris, R.G.M. (1999) Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nature Neurosci.*, 2: 898–905.

Ripps, L.J., Shoben, E.J. and Smith, E.E. (1973) Semantic distance and the verification of semantic relations. *J. Verbal Learning Verbal Behav.*, 12: 1–20.

Rogers, T.T., Lambon Ralph, M.A., Garrard, P., *et al.* (2004) Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol Rev,* 111: 205–235.

Rogers, T.T. and Plaut, D. (2002) Connectionist perspectives on category-specific deficits. In E.M.E. Forde and G.W. Humphreys (eds), *Category Specificity in Brain and Mind*. Hove: Psychology Press, pp. 251–284.

Rose, S.P.R. (1992) *The Making of Memory*. London: Bantam Press.

Rose, S.P.R. (2002) 'Smart drugs': do they work? Are they ethical? Will they be legal? *Nature Rev. Neurosci.*, 3: 1–6.

Rumelhart, D.E. and McClelland, J.L. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: Bradford Books.

Sarnthein, J., Petsche, H., Rappelsberger, P., Shaw, G.L. and von Stein, A. (1998) Synchronization between prefrontal and posterior association cortex during human working memory. *Proc. Natl Acad. Sci. USA*, 95: 7092–7096.

Schacter, D.L. (2001) *The Seven Sins of Memory.* New York: Houghton Mifflin.

Schacter, D.L., Kaszniak, A.W., Kihlstrom, J.F. and Valdiserri, M. (1991) The relation between source memory and aging. *Psychol. Aging*, 6: 559–568.

Schneider, J., Kavanagh, S., Knapp, M. Beecham, J. and Netten, A. (1993) Elderly people with advanced cognitive impairment in England: resource use and cost. *Ageing Society*, 13: 27–50.

Shallice, T. (1988) *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Sherry, D.F. and Schacter, D.L. (1987) The evolution of multiple memory systems. *Psychol. Rev.*, 94: 439–454.

Simons, J.S., Verfaellie, M., Galton, C.J., Miller, B.L., Hodges, J.R. and Graham, K.S. (2002) Recollection-based memory in frontotemporal dementia: Implications for theories of long-term memory. *Brain*, 125: 2524–2536.

Singer, W.(1993) Synchronization of cortical activity and its putative role in information processing and learning. *Annu. Rev. Physiol.*, 55: 349–374.

Singer, W. (1999) Binding by neural synchrony. In R.A. Wilson (ed.), *The MIT Encyclopedia of the Cognitive* Sciences. Cambridge: MA: MIT Press, pp. 81–84.

Smyth, M.M., Pearson, N.A. and Pendleton, L.R. (1988) Movement and working memory: patterns and positions in space. *Q. J. Exp. Psychol.*, 40: 497–514.

Squire, L.R. (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.*, 99: 195–231.

Stuss, D.T., Craik, F.I., Sayer, L., Franchi, D. and Alexander, M.P. (1996) Comparison of older people and patients with frontal lesions: evidence from world list learning. *Psychol. Aging*, 11: 387–395.

Sur, M. and Leamey, C. (2001) Development and plasticity of cortical areas and networks. *Nature Rev. Neurosci.*, 2: 251–262.

Swainson, R., Hodges, J.R., Galton, C.J., Semple, J., Michael, A., Dunn, B.D., Iddon, J.L., Robbins, T.W. and Sahakian, B.J. (2001) Early detection and differential diagnosis of Alzheimer's disease and depression with neuropsychological tasks. *Dement. Geriatr. Cogn. Disord.*, 12: 265–280.

Tang, Y.P., Shimizu, E., Dube, G.R. *et al.* (1999) Genetic enhancement of learning and memory in mice. *Nature*, 401: 63–69.

Treisman, A. (1996) The binding problem. *Curr. Opin. Neurobiol.*, 6: 171–178.

Tsien, J.Z., Chen, D.F., Gerber, D., Tom, C., Mercer, E.H., Anderson, D.J., Mayford, M., Kandel E.R. and Tonegawa, S. (1996) Subregion and cell type-restricted gene knockout in mouse brain. *Cell*, 87: 1317–1326.

Tulving, E. (1972) Episodic and semantic memory. In E. Tulving and W. Donaldson (eds), *Organisation of Memory.* New York: Academic Press, pp. 381–403.

Tulving, E. (1983) *Elements of Episodic Memory.* New York: Oxford University Press.

Tulving, E. (2001) Episodic memory and common sense: how far apart? *Phil. Trans. R. Soc. Lond. B Biol. Sci.*, 356: 1505–1515.

Ungerleider, L.G. and Mishkin, M. (1982) Two cortical visual systems. In D.J. Ingle, M.A. Goodale, R.J.W. Mansfield (eds). Cambridge, MA: MIT Press, pp. 549–586.

Vallar, G. and Shallice, T. (1990) *Neuropsychological Impairments of Short-term Memory.* Cambridge: Cambridge University Press.

Vargha-Khadem, F., Gadian, D.G., Watkins, K.E., Connelly, A., Van Paesschen, W. and Mishkin, M. (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277: 376–380.

Wan, H., Aggleton, J.P. and Brown, M.W. (1999) Different contributions of the hippocampus and perirhinal cortex to recognition memory. *J. Neuroscience*, 19: 1142–1148.

Warrington, E.K. (1975) Selective impairment of semantic memory. *Q. J. Exp. Psychol.*, 27: 635–657.

Warrington, E.K. (1982) Amnesia: A disconnection syndrome? *Neuropsychologia*, 20: 233–248.

Warrington, E.K. and Weizkrantz, L. (1968) New method of testing long-term retention with special reference to amnesic patients. *Nature*, 217: 972–974.

Warrington, E.K. and Shallice, T. (1984) Category specific semantic impairments. *Brain*, 107: 829–854.

Willshaw, D. and Dayan, P. (1990) Optimal plasticity from matrix memories: what goes up must come down. *Neural Communication*: 85–93.

Wilson, B.A. (1999) *Case Studies in Neuropsychological Rehabilitation.* New York: Oxford University Press.

Wilson, B.A. (2003) Cognitive rehabilitation of memory disorders: Theory and practice. In P.P. DeDeyn, E. Thiery, R. D'Hooge (eds), *Memory: Basic Concepts, Disorders and Treatment*. Leuven: Acco Press, pp. 399–412.

Wilson, M.A. and McNaughton, B.L. (1993) Dynamics of the hippocampal ensemble code for space. *Science*, 261: 1055–1058.

Wilson, M.A. and McNaughton, B.L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science*, 265: 676–682.

Yates, F.A. (1966) *The Art of Memory*. London: Routledge and Kegan Paul.

CHAPTER

10

Memory, Reasoning and Learning

Kieron O'Hara, Wendy Hall, Keith van Rijsbergen and Nigel Shadbolt

1 INTRODUCTION

In this chapter we project probable research directions in computer science, with respect to memory, reasoning and learning. We discuss potential synergies with cognitive neuroscience and related disciplines over 5- and 20-year horizons, linked with wider trends in computing. We summarize our suggestions in an Appendix.

1.1 State of the Art

Interesting developments in computer memory include the use of content addressable memory. In this approach, instead of

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

retrieving information via index structures, software exploits the properties of memories to allow a rudimentary form of associative memory.

New methods of organization and retrieval are coming on line as a result of the greater multimedia content of data stores. Content is beginning to be something that systems themselves can determine, as opposed to being imposed from outside. For example, autonomous agents allow interactions with the environment to develop concepts and categories.

Traditionally, logic has underlain automated reasoning. However, we are beginning to understand the relation between logic and more biologically plausible neural net reasoning, as well as formalisms such as non-monotonic logic for modelling 'scruffy common sense' inference.

New methods are arising for reasoning under uncertainty, ranging from neural nets, probabilistic reasoning and methods of harvesting content from web-scale information stores. Furthermore, many new domains for reasoning are coming on stream, based around technologies such as multimedia, the Semantic Web and the eScience grid, and requiring special purpose methods.

To be used efficiently, content should be personalized to the user. Research is developing using modelling technologies and techniques to perform this task. Aspects of neuroscientific models of learning are under scrutiny, and will continue to set a research agenda. Currently, though, most automatic learning is done via statistical methods, by machine learning. Research is leading to interesting developments in information extraction from natural language.

1.2 Open Questions

We need to bear in mind some key distinctions to understand fully the contributions that computing and neuroscience can make to each other. First, there is already successful research developing artificial neural systems. However, we must not lose sight of the fact that such systems exploit only some properties of the neural systems of animals. Therefore we must expect some disanalogies between the two types of system.

Secondly, there are two ways of approaching cognitive modelling. Top-down modelling involves specification of some cognitive properties, and then design of systems to instantiate them.

Bottom-up modelling allows 'cognitive' systems to develop according to a causal route that mimics some natural process such as evolution. Both approaches have had successes, and are in important ways complementary. Thirdly, another distinction, that between the physical and the digital world, is in many ways converging.

There is scope for widening access to memories by enhancing content addressable memory, by determining useful synergies with biologically plausible hardware and software. Multimedia and multimodal memories need methods of representation, organization and retrieval.

We can improve efficiency in storage and retrieval by exploiting neuroscientific mechanisms, such as different types of memory (long-term versus short-term, episodic versus semantic), forgetting or deleting knowledge that is 'past its sell-by date', and developing concepts and categories internally by creating links between patterns in sensory input and particular symbols.

Underpinning the development of reasoning is the use of distributed computing to create large-scale computing power. Some of this power may be special-purpose reasoning directly implemented in hardware.

Research into modelling the highways and byways of logic, such as non-monotonic logic or reasoning under uncertainty, will continue. Multimedia demand non-verbal reasoning, or methods of annotating non-verbal content. The usability of standard systems will be tested by heterogeneous users on the Web, demanding developments in Web navigation, while information-hungry science will place heavy demands on the grid architecture.

A key factor in enhancing the usability of computing technology will be the development of improved user models, perhaps as a 'side effect' of knowledge acquisition from user behaviour. The neurosciences have much to teach us about reinforcing learning, and how to develop plastic architectures that can modify learning abilities at crucial stages. Learning and plasticity will be crucial to the development of autonomic computing in which systems contain a degree of self repair.

Machine learning has always been important in our learning methods. We must continue to develop smart ways of representing information to facilitate reasoning over it. It is perhaps equally important to supplement statistical learning methods with generalizations about the domain to promote helpful biases in learning.

2 SCOPE AND ASSUMPTIONS

In this chapter we review current research from computer science and related fields into memory, reasoning and learning. The aim is to uncover research themes in the area of cognitive science/neuroscience that are important for the development of computer science. Such themes may be problems that have to be solved, or potential opportunities for progress. In this section, we introduce the scope of our research, the assumptions that underlie it, and some of the important distinctions to bear in mind while considering these issues.

2.1 Scope

We consider a 5- to 20-year horizon. We will focus on computer science and its relation to cognitive neuroscience. We address two issues in particular:

- What developments can we expect from computer science? Which of those could benefit from input from neuroscience?
- Given that, what questions should we ask of neuroscience in the longer term?

The issue looks toward the 5-year horizon. The second addresses the longer term. We shall indicate our guesstimates of the work that is plausible on these horizons. We summarize these suggestions in the Appendix.

2.1.1 Type of Research

Most of our discussion concerns computational modelling and software. However, neuroscience and neural computing will introduce hardware issues. We will also discuss these.

2.1.2 Location

Our discussion is concerned with computer science and the potential for development in the UK. This will, however, be understood in the context of the worldwide discipline, and the market forces that will privilege particular lines of development over others.

2.1.3 Background Assumptions

Our chief assumption about the future of computing is the continued development of the Internet and World Wide Web (WWW). This implies:

- An increase in application scale. Retrieval methods in particular will have to cope with Web-scale domains.
- Greater uncertainty about representation formats. As the WWW increases in size, and as the number of users increases, there will be increasing quantities of legacy data – 'old' data often created using out-of-date software or languages, and about which we can make few simplifying assumptions. Much of the legacy data will be in formats developed for other uses. Increasing quantities will be in relatively unstructured natural language. There will also be more data that is not in textual form, but in multimedia formats.
- Greater heterogeneity of users. More users, who in general will be less expert, will require information and communication technology.
- Greater ubiquity. Current trends in computing point towards providing the

individual user with less computing power and a move towards distributed networks of cheap, local processing. Such networks will enable the embedding of computation in many more aspects of life and work.

2.2 Memory, Reasoning and Learning

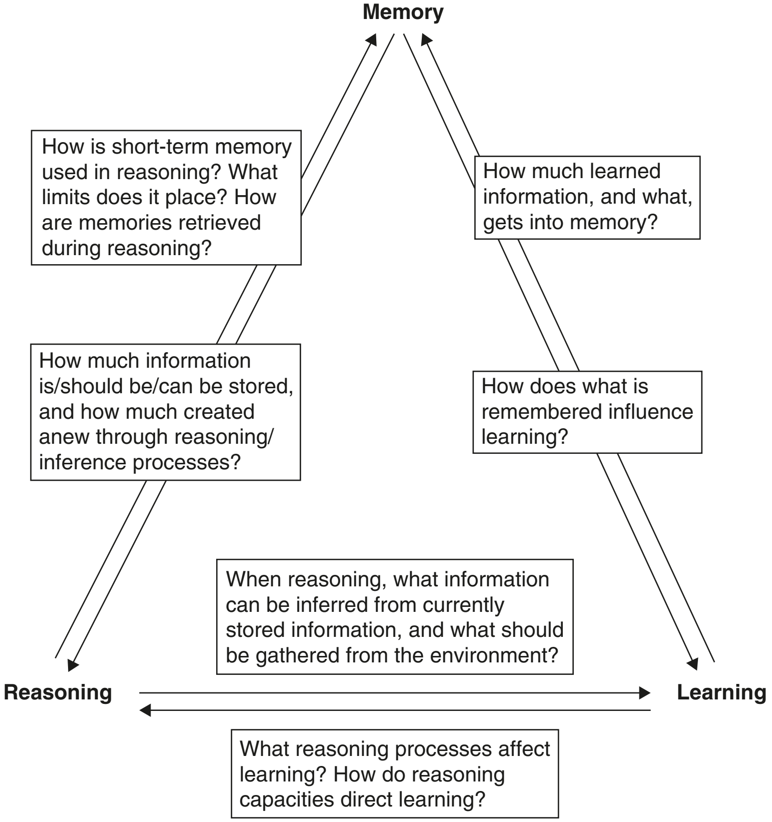
In this chapter, we focus on memory, reasoning and learning. These interrelate in interesting ways. A position on one issue will affect the possibilities on others. We discuss some of these interrelationships, but otherwise we will not mention them explicitly. (We can see some of the relations between the memory, reasoning and learning in Fig. 10.1). In many cases, it is possible to recast a problem in one category, 'memory' say, as a problem in another, e.g. 'learning'.

Orthogonal to these interrelationships are a number of key distinctions. We will discuss these briefly to end our introductory task:

- Human and artificial neural computing
- Top-down and bottom-up approaches
- Physical and digital.

2.3 Human and Artificial Neural Computing

We must beware of expecting too strong a connection between neuroscience and computing. While the human brain is the remembering, reasoning and learning machine *par excellence*, and it is natural to look to its



How is short-term memory used in reasoning? What limits does it place? How are memories retrieved during reasoning?

How much information is/should be/can be stored, and how much created anew through reasoning/inference processes?

How much learned information, and what, gets into memory?

How does what is remembered influence learning?

When reasoning, what information can be inferred from currently stored information, and what should be gathered from the environment?

What reasoning processes affect learning? How do reasoning capacities direct learning?

FIGURE 10.1 Some relations between memory, reasoning and learning.

study for 'hints' on how to make progress, we should bear in mind important differences between brains and computers.

2.3.1 Input from the Neurosciences

Neural net research makes several simplifying assumptions – for obvious and justifiable reasons – about neural computing that may not apply to the human brain. For instance:

- Human neurons receive a large range of inputs, and have many different cellular pathways for them to effect changes. An artificial neuron typically bases its behaviour on a local learning law.
- Human neural assemblies encode information in a variety of ways apart from the strength of synaptic connections. These include the timing and frequency of neural impulses, varieties of oscillatory circuit and reverbatory loops. Artificial systems are only beginning to exploit a variety of encoding.
- The human brain contains many re-entrant connections with a high degree of feedback throughout all areas of the central nervous systems. This feedback is beyond the scope of all but a few artificial neural architectures.
- Human learning is not necessarily connected intimately to learning examples, but often takes place at a distance in time (for example, we may lay down some memories during sleep). Artificial nets are memory efficient, in that they do not require storing a large amount of training information.
- Large-scale modulatory phenomena, such as emotional states, affect human learning. Artificial nets are not subject to these complex contexts.
- The scale difference between the human brain and an artificial net is very large.
- Human neural learning must be 'self-sufficient', while artificial neural nets (or robotic agents) require further intervention, for example through experimentation with network designs, searches for local maxima, adjustments of learning rates, etc. They are often put into artificial situations that maximize their learning potential.

2.3.2 Lessons from the Comparison

None of these reasons invalidates neural research, but equally is no easy route from neuroscience to computer science. We have to make simplifying assumptions. The lessons to take forward are:

- Neural net technologies constitute important paradigms for research in computer science.
- We should beware of expecting a simple export of results from neuroscience to computing. The passage of an idea from neuroscience is more often as a computing metaphor.
- Biological plausibility is always of interest in artificial systems. Making use of such constraints often aids the efficiency and simplicity of such systems.

There are many ways of conceiving the relationship between an artificial neural (or robotic) system and a human cognitive system. There are at least seven respectable dimensions for evaluating an artificial model (Webb, 2001).

- *Relevance* Does the model test and generate hypotheses applicable to biology?
- *Level* Where are the elements of the model, on an ontological hierarchy from atoms to societies?
- *Generality* What range of biological systems does the model represent?
- *Abstraction* How much of the modelled system's complexity appears in the model?
- *Structural accuracy* How well does the model represent actual behaviour-producing mechanisms?
- *Performance* How well does the model's performance match that of the system?
- *Medium* What is the model's physical basis?

When discussing an artificial model, we need to be clear what the aims of the model are in terms of dimensions such as these.

2.3.3 *Metaphors*

Metaphor is an important part of scientific research and of conceptualization. As an example, consider memory research. Different paradigms of memory research are emerging that exploit two distinct metaphors.

The traditional paradigm, laboratory-based, sees memory in effect as a storehouse, focusing on what is 'kept' in there, and 'how it is retrieved'. This paradigm has been challenged recently by a more context-embedded view, seeing memory as a correspondence between past events and their recall, emphasizing the everyday, naturalistic aspects of memory and recall (Koriat and Goldsmith, 1996).

2.4 Top-down and Bottom-up Approaches

A second important distinction is that between top-down and bottom-up approaches to the study and modelling of cognition. Top-down approaches specify behaviour in advance, and try to find architectures that produce it. Bottom-up approaches specify architectures and try to see what behaviour results.

2.4.1 *Logic and Virtual Machines*

One example of a top-down approach is the view that logic is adequate to describe psychological phenomena. This view is attractive: it provides epistemological foundations (the axioms of the logic). However, as we shall see (section 4.1), automatic reasoning based on logic is not enough for a general psychological account. We therefore need different logics for different contexts.

Nevertheless, the idea continues to influence. Although well-defined formalisms are not thought to be good at describing neural computation, there is a view that at some level of abstraction logic is the best description of inference systems. There is therefore the view that it should be possible to describe any such system as a logical calculator.

Different levels of description of computational systems are termed 'virtual machines'. A key question about neuroscience and artificial neural nets, therefore, is whether such substrates can provide material for a virtual machine that supports logic-based reasoning and subject–predicate descriptions of the environment.

2.4.2 *Holistic, Bottom-up Approaches*

Bottom-up approaches to the study of cognition try to emulate the causal factors of cognitive functionality by simulating the evolutionary development that brought it about. Such approaches reject logic, looking instead for explanatory interactions between agents and the world (Pfeiffer, 1999).

Recent work includes the evolution of robotic controllers to produce team-oriented behaviour. This behaviour can evolve despite the constraints on autonomous agents, such as the possibility of the non-optimality of information that may be available only locally. Systems like this can throw interesting light on cognition both in terms of the properties of the individual agents, and in terms of the group behaviour that emerges from cooperative teamwork. Other examples of evolutionary research include the development of visual capabilities to navigate through noisy environments, and the development of capacities to detect unusual phenomena.

Another bottom-up approach is to try to model simple biological systems. Invertebrate systems can show interesting behaviour. There are approaches to the study of cognition modelling single cells, or simple nervous systems. Examples of the results from such research include systems that can navigate by locating sources of sound. The aim of such systems is to produce rich behaviours without complex information processing.

A related area of hardware research is that of evolutionary electronics. Here artificial evolutionary processes (e.g. genetic algorithms) can be used to design systems on reconfigurable silicon chips. Analysis techniques are being developed, and are still required, to give insights into the operation of circuits produced through evolution, and to map out their fault tolerance.

Bottom-up approaches generally involve developing/modelling complete cognitive systems, as opposed to the top-down style of isolating a particular cognitive function and modelling it in isolation. The advantage of the bottom-up approach is that there are no problems of integration/inconsistency with other cognitive functions. The corresponding disadvantage is that the behaviour can miss system requirements, in terms of both complexity and correctness.

The increasing scope of ubiquitous computing, with distributed networks of cheap processing power, means that it could be important to understand the possibilities of simple behaviour. The ability to produce useful behaviour without placing great constraints on memory could be very timely.

2.5 The Physical and the Digital

A third distinction between natural and artificial systems is that between the physical world and the digital world. In some ways it can be glossed over – the digital world can be seen as virtual reality, where the same (or interestingly different) laws can govern life and reproduction, learning and adaptation (see The EQUATOR Project http://www.equator.ac.uk).

In the physical world, we can investigate intelligence, for example, via physically embodied robots. As we wish to study increasingly complex aspects of the world, then the appropriate robotics focuses on small-scale, smart niched robots, that function in zones with a restricted set of properties.

In the digital world, the analogues of robots are 'softbots', adaptive intelligent agents that are goal-oriented, autonomous and communicative. Such agents will have to possess increasingly focused microintelligence as the digital world specification becomes increasingly complex. The response to complexity by specialization in both the physical and the digital world is not a coincidence, as physical and virtual reality share many essential properties.

Computer science often exploits the physical/digital distinction by borrowing metaphors from the physical side to exploit on the digital. One example of this is the idea of autonomic computing, with distributed networks that can repair themselves, or recover from damage. These often use redundancy to allow recovery without a crash. Such systems need to be adaptive, to have goals for system improvement, and to anticipate demands upon them (Gibbs, 2002).

The use of physical metaphors in the digital world has changed the concepts of intelligence that we use in the digital world. For example, in the natural world the phenomenon of speciation means that many simple specialized forms fill small ecological niches. In the digital world, the same is increasingly happening, with the resulting 'speciation' into microintelligences.

3 MEMORY

3.1 Content Addressable Memory and Associative Memory

In a neural net, the pattern of the unit's activation stores the data. If a unit receives an input from a connected unit, it passes output to other connected units. If these units represent attributes of data, then any connected unit can access the data – i.e. the data can be recalled not only by using the indexing system, but by any of the attributes. As the connections represent the unit's content, this sort of mechanism is called content addressable memory. Such memory allows flexibility of recall and retrieval, and can work round errors or corruption of data to reconstruct the original information.

There are also information retrieval methods for going beyond standard indexing in symbolic systems, and conceptions of the symbolic/subsymbolic relationship that postulate intermediate layers in the architecture (see Gardenfors, 2000, discussed in section 3.5).

Content addressable memory can operate as associative memory. This takes particular patterns of data and returns stored data conforming to those patterns. In other words, the user expresses what he or she is looking for, and the memory is retrieved. This is much more like human memory, where an input triggers semantically linked associations (as with Proust's Madeleine), and directly relates (input) data with (retrieved) data.

3.1.1 Five-year Horizon

Much data comes to us in symbolic form. This leads to the question of how to organize and encode symbolic data so that it is amenable to associative recall. This is clearly related to the idea of intermediate architectures between the symbolic and the subsymbolic (Gardenfors, 2000), and the question of the relationship between such architectures (see section 4.1.1).

One example of a hardware implementation of content addressable memory is that of correlation matrix memories (Hodge and Austin, 2001). Their modular structure in particular suggests research to understand retrieval processes (and information-encoding) across distributed, composed memories. Is such a modular structure biologically plausible, or does the neurological system have to be far more interdependent?

3.1.2 Twenty-year Horizon

Can these memory models be usefully informed about the organizational structures found in cortical architectures?

3.2 Multimedia Issues

Multimedia data mainly pose problems of information retrieval. How do you search multimedia data without the relative ease of textual cues? But there is also a related problem of memory and storage. This is an example of how the three themes of this chapter interrelate very tightly.

3.2.1 Five-year Horizon

Multimedia data storage requires an integrated set of storage technologies that facilitate information retrieval. The storage requirements for graphics and video are of course greater than for text. Hierarchies are a potential mechanism here. Furthermore, the real-time support requirements of multimedia need special-purpose algorithms and models.

Data retrieval needs to be sensitive to delays and heterogeneous retrieval times. Index processing of multimedia data is expensive. Different types of data need to be synchronized. Support for data retrieval would ideally be built into multimedia memory structures.

An interesting idea, borrowed from computer graphics and games, is that of multiresolution modelling. Realistic graphics (for example in the representation of bumpy/scruffy surfaces) require complex, detailed models. Rendering such graphics often has to be fast (sometimes interactive), and the complexity naturally slows processing down, so software designers have developed techniques to extract the essential details from a scene and to discard the unnecessary details.

Multiresolution models operate on objects at many different levels of abstraction. They can reconstruct any one of these levels on demand. There are a number of techniques for this. For example 'wavelets' are mathematical functions that can represent data at different scales and resolution. Hence wavelet algorithms process data at different levels, and the scale in which one is interested crucially affects the way data are processed.

Another approach is with simplification envelopes. These are techniques for establishing a global measure of the error of the model in representing its object. In effect, this approach works out a set of limits within which the object is known to exist.

3.2.2 Twenty-year Horizon

Multiresolution modelling may prove to be an interesting interpretation of neural processing. Computer systems now have a full range of sensory input capabilities. The notion of a media type will extend well beyond text, image and audio into meaningful representations of haptic interactions (touch) and olfactory experience (smell).

Understanding how to represent, index and integrate memories that originate in different modalities, will become a pressing question for research.

3.3 Different Types of Memory

3.3.1 *Five-year Horizon*

There will be a pressing requirement for the construction and indexing of large-scale personal and corporate multimedia memories. This in turn will focus attention on the need for semantic enrichment of the content so that it can be integrated and organized in a more useful fashion. This is likely to direct attention to the mechanisms employed to these ends in human memory, which are linked to questions about the relationship between working memory and long-term memory.

3.3.2 *Twenty-year Horizon*

There are many challenges for understanding the relationship between different types of human memory. For example, what are the constraints on the extraction of information from working memory? Understanding this sort of filtering issue, in psychological and neuropsychological terms, would help to cast light on the focusing of attention. How does irrelevant information get removed from working memory? How does our understanding of 'irrelevant' here fluctuate with context? How are objects selected for attention focus in general, and to structure the cognitive functions to minimize the costly switching of focus of attention (Oberauer, 2002)?

A related issue is that of conceptualization of the environment. Survival often means understanding the environment in terms of both the arrangement of objects in the field of vision (e.g. that some rocks and trees appear), and their significance to the subject (e.g. as the path home) (Glenberg, 1997).

Another example of a problem developing from discussion of human memory is that of selecting particular storage mechanisms, or recollection mechanisms, for particular memories or particular tasks.

A final point: much of our knowledge is implicit. I do not explicitly recall that Tony Blair was wearing trousers when I last saw him addressing the House of Commons on TV, but if asked I can deduce that he was, from, among other things, the fact that I would remember if he was not. Memory is a trade-off between laying down memories, thereby creating a retrieval problem, and leaving certain things implicit and performing some reasoning to rediscover them, thereby creating reasoning problems. An interesting synergy with neuroscience would be to find ways to characterize this distinction accurately, to select those pieces of knowledge best 'remembered' implicitly, and to link the distinction with similar distinctions, such as procedural/declarative, conscious/unconscious and verbalizable/non-verbalizable.

3.4 Forgetting

An intriguing topic, and one where computer science has much to learn from neuroscience and psychology, is that of forgetting obsolete data. Forgetting, in human psychology, is more than just a failure of memory; it is an essential mechanism for clearing out useless facts in order to preserve efficiency of recall. The forgotten facts can no longer interfere with search. For computing, the availability of ever-cheaper storage capacity means that 'forgetting' is no longer essential to make room for new knowledge. However, forgetting can still be useful for two reasons:

- to improve search, which is analogous to focusing of attention
- as part of routine knowledge maintenance, information repositories could be stripped of knowledge when its validity conditions failed.

3.4.1 *Five-year Horizon*

Research problems to do with forgetting include:

- Identification of knowledge to forget.
- Deciding how to deal with knowledge related to that which is forgotten; such as knowledge that implies the forgotten knowledge, knowledge whose only justification is the forgotten knowledge, or processes that make essential use of the forgotten knowledge.
- Interpreting the meaning of the term 'forget': should it mean removal from memory, or simply making knowledge harder to find? This issue is related to that of the conception of information as being 'foreground' or 'background'. Psychologically, information in the foreground is more amenable to processing; background information is only processed where necessary.

On this interpretation, the information space could adapt to a user's decisions as they follow a path through it, and hence the information that passes from the background to the foreground is determined implicitly to some extent by the choices made by the user in the past (Campbell and van Rijsbergen, 1996). This is illustrated in Fig. 10.2, in which the user path through an information space alters the space itself.

3.4.2 *Twenty-year Horizon*

A number of psychological mechanisms could be applied, metaphorically at least, to knowledge stored mechanically.

- *Gestalt theory* Memories are adjusted in the light of new knowledge to create a more streamlined, theoretically smooth view of the world, forgetting 'inconvenient' information that contradicts new corroborated theories. This is analogous to abstraction, or processes designed to 'smooth' data in machine learning.
- *Decay theory* Knowledge is forgotten when it has been unused for a period of time, and the brain structures encoding it decay. Alternatively, forgetting takes place when representations of the responses that lead to their recall are truncated (Killeen, 1994). Such a process could be artificially created in computers. An example of this appears in the QuiC dynamic link service, which generates links for users on the basis of the browsing patterns of similar users. Links that are not used are removed (El-Beltagy *et al.*, 2002). (See section 4.5 for more on link separation and external link storage.)
- *Interference theory* Knowledge of different types can interfere with each other and some knowledge can be overwritten. This might be seen as 'consistency checking', where new information caused the removal of inconsistent information already in memory.
- *Forgetful neural networks* On one model of memory, we learn new patterns of activation gradually at the expense of older ones (Hopfield, 1982). This model is important for neural networks, where there is the danger of 'catastrophic forgetting'. This could happen when essential changes wipe out distributed representations in models, when, for example, the model is scaled up. Smolensky's superpositional memory theory might be a useful medium for this (Smolensky, 1991).
- *Retrieval failure* Memory is more reliable when the cues accompanying retrieval are similar to the cues accompanying learning. Associating memories with acquisition context might be a way of making memory more context-sensitive.

3.5 Symbol Grounding

The symbol grounding problem is a problem about how symbols get their reference. For a symbol system to be useful, it has to refer to the world in some respect. For that to happen there needs to be some sort of relationship between a symbol and its referent. However, this relationship cannot be grounded verbally, by definitions, because a definition only provides a link between two symbols.

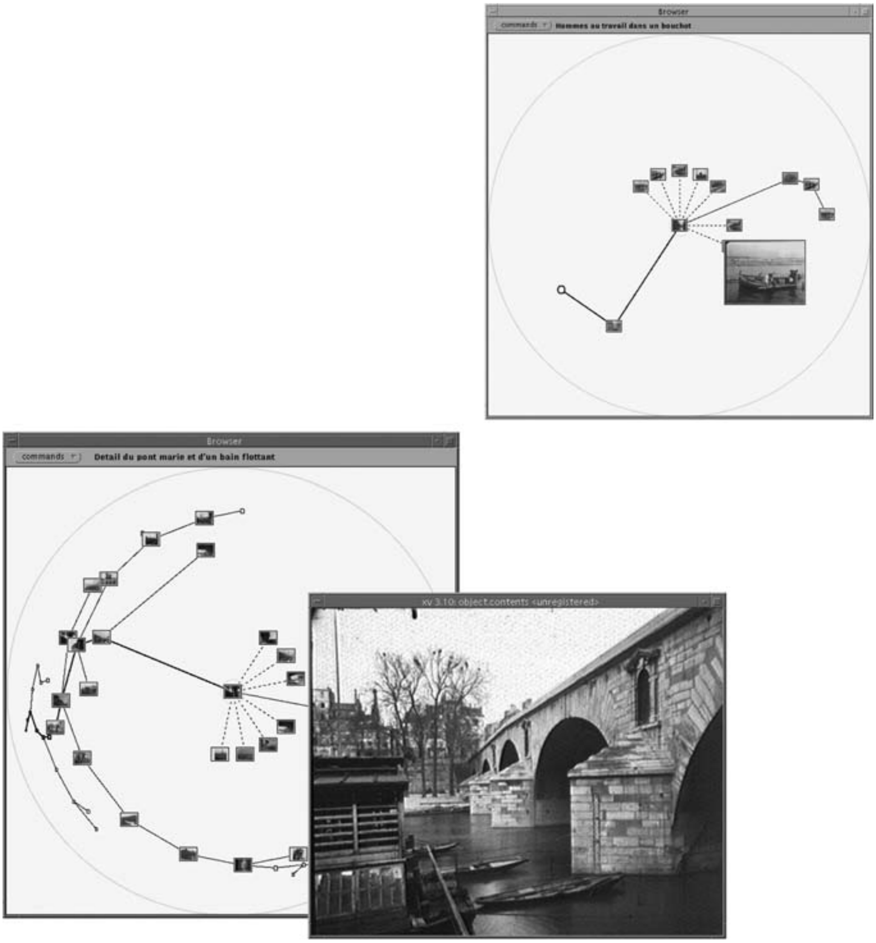

Browser
commands
Hommes au travail dans un bouchot
Browser
commands
Detail du pont marie et d'un bain flottant
xv 3.10: object.contents <unregistered>

FIGURE 10.2 A user entering an information space will experience the choice of a number of objects (images). As the user moves from object to object, tracing out a path, they will face a choice of the most relevant objects. As the user continues, the previously chosen objects are suitably arranged. The arrangement of the objects at any one point is a function of the path through the space. A user can, at any time, start browsing from any one of the objects displayed.

3.5.1 Five-year Horizon

Neural net systems are generally good at classificatory tasks. They have also shown interesting linguistic properties. It can therefore be argued that connectionism is the natural candidate for learning the symbol grounding invariant features, connecting names with the objects they stand for (Harnad, 1990). Another approach is to use autonomous agents to bootstrap ontologies, words, visual categories or other semantic structures, through particularly structured interactions with the environment (Steels and Kaplan, 1999; Steels, 2001).

3.5.2 Twenty-year Horizon

A third approach to symbol grounding is via neural Darwinism, the idea that groups of neurons, clustered together and wired up in a more or less random way, achieve a regular kind of interactive relationship with the outside world. On this view, unfamiliar stimuli 'invade' the brain through the sensory system. Some of these randomly wired clusters of neurons react more strongly than others to the stimuli. If such reactions strengthen connections between the neurons, then those neurons will develop an increasingly strong reaction to stimuli of that type (Edelman, 1990, 1992).

A fourth approach is to understand an intermediate stage between the subsymbolic stratum and the symbolic layer. For example, Gardenfors (2000) argues that symbolic representation is particularly bad for concept learning, and conceptual spaces (geometrical structures based on quality dimensions) are a better framework for representing knowledge at the conceptual level.

4 REASONING

4.1 Logic and Alternatives

4.1.1 The Relation between Logic-based Reasoning and Neural Nets

Automated reasoning systems are among the most mature applications of computer science. Early attempts to produce a general purpose reasoner were widely seen as a dead end. Special purpose reasoners for smaller, tractable problems, such as theorem proving, have superseded them (see e.g. Brewka *et al.*, 1997; Greiner *et al.*, 2001; Chesñevar *et al.*, 2000).

Different logics have developed for different reasoning contexts, such as particular propositional contexts (e.g. reasoning about beliefs, reasoning about time and temporal relations), different levels of expressivity (e.g. propositional logic, first order logic with quantification, fuzzy logic), or particular phenomena (e.g. processes, uncertainty). The theory of such logics is well-understood. They provide an intersection with related disciplines, such as artificial intelligence, philosophical logic, databases, etc.

4.1.2 Five-year Horizon

Reasoning in natural and artificial neural systems is often understood to go on below the level of logic. As a result, the relationship between 'classical' automated reasoning, and biologically plausible reasoning needs elaboration. Can neural nets underlie a virtual machine for logic-based reasoning? There are promising results in this area.

One issue at the interface of computer science and psychology is that of the limitations of a resource-bounded processor with respect to a logic. Logics are 'perfect', mathematical Platonic systems with no real time or memory constraints. Machines that embody logic are resource-bounded, and consequently deviate from their underlying logics in ways that can often be of psychological interest (O'Hara *et al.*, 1995). If we could use a neuronally based virtual machine for logic-based reasoning, it would be interesting to see how such a machine would deviate from the underlying logic, and whether such deviations had interesting or surprising properties.

4.1.3 Non-monotonic Reasoning

It is simplest to reason in environments that are generally 'well-behaved'. This is not always possible. Non-monotonic reasoning tries to model more closely the type of messy, common-sense reasoning that people actually do. It is the logic of contingency rather than necessity (Brewka *et al.*, 1997).

The key assumption of a monotonic, 'well-behaved' logic is that if a set of propositions entails a conclusion, then any addition to that set of propositions still entails the conclusion, that is, if what you know means that something must be true, finding out more will not change it. Non-monotonic reasoning drops that assumption, and allows conclusions to be withdrawn upon the discovery of more information.

Types of non-monotonic reasoning include: belief revision systems, reasoning under uncertainty (section 4.2), handling inconsistency, default reasoning, reasoning from similarities, planning systems, and abductive reasoning (i.e. if a hypothesis would, if true, explain data better than other hypotheses, conclude that the hypothesis is true).

4.1.4 Five-year Horizon

Aspects of non-monotonic reasoning in current research include: the relation of non-monotonic reasoning systems to the actual reasoning that people do; how non-monotonic reasoning can be supported by neural systems; and semantic accounts of non-monotonic reasoning, as well as honing the algorithms (e.g. how to decide on the parameters, and how to assess the parameter values, of abduction algorithms).

4.1.5 *Hardware and Evolution*

There is a move towards micro-intelligences, distributed systems and niched systems. Using logic in computing implies structured understanding of the reasoning to be performed. However, the success of a number of anti-structural approaches – such as neural nets and hidden Markov modelling – in tackling problems such as speech recognition, data-mining and robot navigation, has led to a new appreciation of what brute computational force can achieve, rather than intensive advance analysis of problems.

4.1.6 *Twenty-year Horizon*

In the natural world, much computation is performed by direct implementation in special purpose hardware. Evolution has balanced a number of requirements, playing behavioural requirements off against environments, brains against bodies, to produce such adaptive architectures. As our computing techniques improve, and we can in effect throw computational power at problems, we will need synergistic research with neuroscience and biology to find insights into the best use of this power.

4.2 Uncertainty

We might express the problem of uncertainty as how to reason in situations which are underdescribed. If we do not have values for relevant parameters of a particular inference, how can we make decisions? This is, of course, very common, for example when inferences are required in real time, when there simply may not be time to gather all the desirable information.

4.2.1 *Five-year Horizon*

One approach to reasoning about uncertainty is with neural nets. These are very robust with respect to noisy or incomplete data, they also allow fast search but at the cost of reliability, and can perform combinatorial search.

Uncertainty also features heavily in information retrieval. Information retrieval has a number of orthogonal dimensions, including logics for reasoning about structures of multimedia documents, and associated theories of uncertainty to supplement the formal logic. The uncertainty provides a quantitative measure of the relevance of a stored object to a query. A third dimension is that of providing parallel and probabilistic algorithms to improve retrieval performance under such logical and probabilistic descriptions. This is a well-trodden research path (Crestani *et al.*, 1998).

4.2.2 *Twenty-year Horizon*

On the longer horizon, work from the cognitive neurosciences that might be relevant includes focus of attention. This is now largely understood as a computer vision issue, but this is an instance of the general problem of extracting the important information from input describing the environment. An understanding of the properties of the current task can help to introduce biases into any representation of the environment to increase the chances of discovering the important factors.

4.2.3 *Probabilistic Reasoning*

An alternative, possibly complementary, approach to uncertainty is probabilistic reasoning (de Mántaras, 1990). Whereas standard logic-based reasoning reasons about propositions or other symbols of semantic significance, probabilistic reasoning looks at propositions associated with some degree of belief. The problem then is how to represent and combine degrees of belief in inference.

There is also work in so-called 'graphical models'. These are types of probabilistic networks with their roots in several research communities.

The graphical models framework provides a clean mathematical formalism that suggests that a wide variety of network approaches to computation, including neural networks, are instances of a more general probabilistic methodology (Jordan and Weiss, 2002). Moreover, there are now highly efficient implementations of these models that realize probabilistic inference. Learning algorithms have been built within this approach that some argue are good candidates for describing learning in cognitive systems (Jordan and Jacobs, 2002).

Five-year Horizon

Challenges here include the development of psychologically plausible models of probabilistic reasoning, their implementation on biologically plausible platforms, extending the principles of probabilistic reasoning to distributed and multi-agent environments, and finding techniques and software tools for modelling the environment as probabilities. This last would be useful, for example, in knowledge acquisition from lay people, who are often required to calibrate computer systems using Bayesian or other probabilistic terms.

4.2.4 *Web-scale Applications: Semantic Harvesting*

Five-year Horizon

Using probabilistic reasoning to understand content from the Web could be crucial. Given the scale of the Web, it will be essential to harvest, index and annotate such content, and to apply probabilistic methods for determining its significance. We will need ontologies that we can map concepts onto. We will also need natural language-based information extraction techniques to reach a formal understanding of the text on a web page. The requirement for probabilistic reasoning in this context is caused not only by the scale of the application, but also by the irremediable scruffiness of the data.

The application opportunity here is to develop 'semantic clerks', clerically mundane applications for routine information management requiring complex behaviours and detailed ontological views of the world.

4.3 Demands of Multimedia

The new computing environment includes much data in non-textual form. Much new research has focused on the problem of reasoning over multimedia data.

4.3.1 *Five-year Horizon*

One aspect of the problem of information retrieval concerns the nature and types of multimedia data, in order to allow the development or use of representational and inferential technologies and languages that allow effective retrieval of data in this form. We can then model these technologies and incorporate their descriptions into a logic capable of modelling uncertainty, using an object-oriented paradigm, graph models or other formalisms to understand and express the multimedia concepts. The resulting logic should then be a suitable model for a new generation of multimedia information retrieval systems. A key factor here is the conceptualization of retrieval as reasoning or inference (see also section 3.2, and Crestani *et al.*, 1998).

4.3.2 *Twenty-year Horizon*

If we are to move beyond the traditional media forms of text, graphics audio and video we will need methods of retrieval for these other modalities. The storage and retrieval of sensory memories such as smell are active areas of research in neuroscience. This work could provide ways of thinking about computational analogues.

4.4 Usability

4.4.1 *Five-year Horizon*

The future computing environment will be messy. Users will be heterogeneous, with an increasingly complex set of requirements for the systems with which they interact. Many users will be versed in computing, but as computing becomes more ubiquitous they will generally be less expert. Hence one priority is to develop high-level interfaces to assist users in describing what they do not precisely know, by providing expressivity for describing requirements, displaying retrieved information and reformulating queries.

Furthermore, as data repositories grow, systems will have to deal with increasing quantities of data. Much of this data will be in legacy, perhaps obsolete, formats. Some will be out of date or plain wrong. Such a universe is uncongenial to reasoning systems. The new generation of systems needs to be able to operate in such an environment.

For usability, the key issues include effective and appropriate querying, linking and browsing. Building search engines that are sensitive to query semantics will be an important aid for usability, especially for novice users. Important underlying technologies here include the determination of the context of a query – for example determining reference of terms. Furthermore, it is important to present results to the user, in such a way that he or she can easily determine the relevance of the information to his or her requirements.

It will be important to understand the psychology of search and browsing. For example, do people want to retrieve multimedia data on the basis of properties of the medium (e.g. for pictures, in terms of colour, texture, intensity of light)? Or is content-based searching appropriate? How should multimedia documents be presented to the user? Which technologies support a fast perception of the content of multimedia documents?

4.4.2 Twenty-year Horizon

A substantial challenge is to understand the task context and subtleties of workflow that individuals and groups engage in. This requires innovations in terms of sociological and ethnographic methods, as well as technical progress to understand how to recognize and support these types of context switching.

4.5 Indexing, Navigating and Linking

The Web, with its network of hyper-text links, should provide a series of multi-dimensional navigational paths through which the user can move. In fact there are relatively few such paths. This is because with standard technologies the links are provided by authors, who would have to commit themselves to a large maintenance task to provide a comprehensive set of associative links, for example, weeding out dangling links. The links would also reflect the interests and knowledge of the author, rather than those of the reader (see Cailliau and Ashman, 1999).

4.5.1 Five-year Horizon

The lack of associative linking leads users to search engines. These provide, in effect, navigation by query. But this limits hypertext systems. Associative linking allows navigation by browsing, as opposed to the data-base-like navigation by matching, in the case of search engines, key phrases.

One aim is to recover the possibilities of the Web by reintroducing associative linking in different ways (Hall, 2000). This could be done by, for example, storing links separately in an external database called a linkbase. This approach is well-documented and has clear advantages when applied to open hypermedia (Hall, 1996).

Researchers have discussed various other approaches (Chalmers, 1999). These include information retrieval, workflow, collaborative filtering and paths. These techniques are members of a general family of approaches to information access that use a framework with philosophical and semiological roots to examine them in terms of the phenomena they include and exclude, how information is shared with their communities of use, models of user activity, presentation of results, adaptation of system behaviour and the interrelationships of representational components.

One example is for the user rather than the writer to generate associative links. Monitoring the user's browsing behaviour can suggest which documents users find useful, both via monitoring the browsing trails and noting their expressions of greater interest, such as bookmarking. Following such linking in context can help other users with similar profiles by making that information available to them, which – assuming the linkbase is kept up to date – will largely consist of functioning links (El-Beltagy *et al.*, 2002).

Such an idea is one way of introducing context sensitivity. Determination of context is a serious problem for such applications, and will be a key problem to crack.

As an example of potential approaches, TF-IDF (Term Frequency, Inverse Document Frequency) is an information retrieval technique that calculates the importance of terms in a document (Belew, 2000). If a term is used more frequently in one document than in others, then we can assume that the term is important in determining the context of that document. Determining the context of use, and therefore which users are similar to those whose browsing behaviour provides the links, will be an important development, particularly as context can be specified at a lower level of abstraction (i.e. a more local context specified).

Indexing and context is an issue on the borderline of reasoning and memory. When expressing index terms, how much should be pre-computed, and what should be computed dynamically? How should the structure of index strings be expressed?

One answer to this question is through ontologies and description logics. For example, ontologies of the domain can provide sophisticated conceptual models of the document terms and their interrelationships, and can generate links dynamically on that basis (Carr *et al.*, 2001). A further issue is the desirability of having free vocabularies for indexing, as opposed to keeping the vocabulary restricted.

More research is also required on the best ways to present a semantically suggestive set of index terms for information retrieval. Indexing can be manual or automated. Newer approaches include intelligent or agent-based indexing, and, in the context of a Semantic Web, annotations or metadata about documents can provide more intelligence or context-sensitivity.

A further issue about indexing here is the contrast, or complementarity, of statistical information retrieval, and semantically controlled approaches. Should the approaches be integrated closely, or merely act as two separate components of the information retrieval toolkit? An example here of such a challenge might be the combination of TF-IDF retrieval with a link-based approach to retrieval (e.g. Page *et al.*, 1991).

4.5.2 Twenty-year Horizon

Discoveries about human associative memory will be very important in this context. They will allow the specification of a link lifecycle that is closer to human memory patterns. They will also allow intelligent clustering of links and phrases.

The human brain combines different types of retrieval clues seamlessly. One way to understanding this sort of mechanism could be the exploitation of evidence combining techniques such as the Dempster–Shafer theory of evidence (see de Mántaras 1990: 38–56 for a short introduction).

4.6 eScience and the Grid

Finally, we consider the reasoning requirements from the eScience initiative. This is a new environment of complex assemblies of programming power in open systems. The aim is to develop an open digital infrastructure that can handle rapid changes with minimum administration.

We will need autonomous reasoning systems that can deal with such changes in a straightforward way. The aim of producing an infrastructure by large-scale aggregation of computing resources – allowing scientists to 'plug in' and extract the power they need to tackle data-heavy problems – will set a number of important challenges (Foster and Kesselman, 1998; De Roure *et al.*, 2001).

4.6.1 Five-year Horizon

Managing such autonomous systems, and negotiating between users, their agents, host systems and regulators, will be challenging.

Reasoners must be interoperable across, and in combinations of, domains, and capable of adapting to contexts defined by device capabilities, user activity and the current problem. We will require new models of collaboration and workflow management. Furthermore, researchers will come in with varying levels of expertise, requiring flexible user models.

Grid-based computing is likely to produce data on a very large scale. We need to develop methods for retrieving, reusing, publishing and maintaining content on this scale.

The discovery, assembly and exploitation of the suites of services required for eScience applications will require new ways of representing and manipulating them. Grid services will need to be brokered to users, while users will have to find ways to match services to their needs.

One way forward is to develop problem-solving environments. These would allow users to make resource allocation decisions at a higher level of abstraction than grid management. They allow users to see the issues in terms of the problem semantics, not in the alien terms of the underlying infrastructure.

4.6.2 Twenty-year Horizon

The move to a computing utility model, with ubiquitous and pervasive computing power, will happen over this timeframe. This computing fabric will have to provide significant autonomic capability.

5 LEARNING

5.1 Personalization of Content

The value that users place on knowledge depends greatly on the way in which that information is presented. This leads to the idea of personalization, where systems tailor content to fit a user's preferences. For example, we have seen work that attempts to allow the reader's system to plant hyperlinks in web pages, which can then be individually tailored without much effort by the reader (Hall, 1996; section 4.5 above). As another example, recall the work on information spaces that adapt to user choices (section 3.4 above). Other technologies applied to modelling user's requirements include predicative statistical models, machine learning, plan recognition techniques and generic modelling systems (Kobsa, 2001).

5.1.1 Five-year Horizon

Two different, though interlinked, learning-based issues are related to personalization of content:

- What is the relation between personalization of content and human learning?
- How does the system learn about the user?

The first issue generates interesting research questions. The main problem is how the provision of a series of special-purpose paths through material can aid the user's learning. What adaptations can ease the learning process?

These issues lead to straightforward questions of application. For example, how should we integrate such systems into, say, a classroom environment, acquiring user feedback, and feeding back comments to users? And how can we use such technologies for distributed learning?

The second question is how such systems learn about the user. How are user models created? How can we use available information, for example about browsing patterns, to understand how users learn? Agent architectures have helped to address such questions (see papers in Kobsa, 2001).

5.1.2 Twenty-year Horizon

In general, early user models focused on a set of static attributes (e.g. sex, age). However, the trend has been towards models supplemented by information on how users interact with the system. There is an issue as to which static attributes are important, i.e. which psychological attributes affect learning styles (e.g. study background, personality variables, computing experience). Psychological models of learning will provide input here.

A further question is how a system should move from behavioural description to user model. What aspects of behaviour are relevant for model construction? And given the description of behaviour, how should we feed this into a user model? User browsing behaviour, for example, takes place while learning; how should we describe this learning context?

Another issue is how to adapt user modelling to the new computing environment of localized microintelligences. User models have to be mobile, and adaptable or creatable by the small-scale smart appliances. For instance, a computer in your car will want a limited user model of you as a driver (Kobsa, 2001a). Such niches may not just be technologically determined. They may be influenced by national borders, for example if user modelling came under the purview of data protection regimes of varying strictness.

5.2 Reinforcement Learning

Reinforcement learning is where a system, or agent, tries to maximize some measure of reward while interacting with a dynamic environment. If an action is followed by an increase in the reward, then the system increases the tendency to produce that action.

The idea of reinforcement learning crosses many disciplinary boundaries, and features in, for instance, engineering, artificial intelligence, psychology and neuroscience (Kaelbling *et al.*, 1996). Given the problem's direct roots in human learning, the chief issue is how to adapt human or animal mechanisms for reinforcement learning to the development of artificial autonomous agents which can interact with their environment and maximize utility measures.

Researchers in artificial intelligence, agent technology and neural nets have studied reinforcement learning for some time. Many of the mathematical foundations are in place. An increasing number of mathematical models and algorithms are available (Sutton and Barto, 1998).

Such quantitative models of adaptive optimizing control can be used alongside neuroscientific analyses. There are also links with genetic algorithms, although reinforcement learning focuses on the problem of a single agent's learning rather than evolutionary development over several generations. However, the two fields often discuss the same problems, and interact through domain similarity.

5.2.1 Five-year Horizon

Issues connected with reinforcement learning include (Sutton and Barto, 1998; Kaelbling *et al.*, 1996; Mitchell, 1997):

- *Interactivity* How should we describe the interface of the agent and the environment? How can we understand the contribution of a single agent when it may interact with the environment only with a group of other agents? This is obviously a problem with neural nets, for example, where the system itself rather than the autonomous nodes produces what we understand as the 'behaviour'.
- *Uncertainty* How should an agent behave under uncertainty? Which options are available to describe uncertain conditions? How should we assess risk?
- *Goals and rewards* How can we encode goals in reward structures?
- *Learning in complex environments* Actions often have complex and delayed consequences. How can we know which actions affect which environmental variables? How can we best describe the environment to uncover the specific and non-immediate contributions of the agent?

5.2.2 Twenty-year Horizon

- *Context* Understanding contextual restrictions can reduce effort in appraising the environment. How can we develop this insight? Equally, the efficiency of reinforcement learning increases as the expectations of the environment become more accurate: so there is the same requirement to understand the environment as to understand the agent.
- *Settling on knowledge* There is a tradeoff for agents between exploiting current

knowledge of the environment to increase utility, and exploring the environment to learn more about it.

- *Forgetting* Recall from section 3.4 the decay theory of forgetting, where an unused memory gradually decays. Is there a link between reinforcement theories of learning and decay theories of forgetting?

5.3 Plasticity

The nervous systems of all animals develop and adapt to the bodies and environments they find themselves in. Up to three-quarters of neurons in a structure die during early development. This cell death occurs principally during a phase of neuronal development called target innervation, during which neurons send out axons to make synapses with their intended targets (such as muscles or other neurons). Once contact is established, their targets supply molecular factors, called neurotrophic factors, on which the survival of the innervating neurons depends. If a neuron receives too little factor, it dies. Target innervation and the neuronal death that result are an adaptive mechanism by which a brain can learn about its body.

Recent experimental evidence indicates that, in addition to their role in earlier neuronal death, neurotrophic factors are also implicated in later synaptic growth (the fine-grained connections between neurons) and re-arrangement (McAllister *et al.*, 1999). Such mechanisms support the tuning of a nervous system to its body and to itself as the body and brain undergo development, growth and decay (Purves, 1988, 1994).

Recently, researchers have implemented such biologically inspired models on robots. This shows that considerable advantages can be conferred on artificial systems using methods of plasticity or malleability (Elliott and Shadbolt, 2001, 2002). For example, it can allow a neurocontroller to adapt to variations between robots, or enable a robot to recover if it loses some of its ability to sense its environment.

A variety of cellular changes in the brain can modify the strengths and efficacy of synaptic connections, and as a result can provide important structures for learning. One of the most studied is that of long-term potentiation (Baudry *et al.*, 1991; Baudry and Davis, 1994, 1997). In this process, high frequency stimulation of afferent fibres produces a long-lasting, but decremental increase in synaptic connection efficacy, an increase that is strengthened through repetition (see reinforcement learning).

5.3.1 *Five-year Horizon*

The immediate goal should be to investigate a range of biological plasticity methods in computational contexts. These could be implemented in both software and hardware contexts, and in robotic and non-robotic applications. Plasticity is likely to feature as an important requirement of autonomic computing.

5.3.2 *Twenty-year Horizon*

Questions arise as to whether there are critical periods for learning, developmental phases in which the neural architecture is relatively plastic and can 'learn' so much more quickly, with a corresponding degrading of memory, and whether such techniques could be used in artificial systems (Fazeli and Collingridge, 1998). How long should a neural architecture remain plastic? When should plasticity, with its key role in learning, cease, to allow long-term memory to be laid down? At what level of abstraction should plasticity be visible? Biologically plausible computational neural models are a key tool for investigating the plasticity of learning architectures.

5.4 Machine Learning

Large data stores are proliferating throughout society, and machine learning will remain a vital way of extracting usable information from them.

5.4.1 *Five-year Horizon*

There is a need to integrate machine learning, using statistical techniques, with more informed semantic understanding of contextual knowledge. Machine learning techniques find simple patterns in data, but background knowledge may help them to find the more relevant patterns, and help them to operate on noisier data. For example, helpful biases introduced into large search spaces can improve learning.

Related topics include trying to deal with noisy data in a more robust fashion, and using theories of uncertainty (de Mántaras, 1990) in machine learning, as opposed to standard 'crisp' set theory, in order to try to achieve a greater correspondence with the actual terms and categories that we use in everyday life.

Automatic information extraction from natural language texts is a major area of research. It will be an essential method of harvesting content from large data stores (Pazienza, 1997).

5.4.2 *Twenty-year Horizon*

Machine learning with neural nets is a major area of research. It provides an important connection with the neurosciences (Mitchell, 1997).

A key innovation would be to integrate neural net methods with semantic awareness of domains, as mentioned above, to establish how rules provided by experts can influence neural net learning, and vice versa. The ability to understand an environment in rule-based terms should provide much input into neural net accounts of learning. It is clear that understanding of a domain, and expectations of it, have a heavy influence on human learning (Smolensky, 1991).

5.5 Problem Representation

A final challenge is how to represent problems to make them more amenable for solution. What representations of the world facilitate learning about it?

In particular, where a learning problem involves extremely large search spaces, how can we describe these spaces to facilitate the extraction of useful information without simultaneously creating non-existent regularities? And how can we select the input features that are best redescribed in such circumstances?

5.5.1 *Five-year Horizon*

For example, there is the question of bias learning. How can we choose/specify an agent learner's hypothesis space to make it is large enough to contain the solution to the problem about which it is learning, while simultaneously keeping it compact enough to ensure reliable generalization from reasonably-sized training sets? At present, such biasing is done with input from experts. The (semi-)automation of this task would be helpful.

In general, the inclusion of background knowledge is important in efficient and accurate machine learning. The interesting neuroscientific questions are what the limits are to uninformed learning, how can we approach these, and how can we produce a 'natural' selection of information to inform learning (Mitchell, 1991; Baxter, 2000)?

A second challenge is how to exploit related problems and their properties as sources for helpful biases. For example, in multi-task learning, learning for one task improves if we use information contained in the training signals of other tasks, by learning tasks in parallel using a shared representation (Caruana, 1997).

Further issues include the development of architectures to support variable representations, and dealing with changes of representation that occur over the lifetime of an embedded agent. How should representations be used? Should a representation be used over multiple contexts? How useful is the information that is extracted on the back of a particular representation? And how can we discriminate between solutions depending on different representations and different biasing strategies?

A link with evolutionary biology and computational biology is the important emerging issue of the importance of change of representation in gene expression. Modelling of gene expression is increasingly important in the life sciences and biotechnology as a method of modelling interrelationships among genes, and of exploring the world of gene and protein interactions. Researchers are developing algorithms to learn from and to validate the data, and to change representations to introduce and control the inductive biases discussed above (Hsu, 2003).

References and Further Reading

Baudry, Michel, Davis, Joel L. and Andersen, Per (eds) (1991) *Long Term Potentiation*. Cambridge, MA: MIT Press.

Baudry, Michel and Davis, Joel L. (eds) (1994) *Long Term Potentiation*, vol. 2. Cambridge, MA: MIT Press.

Baudry, Michel and Davis, Joel L. (eds) (1997) *Long Term Potentiation*, vol. 3. Cambridge, MA: MIT Press.

Baxter, Jonathan (2000) A model of inductive bias learning. *J. Artif. Intell. Res.*, 12: 149–198.

Belew, Richard K. (2000) *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW.* Cambridge: Cambridge University Press.

Brewka, Gerhard, Dix, Jürgen and Konolige, Kirt (1997) *Nonmonotonic Reasoning*. Chicago: University of Chicago Press.

Cailliau, Robert and Ashman, Helen (1999) Hypertext in the web: a history. *ACM Comput. Surv.*, 31(4): 35.

Campbell, Iain and van Rijsbergen, Keith (1996) The ostensive model of developing information needs. In P. Ingwersen and N.O. Pors (eds), *Information Science: Integration and Perspective: Proceedings of CoLIS-2*, pp. 251–268.

Carr, Leslie, Hall, Wendy, Bechhofer, Sean and Goble, Carole (2001) Conceptual linking: ontology-based open hypermedia. In *Proceedings of 10th International World Wide Web Conference (WWW10)*. ACM Press, pp. 334–342.

Caruana, Rich (1997) Multitask learning. *Machine Learning*, 28: 41–75.

Chalmers, M. (1999) Comparing information access approaches. *J. Am. Soc. Inform. Sci.*, 50: 1108–1118.

Chesñevar, Carlos Iván, Maguitman, Ana Gabriela and Loui, Ronald Prescott (2000) Logical models of argument. *ACM Comput. Surv.*, 32: 337–383.

Crestani, F., Lalmas, M. and van Rijsbergen, C.J. (eds) (1998) *Information Retrieval: Uncertainty Logics: Advanced Models for Representation and Retrieval of Information*. New York: Kluwer.

de Mántaras, Ramon López (1990) *Approximate Reasoning Models*. Chichester: Ellis Horwood.

de Roure, David, Jennings, Nicholas and Shadbolt, Nigel (2001) *Research Agenda for the Semantic Grid: A Future E-Science Infrastructure*. Edinburgh: National e-Science Centre.

Edelman, Gerald M. (1990) *Neural Darwinism.* Oxford: Oxford University Press.

Edelman, Gerald M. (1992) *Bright Air, Brilliant Fire*. Harmondsworth: Penguin.

El-Beltagy, Samhaa R., Hall, Wendy, de Roure, David and Carr, Leslie (2002) Linking in context. *Journal of Digital Information*, 2.

Elliott, Terry and Shadbolt, Nigel (2001) Growth and repair: instantiating a biologically-inspired model of neuronal development on the Khepera robot. *Robotics and Autonomous Systems*, 36: 149–169.

Elliott, Terry and Shadbolt, Nigel (2002) Developmental robotics: manifesto and application. *Proceedings of the Grey Walters Workshop*. Bristol: Hewlett Packard Labs.

Fazeli, Sam and Collingridge, Graham L. (eds) (1998) *Cortical Plasticity: LTP and LTD*. Oxford: Oxford University Press.

Foster, Ian and Kesselman, Carl (eds) (1998) *The Grid: Blueprint for a New Computing Infrastructure*. San Mateo, CA: Morgan Kaufmann.

Frege, Gottlob (1953) *The Foundations of Arithmetic* (trans. J.L. Austin), 2nd edn. Oxford: Blackwell.

Gardenfors, Peter (2000) *Conceptual Spaces: The Geometry of Thought*. Memory, Reasoning and Learning Research Review, Cambridge, CA: MIT Press.

Gibbs, W. Wayt (2002) Autonomic computing. *Sci. Am.*, 6 May. http://www.sciam.com/article.cfm?chanID=sa004&articleID=000B0152-8C15-1CDA-B4A8809EC588EEDF&pageNumber=1&catID=4.

Glenberg, Arthur M. (1997) What memory is for. *Brain Behav. Sci.*, 20: 1–55.

Greiner, Russell, Darken, Christian and Santoso, N. Iwan (2001) Efficient reasoning. *ACM Comput. Surv.*, 33: 1–30.

Hall, Wendy (1996) *Rethinking Hypermedia: The Microcosm Approach*. New York: Kluwer.

Hall, Wendy (2000) The button strikes back. *New Rev. Hypermedia and Multimedia*, 6: 5–17.

Harnad, Stevan (1990) The symbol grounding problem. *Physica D*, 42: 335–346.

Hodge, V.J. and Austin, J. (2001) An integrated neural IR system. In M. Verleysen (ed.) *Proc. 9th Eur. Symp. Artif. Neural Netw.* 25–27 April: 265–270.

Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci.*, 79: 2554–2558.

Hsu, William H. (2003) Control of inductive bias in supervised learning using evolutionary computation: a wrapper-based approach. In J. Wang (ed.), *Data Mining: Opportunities and Challenges*. IDEA Group Publishing; and at http://www.kddresearch.org/Publications/Book-Chapters/Hs2.pdf.

Jordan, Michael I. and Jacobs, R.A. (2002) Learning in modular and hierarchical systems, in M. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press.

Jordan, Michael I. and Weiss, Yair (2002) Graphical models: probabilistic inference. In M. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press.

Kaelbling, Leslie Pack, Littman, Michael L. and Moore, Andrew W. (1996) Reinforcement learning: a survey. *J. Artif. Intell. Res.*, 4: 237–285.

Killeen, Peter R. (1994) Mathematical principles of reinforcement: based on the correlation of behavior with incentives in short-term memory. *Brain Behav. Sci.*, 17: 105–172.

Kirsh, David (1991) Foundations of artificial intelligence: the big issues. *Artif. Intell.*, 47: 3–30.

Kobsa, Alfred (ed.) (2001) Tenth anniversary special issue of user modeling and user-adapted interaction. http://umuai.informatik.uni-essen.de/anniversary.html.

Kobsa, Alfred (2001a) Generic user modelling systems. *User Modeling and User-Adapted Interaction*, 11: 49–63.

Koriat, Asher and Goldsmith, Morris (1996) Memory metaphors and the laboratory/real-life controversy: correspondence versus storehouse views of memory. *Behav. Brain Sci.*, 19: 167–188.

McAllister, A., Katz, L. and Lo, D. (1999) Neurotrophins and synaptic plasticity. *Ann. Rev. Neurosci.*, 22: 295–318.

Mitchell, Tom M. (1991) The need for biases in learning generalisations. In June Shavlik and Thomas Dietterich (eds), *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Mitchell, Tom M. (1997) *Machine Learning*. New York: McGraw–Hill.

Oberauer, Klaus (2002) Access to information in working memory: exploring the focus of attention. *J. Exp. Psychol.: Learning, Memory and Cognition*, 28.

O'Hara, Kieron, Reichgelt, Han and Shadbolt, Nigel (1995) Avoiding omnidoxasticity in logics of belief: a reply to MacPherson. *Notre Dame J. Formal Logic*, 36: 475–495.

Page, Larry, Brin, Motwani, Sergey Rajeev and Winograd, Terry (1991) The PageRank Citation Ranking: Bringing Order to the Web. Stanford University Working Papers, http://dbpubs.stanford.edu:8090/pub/1999-66.

Pazienza, Maria Teresa (ed.) (1997) *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Berlin: Springer Verlag.

Pfeiffer, Rolf (1999) *Understanding Intelligence*. Cambridge, MA: MIT Press.

Purves, Dale (1988) *Body and Brain: A Trophic Theory of Neural Connections.* Cambridge, MA: Harvard University Press.

Purves, Dale (1994) *Neural Activity and the Growth of the Brain*. Cambridge: Cambridge University Press.

Smolensky, Paul (1991) Connectionism, constituency and the language of thought, in Barry Loewer and Georges Rey (eds.) *Meaning in Mind: Fodor and his Critics.* Oxford: Blackwell, pp. 201–227.

Steels, Luc (2001) Language games for autonomous robots. *IEEE Intelligent Systems,* 16 (5): 16–22.

Steels, Luc and Kaplan, Frederic (1999) Situated grounded word semantics, in T. Dean (ed.), *Proc. 16th Int. Joint Conf. Artif. Intell.*, 2: 862–867.

Sutton, Richard S. and Barto, Andrew G. (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Webb, Barbara (2001) Can robots make good models of biological behaviour? *Behav. Brain Sci.*, 24 (6): 1033–1050.

Further Reading

Bechtel, William and Abrahamsen, Adele (1991) *Connectionism and the Mind: An Introduction to Parallel Processing in Networks.* Oxford: Blackwell.

Boden, Margaret A. (ed.) (1990) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Gross, Richard and McIlveen, Rob (1999) *Memory*. London: Hodder and Stoughton.

IEEE Intelligent Systems Journal

Journal of Experimental Psychology: Learning, Memory and Cognition

Lyotard, Jean François (1979/1984) *The Postmodern Condition: A Report on Knowledge* (trans. Geoff Bennington and Brian Massumi). Manchester: Manchester University Press.

O'Hara, Kieron (2002) *Plato and the Internet*. Cambridge: Icon Books.

Further Browsing

Association of Computing Machinery (ACM) Computing Surveys, portal, http://portal.acm.org

Brain and Behavioral Sciences, including archive, http://www.bbsonline.org/

The Cogprints cognitive science e-print archive, http://cogprints.ecs.soton.ac.uk

For the grid, the Global Grid Forum, http://www.gridforum.org

The University of Alberta Cognitive Science Dictionary, http://www.psych.ualberta.ca/~mike/Pearl_Street/Dictionary/dictionary.html

The EQUATOR Interdisciplinary Research Collaboration, http://www.equator.ac.uk

The Advanced Knowledge Technologies Interdisciplinary Research Collaboration, http://www.aktors.org

6 APPENDIX: OPEN QUESTIONS AND POSSIBLE RESEARCH DIRECTIONS FOR THE FUTURE

Cognitive function	Issue	5-year horizon	20-year horizon
Memory	Content addressable memory and associative memory	Hardware implementations – how biologically plausible? How to understand retrieval in these terms?	Can these memory models be usefully informed about organizational structures found in cortical architectures?
	Multimedia	Facilitating retrieval. Multiresolution models	Multiresolution in neuroscience. Can novel modalities and media types be integrated into our current systems, indexed and retrieved?
	Different types of memory	Indexing and organization of large-scale personal and corporate multimedia memories	Understanding the roles of different memory mechanisms. Focus of attention. Conceptualizing the environment/context. Selecting storage or recollection mechanisms for particular tasks. Implicit knowledge
	Forgetting	Identifying obsolete knowledge. How to delete knowledge. How to deal with knowledge related to that which is forgotten	Understanding forgetting in terms of psychological theories and imputed mechanisms

(*continued*)

Cognitive function	Issue	5-year horizon	20-year horizon
	Symbol grounding	Neural nets as classifiers. Using autonomous agents and robots to bootstrap ontologies and symbol systems	Neural Darwinism and its consequences. More complex architectures than symbolic/subsymbolic distinction
Reasoning	Logic and alternatives	Neural nets as virtual machine for logic. Resource-bound reasoners. Continuing development of nonmonotonic reasoning	Using distributed computing to assemble large-scale computing power. Understanding the possibilities of special purpose reasoning directly implemented in hardware
	Uncertainty	Neural nets for reasoning with noisy data. Understanding uncertainty with respect to multimedia. Using parallel/probabilistic algorithms and 'graphical models'. Psychologically plausible models of probabilistic reasoning. Extending probabilistic principles to distributed/multi-agent environments. KA tools and techniques. Harvesting, indexing and annotating content from web-scale information stores	Focus of attention, biasing representations of the environment
	Multimedia	Retrieval techniques	Understand how to represent, retrieve and exploit modalities other than text, image, audio and video
	Usability	Understanding heterogeneous users. Legacy formats. Querying, linking and browsing	Recognizing and supporting task switching and patterns of workflow
	Indexing	Associative linking. User modelling. Using information retrieval techniques for determining context. Whether to compute index terms dynamically? Statistical v semantic approaches	Specifying a link lifecycle closer to human memory patterns. Combining retrieval techniques

(*Continued*)

Cognitive function	Issue	5-year horizon	20-year horizon
	The grid	Managing autonomous systems over heterogeneous domains and contexts. Retrieving, reusing, publishing and maintaining content. Problem-solving environments. Service brokerage	Move to a computing utility model where computing power is ubiquitous and pervasive. Significant amounts of autonomic capability
Learning	Personalization of content	Understanding the relationship between personalization of content and human learning. User modelling	Understanding the essential user (and context) attributes. Adapting user models to small-scale niche systems
	Reinforcement learning	Understanding the contribution of a single agent from a distributed system. Reinforcement under uncertainty/risk assessment. How to encode goals in rewards? Learning with complex and delayed consequences of action	Understanding context. Trade off between continuing to learn and settling on what the agent has learned. Decay theories of forgetting
	Plasticity	Investigating a range of plasticity methods in computational and robotic contexts	Understanding the contribution of mechanisms such as long-term potentiation to memory. Incorporating plasticity of architecture in the memory lifecycle
	Machine learning	Integrating statistics and semantics. Helpful biases. Fuzzy logic and other theories of uncertainty. Information extraction from natural language	Machine learning with neural nets. Very large-scale learning from internet scale content
	Problem representation	(Semi-)automation of bias learning. Problem understanding as a source of helpful biases. Changing representation of gene expression	Understanding and exploiting representations that use temporal encoding. Developing useful analogical representations

SECTION 5

Science Applied

Section Contents

This page intentionally left blank

CHAPTER

11

Advanced Neuroscience Technologies

Bashir Ahmed, Jon Driver, Karl Friston, Andrew Matus, Richard Morris and Edmund Rolls*

1 INTRODUCTION: TECHNOLOGICAL DEVELOPMENTS IN COGNITIVE AND IMAGING NEUROSCIENCE

Richard Morris

Unlike other Research Reviews in this series, this chapter relates to advanced techniques. The neurosciences are not unusual in science in having periodically been transformed by the advent of new techniques. Striking examples in recent memory include the advent of single-unit recording in awake animals by Hubel and Wiesel in the early 1960s, and the introduction of patch-clamping to observe single-channel ion-currents by Sakmann and Neher in the 1980s. Both developments were justly rewarded with Nobel Prizes to these pioneers.

* Authors are listed in alphabetical order. Authorship of specific sections is indicated in the text.

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

While a mature theoretical understanding of many issues in the brain sciences must marry understanding at different levels of analysis, research often arrives at obstacles that require new techniques to circumvent them. In this summary, we highlight three examples of new technologies that offer considerable promise for the future.

First, the widespread use of non-invasive human brain imaging has transformed cognitive neuroscience. Instead of relying on techniques from experimental psychology in the analysis of neurological patients, human brain imaging allows us to derive a spatial picture of brain activation in normal subjects as they undertake different tasks. Second, single-cell and multiple single-neuron recording technologies are advancing to the point where we can simultaneously record from large ensembles of cells, sometimes in different brain areas. In this way we can obtain data about how the firing of one or more cells may influence or be influenced by that of other cells. Third, optical imaging provides a new window on the brain, enabling microscopic techniques to reveal the molecular dynamics of activity within individual neurons. A wide variety of optical techniques are available, with the most recent using two-photon confocal microscopy.

Each technique can be coupled to others. Indeed, they often have to be. Functional brain imaging is of limited value without ingenious neuropsychological tests, refined to suit the constraints and opportunities afforded by brain scanning. Single-neuron recording must likewise be coupled to analytic behavioural tests in primates, rodents or other species before we can draw useful inferences about the significance of different patterns of cell firing. Optical techniques on their own are similarly limited, but can provide remarkable information when used in conjunction with appropriate physiological procedures or molecular-genetic probes such as green fluorescent protein.

In focusing on these three techniques, we do not mean to imply that they alone will transform neuroscience. Many other techniques will also be very important, ranging from novel behavioural techniques through to targeted manipulations of individual genes. Rather, with these three examples we aim to illustrate the importance of advanced technological development in a seemingly mature science such as neuroscience.

2 IMAGING THE HUMAN BRAIN

Karl Friston and Jon Driver

2.1 State of the Art

The past decade has seen a profound paradigm shift in cognitive neuroscience. This has been enabled, in large part, by several spectacular technological advances in functional brain imaging that offer entirely new ways of relating cognitive processes to their neural substrates. The ability to map measures of brain activity in response to particular cognitive or sensorimotor challenges now allows cognitive neuroscientists to think about brain function in explicit neurobiological terms (Friston, 2002). Moreover, the agenda of cognitive neuroscience has been transformed. The focus is no longer merely on the fractionation of mental processes into putative components, a feature of the era when the only obtainable data was from the careful study of neurological patients, but on how neuronal information processing and cognition actually proceed in the brain, and how different components, and different brain areas, influence each other.

Key developments in neuroimaging began in the 1980s with the application of positron emission tomography (PET) to measure evoked changes in regional cerebral blood flow. In the 1990s, functional magnetic resonance imaging (fMRI) came to the fore, using changes in cerebral haemodynamics as an intrinsic signal for measuring brain activations.

fMRI is increasingly supplanting PET in brain imaging. This is due not only to the wide availability of MRI scanners in medical and academic settings, but also to its higher spatial and temporal resolution. It also permits event-related designs that were previously impossible with PET, which could measure only summed activity over periods of 90 seconds or so.

Parallel developments in the recording of electrical brain signals, with electroencephalography (EEG), and magnetic brain signals, with magnetoencephalography (MEG), now allow us to measure evoked responses, either with a fine temporal resolution on the order of milliseconds (EEG and MEG), or with a spatial resolution of a few millimetres.

State-of-the-art fMRI neuroimaging allows us to acquire an image of the whole brain within a second or so. Thus we can generate several hundred images in a scanning session of half an hour or so. This is sufficient for brain mapping because the haemodynamic sequelae of neuronal responses have time-constants in the order of 4–8 seconds, meaning they only have to be sampled every second or so. Typical evoked responses are on the order of 0.1 to 1% of the average MRI signal, although scanners with higher magnetic-field strengths may offer not only higher signal-to-noise ratios but also larger percentage signal changes in some cases. This relatively small absolute magnitude of the changes observed is a serious worry to some critics of the use of fMRI, but robust statistical techniques can ensure that, although small, the reported changes are reliable and potentially replicable.

In a number of specialist units, concurrent EEG recording during fMRI is technically feasible, allowing temporally precise and spatially precise measures to be combined, but this still requires considerable expertise to remove the artefacts induced by MRI acquisition. In addition, several laboratories can now combine fMRI scanning with transcranial magnetic stimulation (TMS) of localized brain regions. This allows an assessment of the effects of disrupting one brain region on activations in remote but connected brain regions.

The more psychological features of the experimental design in neuroimaging have now become quite sophisticated, with some consensus about optimal approaches. Most designs are multifactorial and are often motivated by well-established paradigms in psychology, psychophysics or psychopharmacology. These designs may also relate to studies of brain function in animals from basic neuroscience.

It is now difficult to think of an area in cognitive or clinical neuroscience that has not been advanced by human functional neuroimaging at many levels. To give just a few examples, functional imaging has already substantially influenced research on perception, cross-modal integration, attention, spatial cognition, plasticity and learning, memory, language, intelligence, individual differences, and motor control in the normal brain. It has also advanced understanding of the effects of brain injury and neurological disease, and of pharmacological (drug) manipulations. We think that functional neuroimaging will play an increasing role in the study of cognitive and social development.

The prospects for non-invasive human brain imaging are exciting and diverse. Many open questions remain to be resolved at the technical, theoretical, empirical and applied levels. We will deal with these under the headings of neurophysiology, multimodal integration, biomathematics, computational neuroanatomy, clinical applications and new technologies. The final section on new technologies draws on some of the open questions established in preceding sections.

2.2 Neurophysiology

Fundamental questions in imaging neuroscience centre on the relationship between neuronal responses, the haemodynamic signals measured by fMRI, and the electrical and magnetic signals measured by EEG and MEG. These issues are important because they link non-invasive brain imaging to neuronal information-processing, both at a theoretical level and in relation to multi-neuron electrode recordings of neuronal responses, as in animal studies. The latter are clearly central to understanding non-invasive imaging results in terms of invasive work in basic neuroscience.

Initial work along these lines (see Logothetis and Wandell, 2004) has used either conjoint recording of electrical single neuron activity and fMRI signals in monkeys; or related measures of haemodynamics in smaller animals – e.g. optical imaging and laser-flow Dopplermetry. Although vital, these empirical set-ups are technically challenging. Only a handful of research groups around the world can implement them.

A major open question here is: 'Which exact aspects of neuronal responses produce the signals measured by non-invasive brain imaging, such as fMRI?' The possibilities currently range from simple things such as mean synaptic activity, to more complicated relationships that speak to the temporal dynamics of signal generation – e.g. specific fMRI correlates of oscillation at different frequencies in neuronal assemblies. A full understanding of these relationships will be predicated on plausible mathematical models of neuronal populations, and of how their microscopic organization leads to the emergence of measurable ensemble responses. We return to this theme below.

2.3 Multimodal Integration

One way forward is to combine different non-invasive measurements of human brain activity. In doing so, it is self-evident that fMRI and EEG/MEG have complementary strengths and weaknesses in terms of spatial versus temporal resolution. Harnessing the resolution of multiple modalities concurrently, within a single multimodal observation, is a major technical aim in the field.

While there has been some progress in coupling different measurements, a central outstanding issue remains to be addressed. To date, most fusion approaches merely harness the spatial precision of fMRI to provide constraints on the inverse source-reconstruction problem posed by EEG and MEG, i.e. resolving where the temporally well-specified EEG and/or MEG components come from spatially in the brain. In this context, under most current approaches, functional neuroimaging simply provides a prior spatial constraint. While this has certainly finessed the interpretation of EEG and MEG results, it does not represent true integration. A proper integration should allow the estimation of some aspect of functional anatomy that was hitherto inaccessible using either technique alone.

More formally, one would like to use multimodal data to estimate the parameters of a model that would otherwise be inestimable. This relies upon the construction of forward models of neuronal populations that can generate both electromagnetic and haemodynamic signals, an ambitious but not unthinkable objective for the next few years.

We have already made much progress on biophysical models, linking dendritic currents and local-field potentials to EEG signals measured at the scalp. There has been similar progress with models of how changes in mean synaptic activity are expressed haemodynamically, and thus detected by fMRI, through changes in blood volume, flow and oxygen content. Combining these two components with a model of interacting neuronal populations could lead to a complete forward model. The biologically meaningful parameters of this model would then be obtained by finding those parameters that produced a best match to observed data in both approaches.

The open questions here pertain to the nature of the underlying neuronal model. It is likely that neural mass and mean-field approximations will be useful (e.g. David and Friston, 2003). These approaches, based upon statistical physics and thermodynamics, rest upon modelling the dynamics not just of the states of a neuronal system, but rather of the probability densities of those states. Clearly, this vital enabling endeavour will require close collaboration between imaging neuroscientists and biologically invested physicists. Indeed, such interdisciplinary collaboration has been critical in every major breakthrough in functional neuroimaging to date.

2.4 Biomathematics and Functional Integration

Biomathematics underpins imaging neuroscience at two levels:

- It enables the development of increasingly sophisticated methods of data analysis.
- It drives the interpretation and motivation of brain mapping experiments from the viewpoint of computational neuroscience and theoretical neurobiology.

Open questions in the analysis of brain imaging data now particularly address the integration of anatomically dispersed brain systems – i.e. not just which brain areas are activated in a particular condition overall, but how does activity in one brain area influence activity in connected brain areas? This specific issue in neuroimaging relates to an increasing interest in neuroscience, and cognitive science more broadly, in how different components of the mind and brain work together. The notion of network operations is supplanting old ideas of strictly modular processes. In the neuroimaging context, the issue, referred to as 'functional integration', is usually thought of in terms of the effective connectivity among neuronal populations or cortical regions.

While many of the classical inference problems in analyses of spatially extended data (e.g. functional brain images) have now been resolved (by use of Gaussian field theory, for example), further work is required on the outstanding issue of how best to approach functional integration. There has been a recent move towards Bayesian inference and the incorporation of prior knowledge into estimation and inference procedures for imaging analysis (e.g. Friston *et al.*, 2003). This is particularly important in relation to the analysis of functional integration because the latter is a fundamental ill-posed problem that requires priors for its solution.

Research over the next few years will investigate such questions as 'How do the evoked responses in one brain area depend upon the state of others?' To answer such questions, we require biomathematical models of interacting brain areas that will likely harness the same mean-field approaches we described in the previous section. Models and estimation schemes of this sort are central to understanding the complex functionality of the brain. Not only will they enable understanding of how remote brain areas can influence each other in the normal brain, in different cognitive states, but they should also shed light on the effects of different forms of localized brain damage, and of pharmacologically induced neurotransmitter-specific manipulations.

At present, the forward or estimation models used for data analysis and characterizing empirically observed brain responses are not linked to mathematical models of brain function (e.g. generative models or predictive coding models of brain function). There is a wealth of understanding and expertise in machine-learning and computational neuroscience that may provide important mathematical constraints on possible functional brain architectures and how these might work.

A fundamental development over the next decade will involve attempts to use such formally specified theoretical models as explicit explanations for observed brain responses. To enable this, mathematical models from computational neuroscience of how we learn and perceive things, will have to be re-formulated in terms of the kind of estimation models used in data analysis. Again, such an approach will require collaboration between different disciplines, here including theoretical neurobiologists, computational neuroscientists, functional imagers and mathematicians/statisticians.

2.5 Computational Neuroanatomy

In addition to functional brain imaging of changes in neural activity under particular conditions, there have been substantial developments in the computational characterization of structural brain images of anatomy. This was the traditional clinical use of MRI prior to the advent of fMRI.

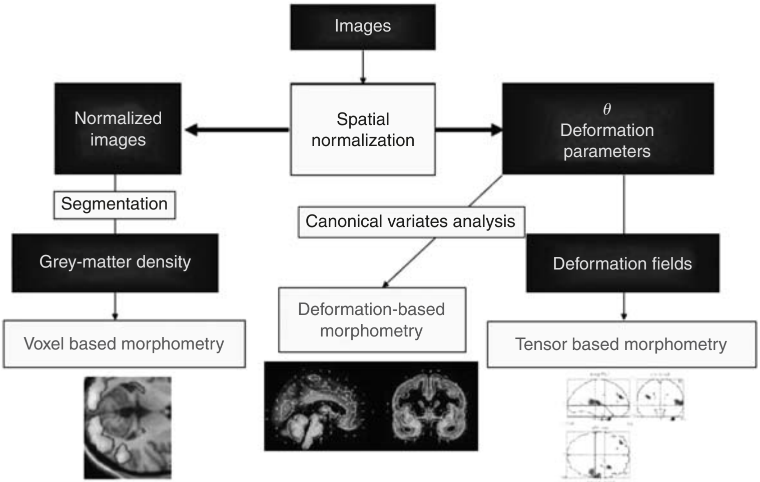


FIGURE 11.1 Computational neuroanatomy. (This figure appears in the colour plate section)

Initially, functional imagers saw variations between subjects in neuroanatomy largely as a 'nuisance' factor that had to be removed by analysis procedures prior to pooling functional data in a common or normalized brain-space. However, over the past few years, some analysis techniques used to remove such differences in individual anatomy (e.g. through non-linear warping and other morphological operations) have been applied to longstanding questions about anatomy *per se*.

The application of these techniques has become known as computational neuroanatomy. From the point of view of clinical neuroscience, this may have been as important as developments in functional brain imaging. For example, we can now detect, objectively, very subtle and previously undetectable changes in the neuroanatomy of various patient cohorts – e.g. schizophrenic, depressive, normal elderly, Alzheimer's, fragile-X etc (e.g. Good *et al.*, 2001). This is important because it provides an unbiased assessment of pathology that can be applied to the entire brain. Moreover, it can assess changes that are distributed in a structured fashion over multiple brain areas.

These techniques of structural anatomy rely on the statistical analysis and characterization of warping fields that map an individual's brain on to some reference template. Over the coming years, computational analysis techniques developed independently in machine vision and image processing could have a profound impact on the characterization of these warps, and should find important applications in clinical neuroscience. These developments will again rely upon interdisciplinary collaboration – e.g. here between clinical neuroscientists, neuroradiographers and experts in image processing.

Technological advances in diffusion-weighted imaging may enable a substantial further development in this field. Diffusion-weighted imaging allows one to assess the anisotropic diffusion of water in the brain. It can thus be an indirect measure of the integrity and orientation of white-matter tracks beneath the grey-matter formed by neurons – i.e. bundles of nerve processes that connect different brain regions, including areas that are otherwise anatomically remote. This is important because the ensuing data are high dimensional (tensor-fields) that may be sensitive not only to regionally specific changes in brain anatomy, but also to anatomical connections amongst these regions.

It will be impossible to analyse these data sets by eye. We will rely upon the continued development of computational neuroanatomic techniques. This is important for a clinical understanding of many neuropsychiatric disorders that may relate in part to subtle 'disconnections' between remote brain areas – as perhaps in dyslexia. It should also furnish critical evidence in the study of functional, rather than purely anatomical, integration between remote brain areas.

2.6 Potential Clinical Applications of Neuroimaging

The potential for applied clinical uses of recent technical advances in functional and structural neuroimaging seems enormous. Notable examples include the study and treatment of deficits following localized brain injury (e.g. caused by stroke or tumour) or other neural pathology (as in Alzheimer's etc). It could also be useful for psychiatric disorders (e.g. schizophrenia, depression, anxiety), developmental disorders (ranging from autism to dyslexia), including those of known genetic origin (e.g. fragile-X or Williams' syndrome), and for studying the functional neural effects of various treatment options, including pharmacological interventions.

Some advances we described in the previous section should lead to an increasingly sophisticated assessment of structural anatomical pathology in different clinical populations. Even more exciting is the potential for functional neuroimaging to shed light on how brain processes are altered in such populations, and how this constrains the observed deficit and the possibilities for recovery.

Sophisticated invasive studies of brain function in animals have, for many years, shown that lesions or disruptions to one brain region, or to fibres of passage there, can disrupt or modulate brain function in remote but interconnected areas, thus documenting how a functional network is influenced. Before the advent of functional imaging, such an approach was largely impossible in studies of human brain function. Such work was mainly restricted to correlating the local lesion-site with cognitive or behavioural impairment, as if each brain area operated in isolation. With fMRI we can now study how damage in one area or set of areas affects functional neural activity elsewhere, and how this correlates with the clinical deficit, relates to recovery and the effects of treatment.

It is a common cliché in neurology textbooks that an experienced clinician is the most sensitive instrument of measurement for pathology. However, one only has to look at how structural brain imaging has supplanted clinical examination as the method of choice for investigating lesion-site, to see how the textbooks may have to be rewritten to accommodate the new technology.

With fMRI, in principle we can now relate the success or failure of brain implants, surgery or stimulation to the functional networks found to be differentially activated. Likewise, we can now monitor pharmacological effects in the intact and damaged brain.

Unlike PET, we can apply fMRI repeatedly. We can track changes in relative brain activity under different conditions across days, months or even years. Indeed, such experiments have already documented previously unsuspected plasticity in the adult brain, both in the course of so-called 'spontaneous', but actually rather slow, recovery, and as a result of prolonged training. While conventional fMRI seems unlikely to be applied to very young babies – although fMRI in utero is now established – other imaging methods, such as optical imaging, can be administered at birth to assess brain function in relation to risk-factors such as oxygen deprivation etc.

In addition to providing critical information on how different forms of pathology influence neural networks, and how they respond to treatment interventions, we can foresee that functional imaging might also be routinely used in basic diagnosis of deficit for many clinical populations, and possibly even in legal contexts. For example, it could supplement behavioural tests of whether brain damage has produced, say, visual-field cuts or attentional deficits; memory-retrieval versus memory-consolidation deficits; hysterical or real paralysis; different forms of acquired or developmental dyslexia; even the extent of psychosis or psychopathy, and so on.

In addition to its clinical relevance, functional imaging of clinical populations should also provide information on issues that are fundamental to basic research on cognition. To give just one example, clinical studies of lesions have established that certain areas of frontal cortex seem intimately involved in 'executive processes' concerned with the planning and coordination of temporally extended behaviours in daily life. However, at present we know very little about how these frontal areas modulate brain activity elsewhere to produce such coordination. Once again, progress on such major issues is likely to require interdisciplinary collaboration – here between clinicians and basic researchers, involving both neuroscientists and computational modellers.

2.7 New Technologies

Pursuing the theme of connections between remote areas in the brain, the use of manganese as an MRI axonal tracer is currently being explored. Manganese can pass through as many as five synapses, thus allowing polysynaptic circuits to be imaged. Manganese gives good T1-based contrast for neuronal pathways in MRI under non-toxic low doses that might possibly be acceptable to humans. In addition, manganese can act as a long-lasting MRI stain for neuronal activity after temporary blood–brain barrier removal with mannitol or other appropriate agents. In this way, it might be used as a functional rather than purely anatomical tracer of connections that are more or less active under different cognitive conditions. This is a secondary application that will probably be limited to animal studies. The ability to do non-invasive studies of anatomical connections in humans could offer an enormous advance in understanding the way that the brain is connected and the anatomical constraints on functional architectures.

Conventional fMRI currently relies upon intrinsic signals mediated haemodynamically throughout the brain. But it is possible to use relatively low-abundance tracers – such as specific neurotransmitters, receptors and antagonists – by enhancing their signal. The use of labelled organic compounds – parahydrogen-based hyper-polarized carbon-13 – is being explored as one way of doing this. This could endow functional neuroimaging with a neurotransmitter or neuroreceptor specificity with enormous implications for pharmacological research on brain processes that relate to cognition.

Researchers can also use functional Magnetic Resonance Spectroscopy (MRS) using oxygen-17, carbon-13 and nitrogen-15 labelling and proton spectroscopy to investigate *in vivo* metabolic pathways in the human brain. This is potentially an extremely important area of research. The key issues are the cost of labelled substances and the dependency on high-field MRI systems.

The use of targeted MR contrast agents that can map gene expression and calcium concentration is a further area of future research. Although there is no current research in this area in the UK, there has been some progress the United States. The issues here are obtaining synthesis recipes and development of good organic chemistry at MRI sites in the UK.

It is possible that dendritic currents induce magnetic fields that could cause intra-voxel de-phasing and generate MRI signals by themselves (Xiong *et al.*, 2003). If this is true, MRI could detect the magnetic changes associated directly with neuronal responses (c.f. MEG). In principle, this would give the spatial resolution of MRI with the temporal resolution of MEG. This is potentially a very exciting prospect. However, the issues here are the size of the magnetic signal in relation to noise and the enormous signal averaging that would be required to measure an evoked response.

2.8 Interim Conclusion

In summary, many open questions centre on integrating different measurement modalities, or bringing together different disciplines in a common research endeavour, either to understand the nature of measured brain signals or to finesse the spatio-temporal resolution of those measurements. Both are predicated on more refined and plausible mathematical models of neuronal dynamics, for which advances in machine learning and computational neuroscience will be extremely valuable.

All these endeavours require a close link between imaging neuroscience, biomathematics and physics. New technologies are primarily aimed at conferring a pharmacological or anatomical specificity on the relatively non-specific techniques that are now available. This specificity will lend our understanding of brain function a much more mechanistic basis that will be central to developments in basic cognitive and clinical neuroscience.

3 MULTIPLE SINGLE-NEURON RECORDING: HOW INFORMATION IS REPRESENTED IN THE BRAIN – THE NEURAL CODE

Edmund Rolls

3.1 Introduction

While non-invasive functional brain imaging has made remarkable progress, it is sobering to reflect on the uncomfortable fact that they provide indirect measurements of brain activity, such as haemodynamic signals. This is not quite the same as recording from neurons themselves.

In order to understand better how the brain operates, we need to know how information is represented in the brain. Rapid advances in understanding the neural code are now on the horizon, because new techniques enable recordings from large numbers of single neurons simultaneously (e.g. 100), and because new techniques in information theory enable quantitative understanding of the neural code from such recordings.

Interdisciplinary teams of empirical neuroscientists and theoreticians trained in quantitative approaches in the physical sciences or mathematics need to come together to address the issue of neural encoding. Considerable resources are needed to record the activity of large numbers of single neurons simultaneously while large numbers of different stimuli are presented.

3.2 How do Populations of Neurons Represent Information?

Single neurons are the computational elements of the brain. They transmit information to other neurons by sending 'all-or-none' action potentials along axons to other neurons. Recording the activity of many neurons simultaneously would enable us to answer questions of the following type.

To what types of input do different neurons respond by altering their spiking activity? To perform computations, brain areas receive different types of input, with some inputs represented by the activity of a quite small proportion of neurons of each type. For example, the orbitofrontal cortex contains only 3.5% of neurons whose spiking activity is related to a mismatch between an expected reward and the actual reward obtained. However, while these neurons may be a small proportion, mixed with many others that respond to visual and taste stimuli, they appear to be an important part of its computation.

Timing At what time do neurons start to respond after a stimulus? And for how long do they respond? Do neurons in certain cortical areas maintain their activity in a stimulus-selective way after a stimulus has been removed, thus implementing short-term memory?

Information content How much information is in the number of spikes emitted by each neuron, and how much by synchronization? Is information contained primarily in the number of spikes that each neuron emits in a short time period (a rate code), or is there additional information present in the relative time of firing of different neurons, as has been championed by Singer (e.g. 1999)? For example, if two neurons fired together (i.e. became 'synchronized') during stimulus 1 but not stimulus 2, could this provide information that only stimulus 1 had occurred? The relative amount of information contributed in these two ways can now be addressed quantitatively.

The answer is to use information theory, originally developed by Shannon (1948). This is the only way to measure on an equal basis what we can learn about a stimulus or event from different sources in the spike code. In doing this, it is essential to measure how much evidence stimulus-dependent synchronization adds to the spike count code. Any synchronization information present may be redundant or quantitatively small compared to the information encoded in the spike trains of different neurons in an ensemble of neurons.

Applying information theory to the question of how information is contained in the number of spikes of different neurons is quite difficult, because of the sampling problem of obtaining enough trials of data to accurately estimate all the probabilities involved for all neurons having particular rates and degrees of synchronization for each stimulus in the set. Until recently, this prohibited the application of information theory techniques for rate *vs* synchronization to cases involving more than a very few neurons (Rolls *et al.*, 2003). However, a new approach uses decoding from neuronal responses and their relative times of firing to each stimulus of which stimulus has been seen (Franco *et al.*, 2004). The mutual information can then be calculated straightforwardly between the decoded stimulus, and that which was actually shown, in the way described by Rolls *et al.* (1997).

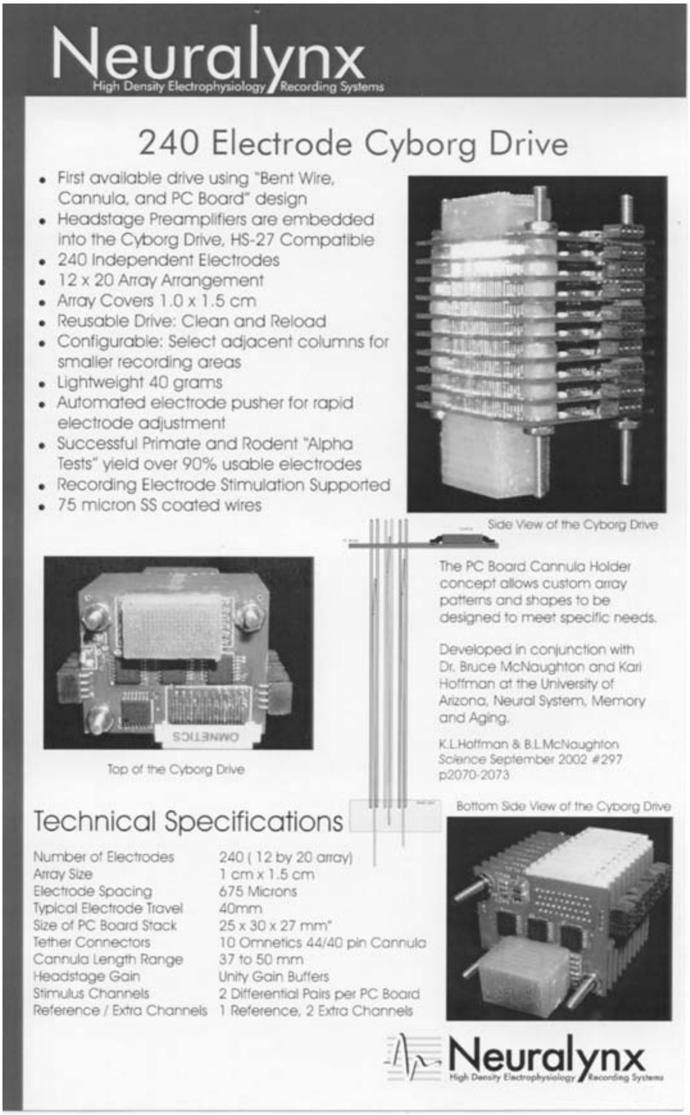
This new approach can apply to very large numbers of neurons, each with as many spikes as wished. The background to the measurement of information from populations of neurons, and many of the results obtained so far, are described by Rolls and Deco (2002).

Population coding How much information is encoded by populations with different numbers of neurons? The fundamental issue here is whether and how information increases as more neurons are added to the ensemble. Is there redundancy across the activity of different neurons, or does information increase monotonically (even linearly?) with the number of neurons in the ensemble?

Notwithstanding the excitement many feel and we have expressed above about non-invasive human brain imaging, it is also right to bring into the open the concerns of scientists who actually record signals from neurons. Specifically, neuroimaging with techniques such as fMRI does not, indeed cannot, answer issues of the timing of firing of coupled neurons (hence order), redundancy between neurons, the numbers of neurons activated for a given type of input, and response synchronization – none of which fMRI can measure. For this reason, fMRI, and neuroimaging in general, may not provide quite the hoped for route to understanding the neuronal network mechanisms that underlie neural computation. To understand how the computation works, it is necessary to know the details of the spiking activity of large numbers of neurons in each brain region.

3.3 New Methods for Recording Spiking Activity of Many Neurons

Recent developments enable simultaneous recording from many neurons to address the above issues. Some of these build on the 'stereotrode' or 'tetrode' developed by O'Keefe and his colleagues at University College London. One method provides up to 240 independently movable microelectrodes with the associated electronics, the Cyborg drive, http://www.neuralynx.com.



Neuralynx
High Density Electrophysiology Recording Systems

240 Electrode Cyborg Drive

- First available drive using "Bent Wire, Cannula, and PC Board" design
- Headstage Preamplifiers are embedded into the Cyborg Drive, HS-27 Compatible
- 240 Independent Electrodes
- 12 x 20 Array Arrangement
- Array Covers 1.0 x 1.5 cm
- Reusable Drive: Clean and Reload
- Configurable: Select adjacent columns for smaller recording areas
- Lightweight 40 grams
- Automated electrode pusher for rapid electrode adjustment
- Successful Primate and Rodent "Alpha Tests" yield over 90% usable electrodes
- Recording Electrode Stimulation Supported
- 75 micron SS coated wires

Side View of the Cyborg Drive

Top of the Cyborg Drive

The PC Board Cannula Holder concept allows custom array patterns and shapes to be designed to meet specific needs.

Developed in conjunction with Dr. Bruce McNaughton and Kari Hoffman at the University of Arizona, Neural System, Memory and Aging.

K.L.Hoffman & B.L.McNaughton
Science September 2002 #297
p2070-2073

Bottom Side View of the Cyborg Drive

Technical Specifications

Number of Electrodes	240 (12 by 20 array)
Array Size	1 cm x 1.5 cm
Electrode Spacing	675 Microns
Typical Electrode Travel	40mm
Size of PC Board Stack	25 x 30 x 27 mm"
Tether Connectors	10 Omnetics 44/40 pin Cannula
Cannula Length Range	37 to 50 mm
Headstage Gain	Unity Gain Buffers
Stimulus Channels	2 Differential Pairs per PC Board
Reference / Extra Channels	1 Reference, 2 Extra Channels

Neuralynx
High Density Electrophysiology Recording Systems

FIGURE 11.2 Neuralynx. (This figure appears in the colour plate section)

This technique has successfully analysed the activity of neurons in the parietal cortex (Hoffman and McNaughton, 2002).

Another method uses fixed arrays of 100 silicon-mounted electrodes for cortical recording (Donoghue, 2002; Nicolelis and Ribeiro, 2002). These developments reach beyond what is now being achieved in the UK (e.g. Baker and Lemon, 2000; Lenck-Santini *et al.*, 2002; Rolls *et al.*, 2003). Both systems, however, involve considerable logistics. If the field advances to the point where these technologies are essential, teams in the UK will need adequate support to enable them to perform such investigations.

3.4 Understanding the Computational Properties of Population Codes

Advances in neuron recording techniques will not only allow the issues raised above to be investigated, but will also enable us to understand the computational properties of the code used by the brain.

Robustness seems to be a key property of the population coding strategy that appears to be used by the cerebral cortex. Damage to a single cell will not, in general, have a catastrophic effect on the encoded representation because the information is encoded across many cells. This is not to deny that the activity of single or small numbers of cells can be important, as has been established by work on the perception of motion after microstimulation of neurons in area MT (V5).

However, population codes turn out to have other computationally desirable properties, such as mechanisms for noise removal, generalization to similar patterns, completion from a part, and the instantiation of complex, non-linear functions. Understanding the coding and computational properties of population codes has therefore become a main goal of computational neuroscience.

A concept of interest, which requires further development, is that if neurons have smooth Gaussian-like tuning to a stimulus dimension, these tuning functions may provide a basis for a non-linear mapping, which is the type of computation that is very important in brain function (Pouget *et al.*, 2000).

Another concept of interest is how the neurons that receive the activity in a cortical code can decode, or interpret, it. It appears that an operation by a neuron of a function as simple as forming a weighted sum of its inputs (dot-product decoding) can extract much of the information that is available in a population code (Robertson *et al.*, 1999; Panzeri *et al.*, 1999). It will be important in future to determine whether neurons could, and do, make more use of information that may be available in the firing of the sending neurons.

4 VISUALIZING MOLECULAR EVENTS AND INTRINSIC SIGNALS IN LIVING NEURONS

Andrew Matus and Bashir Ahmed

4.1 Introduction

The brain is alive, yet one might not realize this from looking at textbook pictures of stained neurons. Just as non-invasive brain imaging opened up a new world to cognitive neuroscientists ten years ago, optical imaging could be about to usher in a new revolution.

From its very beginning, the invasive investigation of brain function has depended on visualizing how nerve cells are arranged into functional circuits. The modern era in neuroscience began in the late nineteenth century, with the first histological techniques for seeing individual neurons and the fibre-like extensions, axons and dendrites that connect them.

Much of our current understanding of brain function has come from progressive exploration of neural circuits in different parts of the brain. These measurements use a combination of electrophysiological techniques to record the activities of individual neurons coupled with anatomical methods for tracing their connections.

Until recently, the anatomical part of this enterprise was limited to dead tissue. Although microelectrodes were used to record the activities of living neurons, the tracing of connections was always performed on dead tissue that had been fixed using chemical preservatives and then stained with coloured dyes to make nerve cells and their connections visible. These are severe limitations because the most important features of the brain's functional anatomy can be observed only in the living state. In particular, brain circuits are characterized by the property of plasticity by which functional relationships between nerve cells change, either at the level at which information is transmitted through existing connections or by the breaking and reforming of connections in new patterns. The lack of methods for visualizing the structure of living neurons has hitherto made it impossible to investigate this second, anatomical, aspect of plasticity – a crucial element in the brain mechanism of learning and memory.

Within the past few years, this situation has changed dramatically thanks to methods for making individual protein molecules visible within living cells, including brain neurons. Central to this development has been the discovery of genes, from marine animals such as jellyfish, which produce brightly fluorescent proteins inside living cells. Techniques from molecular genetics allow the introduction of these genes into the cells of mammalian species such as mice. There they produce fluorescent proteins that allow us to follow anatomical changes in live neuronal circuits.

More importantly, these same genes allow us to follow individual molecules inside nerve cells, opening up the possibility of discovering where and when molecular events underlying learning and memory take place in the adult brain. The potential of these new techniques for exploring the molecular and cellular mechanisms of brain function is vast. Equally promising is their scope for investigating disease states such as mental retardation or schizophrenia, where disturbances of nerve cell anatomy are implicated but still barely understood.

4.2 Basic Aspects of the Technical Advances

Several advances in different fields embracing molecular biology and microscopy have come together to make possible live cell imaging of brain neurons.

4.2.1 'Tagging' Neuronal Structures with Fluorescent Proteins

Genes derived from marine organisms that produce fluorescent proteins are indispensable for these new techniques. The archetype is green fluorescent protein (GFP), derived from the jellyfish *Aequorea victoria*. As its name implies, GFP radiates green light when illuminated at an appropriate frequency. The molecular mechanism underlying this effect, now well understood, is the basis for producing mutated versions of GFP with improved spectral properties. The first satisfactory red fluorescent protein (RFP) has recently been described (Campbell *et al.*, 2002).

Equally important for the use of fluorescent proteins is the unexpected discovery that these genes can be joined to a wide range of non-fluorescent genes without disturbing their natural function. This makes it possible to produce gene fusions – that is, a fusion protein in which a neuronal protein is joined directly to the GFP protein. The result is a combination of the two proteins in which the GFP works as a fluorescent 'tag' that makes the tagged neuronal process visible in a microscope designed to capture fluorescent light (see below).

An unexpected benefit of this approach is that even difficult proteins such as actin – difficult because it is highly sensitive to changes in its chemical structure – appear to work normally inside nerve cells when joined to GFP (Fischer *et al.*, 1998). Proof of concept has recently been obtained that we can use GFP and RFP to visualize simultaneously two proteins associated with different cytoskeletal structures inside live nerve cells that are associated with separate phases of circuit plasticity. The potential for extending this approach to visualize a larger number of gene products is clear.

4.2.2 Recent Advances in Microscopy Techniques

Fluorescent staining has been used to study cell structure for almost 50 years. However, capturing images from the minuscule amounts of light emanating from GFP-tagged proteins inside living cells makes special demands. This challenge has been met with advanced, and in some cases radically new, microscopy techniques. The most important of these is the use of highly sensitive electronic cameras based on sophisticated versions of the same charge-coupled devices (CCDs) used in modern video cameras.

Compared to conventional colour film, the sole capture medium available until a few years ago, electronic capture devices have several important advantages:

- They are far more sensitive than film to low levels of light, allowing the capture of faint images from living objects (cells) that are quickly damaged when illuminated by the strong light sources needed to produce images on conventional film.
- The signals produced by CCD cameras allow for electronic compensation to subtract the noise, the large constant background in a faint image. The remaining signal – the part of the image that actually represents the object being studied – can be amplified even before the image is recorded. Conventional film records all the light emerging from an image source. The weak image signal from a single GFP-tagged protein inside a fluorescent cell would be lost against the high background noise.
- Following on from the use of electronic capture devices, computer-based techniques have been developed, and are constantly being improved, that allow automated processing of raw image data to improve image quality and extracts precise measurements.

Another advance with great potential, arising from the application of electronic image capture devices, is the development of new types of microscopes that can form images from light emitted from cells deep inside brain tissue. These instruments operate in the principle of confocal microscopy and can form high-definition images of nerve cells inside brain tissue while ignoring out-of-focus light from other fluorescent cells above or below the cell being studied. The most advanced of these devices uses a quantum effect known as two-photon microscopy.

The principle of the two-photon microscope is that photons of light within a powerful illuminating beam combine at the focal point within the tissue to form photons at half the wavelength of the original. This has two decisive advantages. First, because the wavelength of the illuminating beam does not excite GFP, interference from out-of-focus fluorescent light above and below the point of focus is minimized. Second, the double-wavelength illuminating light used in two-photon microscopy is in the infrared region, where brain tissue is significantly more transparent compared to illuminating light used in conventional confocal microscopy. In combination, these two factors allow the imaging of structures significantly deeper within brain tissue than has previously been possible.

4.2.3 *Time-lapse Imaging*

An additional feature of modern microscopy – made possible by the widespread use of computer systems to capture, store and process biological images – is time-lapse recording. In this procedure images of living cells are taken repeatedly at set intervals. 'Played back' as a continuous 'movie' they can reveal dynamic changes in cell morphology or molecular dynamics. Examples of rapid changes in molecular dynamics in nerve cells or slow changes in cell shape in synaptic connections appear on the Matus and Svoboda websites: http://www.fmi.ch/members/andrew.matus/.

4.2.4 *Biosensors*

Several of the fluorescent proteins we have discussed are natural biosensors. They respond to physiological events with changes in fluorescence intensity. For example, some mutant forms of GFP are pH-sensitive and are less fluorescent in the acidic environment of an intracellular compartment such as a synaptic vesicle. Pilot studies have demonstrated the potential of this property for visualizing synaptic vesicle release (Miesenbock *et al.*, 1998). Another fluorescent protein, DsRed, changes colour from green to red as it 'matures' over a period of about 24 hours after synthesis suggesting that it may be useful as a 'fluorescent timer' for measuring turnover rates of physiologically important proteins or for tracing the 'history' of structures with which they are associated (Terskikh *et al.*, 2000).

Based on these naturally occurring examples, researchers have set out to design biosensors with enhanced sensitivities and with specificities. Among those that may prove valuable in neurobiological applications are engineered proteins for detecting local changes in membrane potential or the activity states of ion channels, for visualizing intracellular signalling molecules such as calcium and for determining the redox or the phosphorylation states of proteins (for further discussion see Zhang *et al.*, 2002). These applications still require extensive development and optimization before we can use them routinely in simple cell systems. It is presently impossible to predict when, if at all, they will be ready for analysing neuronal circuitry.

4.3 Optical Imaging of Intrinsic Signals in the Brain

Many of the techniques used on brain tissue *in vitro* can also apply *in vivo* to the living, dynamic brain. These generally use intrinsic signals, although the *in vivo* imaging of mice with GFP engineered into a subset of neurons is underway (Trachtenberg *et al.*, 2002). Technological advances, coupled with mathematical and statistical techniques, permit the detection of intrinsic signals where the signal magnitude embedded in noise is extremely small (<0.01%). The relevant imaging technologies are based on detection of signals emanating from optical variations (e.g., Intrinsic Signal Optical Imaging), magnetic perturbations (e.g., MEG), or radiotracer emissions (Single Photon Emission Tomography).

Intrinsic imaging exploits changes in the properties of reflected or transmitted light that can be observed through processes in the tissue. The earliest experiments involved light scattering from bundles of axons during the passage of action potentials. More recently, intrinsic signal imaging has been applied to exposed cortical tissue to delineate functional activations and has gained prominence in brain research through its association with fMRI. Extrinsic imaging is based on optical measurements that follow changes, such as the wavelength of reflected light, in the physical properties of a substance, usually a dye, as a function of stimulation of a tissue.

A new method for optical imaging, exploited in the Cognitive Neuroscience Research Centre at Oxford and other research laboratories, uses intrinsic signals to monitor activity over a wide area of the cerebral cortex (Grinvald *et al.*, 1986). In this method, light from a halogen source, a microscope lamp, passes through a narrow-band filter and illuminates the surface of the cortex. The reflected light focuses on a detector, originally a matrix of photodiodes. More recently, video cameras have become popular. They provide many more near simultaneous images (see figure in Bonhoeffer and Grinvald, 1996).

Put these basic units together with computer controlled stimulation and recording equipment, and it is possible to record optical images that reflect changes in the absorption of the light by the brain tissue. This absorption depends on external stimuli and is thus a way of looking at the brain's response to those stimuli. This system has been used *in vivo* in both anaesthetized and behaving animals (Grinvald *et al.*, 1999).

The primary advantage of these techniques is that they provide an overview of functional organization from multiple sites at a fine spatial resolution. In some cases, recordings are from a broad area of tissue, as much as tens of square millimetres, in an attempt to show the spatial relationship between many active neurons.

In other cases, we can make measurements at much higher magnification and can show regional variations in activity within a neuron, or between small numbers of neurons. In the latter configuration, the neurons are typically incubated in a dye that is sensitive to voltage, pH or an ion (e.g. Ca++). These dye-related signals can be very fast, with time constants as short as 1–2 microseconds. They thus offer advantages over other recording methods, such as intrinsic signal imaging.

A further advantage of optical image can be its non-invasive nature. In this respect, intrinsic signal imaging is clearly preferable. Some intrinsic signal experiments have been performed on the exposed human cortex.

A final advantage of optical imaging is the possibility of addressing many questions in a single experiment. In contrast to some methods (e.g. 2-DG, c-fos), researchers can present some stimulus, observe the response, then change the configuration of the stimulus. Experiments can also combine optical imaging with other methods. Various studies have observed brain dynamics using single-unit recording, and local drug or electrical stimulus application, in addition to natural stimulation.

Presently, with intrinsic signal imaging, resolution is approximately 50 × 50 micrometres over 8 × 8 millimetres of cortex, but with poor temporal resolution (sub-seconds range). Newer protocols, for example, presentation of temporally periodic stimuli or time-reversed stimuli, are increasing the spatial resolution and dramatically reducing the acquisition times (Kalatsky and Stryker, 2003).

With the use of dyes – voltage-sensitive dyes, for example (Wenner *et al.*, 1996; Antic *et al.*, 1999; Zochowski *et al.*, 2000; Arieli and Grinvald, 2002; Slovin *et al.*, 2002) – temporal resolution has reached the sub-millisecond range but a smaller region of cortex has to be imaged, a few millimetres square. In both cases, though, only a surface view is possible and events through the cortical depth are combined over a depth of 500–800 micrometres. Developments in camera/photodiode arrays, together with depth-scanning-based technology, would greatly aid this area of research, possibly allowing depth sectioning from the surface to the white matter. This will give a 3D image of activity within cortical tissue together with the temporal order of activation ('4D' images). There has already been some success in this direction (Maheswari *et al.*, 2003).

4.4 Non-invasive Optical Techniques: Near-infrared Spectroscopy

A narrow beam of light introduced at a point can penetrate the surface. The light is scattered within the medium, with a tiny proportion of this light emitted from the surface. The emitted light takes a quasi-semicircular path through the medium. A number of techniques exploit this phenomenon to detect anatomical landmarks or functional changes without invasive procedures. Complex time-resolved measurements of light intensity at the surface – to continuously monitor cerebral oxygenation and haemodynamics non-invasively, for example (Jobsis, 1977) – allows us to study cerebral oxygenation in newborn infants (Brazy *et al.*, 1985). Furthermore, greater depth of penetration is possible with light wavelengths in the near infrared (750–1000 nm) (http://www.medphys.ucl.ac.uk/research/borg/).

One non-invasive optical imaging technique is Optical Coherence Tomography (OCT). Laser-diodes, emitting light at wavelengths around 690–850 nm and at power levels from a few milliwatts to tens of milliwatts. The light from the diodes is intensity-modulated at high frequency (100–250 MHz). Optical fibres with a diameter less than 1 mm take the laser light to the surface: additional fibres a few millimetres in diameter convey the light emitted from the surface back to photomultipliers, CCD cameras or photodiodes to detect the emitted light.

There are two basic methods of OCT. A continuous method emits a constant light and detects changes in total light. The second, time-resolved, method uses light that is either intensity modulated at radiofrequency, the frequency domain method, or intensity modulated in the picosecond range, the time-domain method.

In the time-resolved methods, the frequency domain case measures the modulation amplitude and phase delay in response to an intensity-modulated signal. The time domain technique measures the temporal distribution of photons emitted between points on the surface in response to illumination by an impulse of light (Gratton *et al.*, 1994; Gratton and Fabiani, 2001; Obrig and Villringer, 2003; Hebden *et al.*, 2002). The measurement apparatus is relatively light and can be made into a head device, e.g., MONSTIR (multi-channel opto-electronic near-infrared system for time-resolved image reconstruction) (Hebden *et al.*, 2002).

To obtain high resolution with this approach, 3D images will require a very large number of recording channels, increasing the number of source-detector pairs, refined arrangement of probe arrays (Obrig and Villringer, 2003). With refinement of recording and analysis methods (TOAST, temporal optical absorption and scattering tomography; Arridge and Schweiger, 1997), it may be possible to attain on-line, high-temporal and spatial resolution of functional activity through a full cranial helmet (Gratton and Fabiani, 2001; Hebden *et al.*, 2002).

4.5 State of the Art in Optical Imaging

Although commercial instruments for two-photon microscopy are becoming available, the technique's use in cutting edge applications requires substantial investment in materials and manpower. It could benefit substantially from continued technical development.

Recent imaging studies in American laboratories have shown that it is possible to follow experience-induced changes in numbers of synapses in the brains of living mice over periods of days (Trachtenberg *et al.*, 2002). While exciting, this time scale is still slow compared to the dynamic events thought to underlie learning and memory mechanisms in the brain. Moreover, the techniques can as yet be applied only to anaesthetized and restrained animals. However, the rapid pace of development suggests that successors to these techniques will increasingly allow realistic exploration of dynamic events in brain circuits. For example, techniques are being developed for attaching a miniature fibre-optic confocal microscope to the skull of a live rat (Helmchen *et al.*, 2001). The technical challenges are formidable, but the potential for illuminating, literally, molecular events inside the brains of freely moving animals suggests that, if successful, such techniques will revolutionize our understanding of brain function.

Developments that may contribute significantly to progress will be the adaptation of techniques using green fluorescent protein to animals biologically and behaviourally more closely related to humans. For example, a programme to develop transgenic primates, such as marmosets, for GFP-based imaging in neurons in the central nervous system via collaboration between a dedicated Brain Neuron Imaging Group, and a consortium working with tractable primates, such as the European Marmoset Research Group, may provide considerable benefits, such as bringing diverse groups together: http://www.dpz.gwdg.de/emrg/emrgcons.htm.

Such an initiative may also provide early identification of opportunities for diagnostic and therapeutic approaches. We should not discount the possible application of fluorescent imaging in human surgical procedures, complementing present diagnostic technology such as fMRI.

It is impossible to predict either the ultimate form nor the likely success of such advanced techniques. However, the power and scope of the overall approach argues strongly for an investment in the underlying imaging technology.

Further Reading on Optical Imaging

Matus, A. (2000) Actin-based plasticity in dendritic spines. *Science*, 290: 754–758.

Zhang, J., Campbell, R.E., Ting, A.Y. and Tsien, R.Y. (2002) Creating new fluorescent probes for cell biology. *Nat. Rev. Mol. Cell Biol.*, 3: 906–918.

Trachtenberg, J.T., Chen, B.E., Knott, G.W., Feng, G., Sanes, J.R., Welker, E. and Svoboda, K. (2002) Long-term *in vivo* imaging of experience-dependent synaptic plasticity in adult cortex. *Nature*, 420: 788–794.

5 GLOSSARY OF TERMS

biosensor A biological molecule, usually a protein, that produces a measurable signal, usually optical, in response to a change in cell physiology.

confocal microscopy A system of illumination that rejects out-of-focus light from biological images. There are several different confocal systems that varying in technical complexity and suited to different applications.

cytoskeleton/cytoskeletal Structures composed of protein filaments that support and control the shapes of cells. The cytoskeleton is important in nerve cells for its role in controlling circuit connectivity.

ensemble recording Electrophysiological recordings from a large number of single-cells simultaneously. This is not to be confused with multi-unit recording in which single-cells cannot be identified.

fixed/fixation Fixation is the process of chemically treating biological tissue to preserve their structure after death. Fixed tissue can be stored for long periods and, with appropriate techniques, the detailed anatomy of cells can be preserved down to the level of molecular structure.

functional magnetic resonance imaging (fMRI) A neuroimaging method that uses the different magnetic susceptibility of deoxyhaemoglobin compared to oxyhaemoglobin to measure changes in the activity of a brain area, which are reflected in its metabolism and thus need for oxygen. Typical spatial resolutions in human studies are 1–3 mm × 1–3 mm × 1–3 mm, and the temporal resolution is in the order of seconds. The method is non-invasive, and can be repeated on an individual subject many times.

MONSTIR Multi-channel opto-electronic near-infrared system for time-resolved image reconstruction; images brain tissue non-invasively.

positron emission tomography (PET) A neuroimaging method that uses radioactive isotopes to estimate the blood flow in a brain region, which reflects the metabolism of the brain region and through this the neural activity. Typical spatial resolutions in human studies are 5 mm × 5 mm × 5 mm, and the temporal resolution is in the order of 90 seconds.

staining With a few notable exceptions, the molecules that make up biological tissue are colourless. To render them visible in the microscope, fixed tissues are treated with reagents that selectively stain individual structures within a tissue, usually by reacting specifically with a particular molecule such as a protein.

synchronization A brain-state in which neurons that are spatially separated show spiking activity that is closely temporally coupled, typically with spikes of one neuron occurring within a few milliseconds of the spikes from another neuron. The existence of synchronized states is widely thought to be relevant to the binding problem in cognitive neuroscience.

two-photon microscopy A system of confocal microscopy that exploits a quantum effect to enhance the selective imaging of in-focus light while additionally increasing the penetration of illuminating light into biological tissue.

References

Antic, S., Cohen, L.B., Lam, Y-W., Wachowiak, M., Zecevic, D. and Zochowski, M. (1999) Fast multisite optical measurement of membrane potential: three examples. *FASEB J*, 13: S271–S276.

Arieli, A. and Grinvald, A. (2002) Optical imaging combined with targeted electrical recordings, microstimulation, or tracer injections. *J. Neurosci. Methods*, 116: 15–28.

Arridge, S.R. and Schweiger, M. (1997) Image reconstruction in optical tomography. *Phil. Trans. R. Soc. Lond. Ser. B, Biol. Sci.*, 352: 717–726.

Baker, S.N. and Lemon, R.N. (2000) Precise spatiotemporal repeating patterns in monkey primary and supplementary motor areas occur at chance levels. *J. Neurophysiol.*, 84: 1770–1780.

Bonhoeffer, T. and Grinvald, A. (1996) Optical imaging based on intrinsic signals: the methodology. In A. W. Toga and J. C. Mazziotta (eds), *Brain Mapping: The Methods.* San Diego, CA: Academic Press, pp. 55–97.

Brazy, J.E., Darrell, V., Lewis, M.D., Mitnick, M.H. and Jobsis, F.F. (1985) Noninvasive monitoring of cerebral oxygenation in preterm infants: preliminary observations. *Pediatrics*, 75: 217–225.

Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proc. Natl Acad. Sci. USA*, 99: 7877–7882.

David, O. and Friston, K.J. (2003) A neural mass model for MEG/EEG: coupling and neuronal dynamics. *Neuroimage*, 20:1743–1755.

Donoghue, J.P. (2002) Connecting cortex to machines: recent advances in brain interfaces. *Nature Neurosci. Supplement* 5, 1085–1088.

Fischer, M., Kaech, S., Knutti, D. and Matus, A. (1998) Rapid actin-based plasticity in dendritic spines. *Neuron*, 20: 847–854.

Franco, L., Rolls, E.T., Aggelopoulos, N.C. and Treves, A. (2004) The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons. *Exp. Brain Res.*, 155: 370–384. Reprint available at http:// www.cns.ox.ac.uk.

Friston, K. (2002) Functional integration and inference in the brain. *Progr. Neurobiol.*, 68: 113–143.

Friston, K.J., Harrison, L. and Penny, W. (2003) Dynamic causal modelling. *Neuroimage*, 19: 1273–1302.

Good, C.D., Ashburner, J. and Frackowiak, R.S. (2001) Computational neuroanatomy: new perspectives for neuroradiology. *Rev. Neurol. (Paris)*, 157: 797–806.

Gratton, G. and Fabiani, M. (2001) The event-related optical signal: a new tool for studying brain function. *Int. J. Psychophysiol.*, 42: 109–121.

Gratton, G., Maier, J.S., Fabiani, M., Mantulin, W.W. and Gratton, E. (1994) Feasibility of intracranial near-infrared optical scanning. *Psychophysiology*, 31: 211–215.

Grinvald, A., Shoham, A.D., Shmuel, A., Glaser, D.E., Vanzetta, I., Shtoyerman, E., Slovin, H., Wijnbergen, C., Hildesheim, R., Sterkin, A. and Arieli, A. (1999) In-vivo optical imaging of cortical architecture and dynamics. In U. Windhorst and H. Johansson (eds), *Modern Techniques in Neuroscience Research.* Berlin: Springer Verlag, pp. 893–969.

Grinvald, A., Anglister, L., Freeman, J.A., Hildesheim, R. and Manker, A. (1984) Real time optical imaging of naturally evoked electrical activity in intact frog brain. *Nature*, 324: 361–364.

Grinvald, A., Lieke, E., Frostig, R.D., Gilbert, C.D. and Wiesel, T.N. (1986) Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature*, 324, 361–364.

Hebden, J.C., Gibson, A., Yusof, R.Md, Everdell, N., Hillman, E.M.C., Delpy, D.T., Arridge, S.R., Austin, T., Meek, J.H. and Wyatt, J.S. (2002) Three-dimensional optical tomography of the premature infant brain. *Phys. Med. Biol.*, 47: 4155–4166.

Helmchen, F., Fee, M.S., Tank, D.W. and Denk, W. (2001) A miniature head-mounted two-photon microscope. high-resolution brain imaging in freely moving animals. *Neuron*, 31: 903–912.

Hoffman, K.L. and McNaughton, B.L. (2002) Coordinated reactivation of distributed memory traces in primate neocortex. *Science*, 297: 2070–2073.

Jobsis, F.F. (1977) Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198: 1264–1267.

Kalatsky, V.A. and Stryker, M.P. (2003) New paradigm for optical imaging: temporally encoded maps of intrinsic signal. *Neuron*, 38, 529–545.

Lenck-Santini, P-P., Muller, R.U., Save, E. and Poucet, B. (2002) Relationships between place cell firing fields and navigational decisions by rats. *J. Neurosci.*, 22: 9035–9047.

Logothetis, N.K. and Wandell, B.A. (2004) Interpreting the BOLD signal. *Ann. Rev. Physiol.*, 66: 735–769.

Maheswari, R.U., Takaoka, H., Kadono, H., Homma, R. and Tanifuji, M. (2003) Novel functional imaging technique from brain surface with optical coherence tomography enabling visualization of depth resolved functional structure in vivo. *J. Neurosci. Methods*, 124: 83–92.

Matus, A. (2000) Actin-based plasticity in dendritic spines. *Science*, 290: 754–758.

Miesenbock, G., De Angelis, D.A. and Rothman, J.E. (1998) Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins. *Nature*, 394: 192–195.

Nicolelis, M.A.L. and Ribeiro, S. (2002) Multi-electrode recordings: the next steps. *Curr. Opin. Neurobiol.*, 12: 602–606.

Obrig, H. and Villringer, A. (2003) Beyond the visible – imaging the human brain with light. *J. Cerebr. Blood Flow Metabol.*, 23: 1–18.

Panzeri, S., Treves, A., Schultz, S. and Rolls, E.T. (1999) On decoding the responses of a population of neurons from short time epochs. *Neural Computation*, 11: 1553–1577.

Pouget, A., Dayan, P. and Zemel, R.S. (2000) Information processing with population codes. *Nature Rev. Neurosci.*, 1: 125–132.

Robertson, R.G., Rolls, E.T., Georges-François, P. and Panzeri, S. (1999) Head direction cells in the primate pre-subiculum. *Hippocampus*, 9: 206–219.

Rolls, E.T. and Deco, G. (2002) *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rolls, E.T., Franco, L., Aggelopoulos, N.C. and Reece, S. (2003) Application of an information theoretic approach to analysing the contributions of the firing rates and correlations between the firing of neurons. *J. Neurophysiol.*, 89: 2810–2822. Reprint available at http://www.cns.ox.ac.uk.

Rolls, E.T., Treves, A. and Tovee, M.J. (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res.*, 114: 149–162.

Shannon, C.E. (1948) A mathematical theory of communication. *AT&T Bell Labs Tech. J.*, 27: 379–423.

Singer, W. (1999) Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24: 49–65.

Singer, W., Engel, A.K., Kreiter, A.K., Munk, M.H.J., Neuenschwander, S., Roelfsema, P. (1997) Neuronal ensembles: necessity, signature and detectability. *Trends Cogn. Neurosci.*, 1: 252–261.

Slovin, H., Arieli, A., Hildesheim, R. and Grinvald, A. (2002) Long-term voltage-sensitive dye imaging reveals cortical dynamics in behaving monkeys. *J. Neurophysiol.*, 88: 3421–3438.

Terskikh, A., Fradkov, A., Ermakova, G. *et al.* (2000) 'Fluorescent timer': protein that changes color with time. *Science*, 290: 1585–1588.

Trachtenberg, J.T., Chen, B.E., Knott, G.W., Feng, G., Sanes, J.R., Welker, E. and Svoboda, K. (2002) Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. *Nature*, 420: 788–794.

Wenner, P., Tsau, Y., Cohen, L.B., O'Donovan, M.J. and Dan, Y. (1996) Voltage-sensitive dye recording using retrogradely transported dye in the chicken spinal cord: staining and signal characteristics. *J. Neurosci. Methods*, 70: 111–120.

Zhang, J., Campbell, R.E., Ting, A.Y. and Tsien, R.Y. (2002) Creating new fluorescent probes for cell biology. *Nature Rev. Mol. Cell Biol.*, 3: 906–918.

Xiong, J., Fox, P.T. and Gao, J.H. (2003) Directly mapping magnetic field effects of neuronal activity by magnetic resonance imaging. *Human Brain Mapping*, 20: 1–9.

Zochowski, M., Wachowiak, M., Falk, C.X., Cohen, L.B., Lam, Y-W., Antic, S. and Zecevik, D. (2000) Concepts in imaging and microscopy: imaging membrane potential with voltage sensitive dyes. *Biol. Bull.*, 198: 1–21.

Web Sites

http://svobodalab.cshl.edu/

This site contains time-lapse recordings of changes in dendritic spine numbers in brains of living mice over periods of several days.

http://www.dpz.gwdg.de/emrg/emrgcons.htm

The URL of the European Marmoset Group.

http://www.opt-imaging.com

An optical imaging website.

CHAPTER

12

Applications and Impact

Bill Sharpe

Lights that come on when there are prowlers about can hardly lay claim to intelligence, but they are an early step on the way towards cognitive systems with increasing powers to sense, understand and react to the world around them. Imagination is the only limit to the technologies that could arise from our growing understanding of how living cognitive systems behave, and the ability to use that knowledge in the design of artificial systems.

1 SETTING THE SCENE

1.1 What are Cognitive Systems?

Cognitive systems – natural and artificial – sense, act, think, feel, communicate, learn and evolve. We see these capabilities in many forms in living organisms around us. The natural world shows us how systems as different as a colony of ants or a human brain achieve sophisticated adaptive behaviours.

Until recently we would not have used use such terms to describe artificial systems, but as computing power grows, and pervades more and more things, we find a growing convergence of interests and vocabulary between the understanding of natural cognitive systems and the emerging opportunities and challenges of building 'smart things'.

By definition, therefore, the field of cognitive systems involves the interplay of interests and outcomes across the life sciences, physical sciences and real-world engineering. If you are a life scientist then

Cognitive Systems: Information Processing Meets Brain Science ISBN 0-12-088566-2

you want to know how the naturally occurring systems work, and perhaps how to fix problems they develop. If you are a physical scientist or engineer, then you are trying to build and understand new things. Naturally occurring cognitive systems are a source of inspiration and challenge – you might want to find out how to do what they do, interact with them, simulate or enhance them, or to beat them at their own game.

Here I provide some overall context for discussing this interplay of interests between natural and artificial cognitive systems, and underpin the discussion of applications. There are three sections:

- the biological computer and the artificial brain: basic parameters of brains and computers that explain why the next 30 years are a defining time for the field
- characteristic capabilities: a high-level description of what a cognitive system is, at least for the purposes of this review
- motivations: a taxonomy of the variety of motivations and interests that we can think of as lying within the field of cognitive systems, and hence within the scope of this review.

1.2 The Biological Computer and the Artificial Brain

1.2.1 Hofstadter's Law

> It always takes longer than you expect, even when you take into account Hofstadter's Law. (Hofstadter, 1979: ch. 5).

How powerful is the human brain, and how should we compare it with the power of a computer? How much computational power do you need to drive a car, to recognize a terrorist, to speak and comprehend fluently? How powerful are different animal brains? For example, how powerful is the brain of an insect – what is it good at, how does it do it, and can we build things that work that way? Would it be good if we could?

These questions are fundamental to assessing the rate at which new applications of cognitive systems will appear. We are still a long way from having the answers. At present we can build our roadmaps and predictions around key observations:

- We have a massive information explosion about the detailed operations of natural nervous systems at the micro or macro level – but this does not yet answer the simple question of how they *work*. All that we can be confident of is that they are very different from digital computers.
- Orders of magnitude comparisons suggest that cheap pervasive computing power will come within the range of humans over the coming 30 years, and will certainly overtake simpler creatures before then.
- Artificial systems may supersede many areas of natural cognition once we get within appropriate performance bands – even though they do things differently. Many apparently hard things, like playing chess, will turn out to be easy.
- Some areas of natural performance will prove surprisingly resistant to emulation. Many apparently easy things, like getting around in the world, will turn out to be very hard.

Von Neumann provided fundamental arguments for comparing computing power between computers and brains: he was one of the first to attempt to make specific estimates of human power in computational terms. The basic approach is to compare power in terms of number of computations per second, based on an abstract equivalence of different underlying architectures. One example of this calculation (Merkle, 1989) puts the upper bound of human brainpower at around 2 × 10^{16} calculations per second:

Number of synapses	= 10^{15}
Operations per second	= 10
Synapse operations per second	= 10^{16}

To put this in perspective, if we assume that computational power continues to grow according to Moore's law – with the processing power of an integrated circuit doubling every two years or so – then in 2030 you will be able to buy a £1000 PC that computes at around 10^{16} instructions/second. How should we use such calculations in predicting the future of cognitive systems?

Optimistic commentators conclude that there are no conceptual or practical barriers to prevent artificial cognitive systems superseding the capabilities of the human brain on roughly the same timescale that this basic comparison of power suggests (see for example http://www.kurweilai.net). However, there is consensus in the research community that it is fundamentally simplistic to use this simple power comparison to draw any general conclusions about computers matching brains. At present, we lack the most basic understanding of many aspects of the performance of neural systems. We already know a lot about such things as neural architecture, the firing rates of different kinds of cells, and the spatial distribution of the many distinct types of synaptic inputs. Then there are the global and site-specific effects of neurotransmitters.

Advances in molecular neuroscience are teaching us more every day about signal transduction pathways, gene activation and protein synthesis in the brain. So, even if (a big if) we were to find a way to match the physical structure of certain networks of the brain, we may well find that we need to understand whole dimensions of performance before we can reach the same type of power.

Even supposing that we have matched the basic computational power of natural nervous systems, then we are still at the very early stages of understanding the architecture that will turn 10^{16} instructions/second into the extraordinary range, diversity and adaptability of performance of the cognitive systems we see around us. After all, we already have a lot more computing power available to us now in grid networks, but apart from a few high profile specialized tasks, such as chess playing, we are a very long way from knowing how to put it together into high-level cognitive performance.

However, even with all these caveats, we are already in a very different situation than in the past. This is because it is reasonable to suppose that the computing power available to us is approaching that of natural cognitive systems. We will, therefore, be tackling problems of the organization and architecture of cognitive systems knowing that we have the basic engine power required. This may allow what we might think of as a 'Cambrian explosion' of new systems and applications.

From the perspective of discussing the evolution of applications of cognitive systems, the most important change ahead is that the development of cognitive systems will be part of a diversification of computing into many lines of development – each serving as a focus for huge areas of commercial and societal investment.

An essential aspect of Moore's law is that it is a self-fulfilling prophecy. The industry commits itself to an agenda and marshals the resources needed to achieve it. While this continues to generate smaller, faster, cheaper computing power, it is being progressively coupled with other technologies and specialized into new branches of technology. Each branch then becomes a new focus of societal development – harnessing resources from research, industry and society at large.

Robotics is a major example of a 'branch' of cognitive systems. The overall tree may look something like the diagram in Fig. 12.1.

However, while this structure looks reasonable, we cannot predict with any confidence the 'shape of things to come'. As stated above, we currently lack a good characterization of how natural cognitive systems work. So we have no straightforward way to predict progress on higher-level capabilities in terms of fundamental computing power. Nevertheless, in the remaining sections I will attempt to provide an overview of how fundamental work on cognitive capabilities will relate to these various branches.

1.3 Characteristic Capabilities

A constructive way of thinking about cognitive systems is to characterize them in terms of a short list of high-level capabilities.

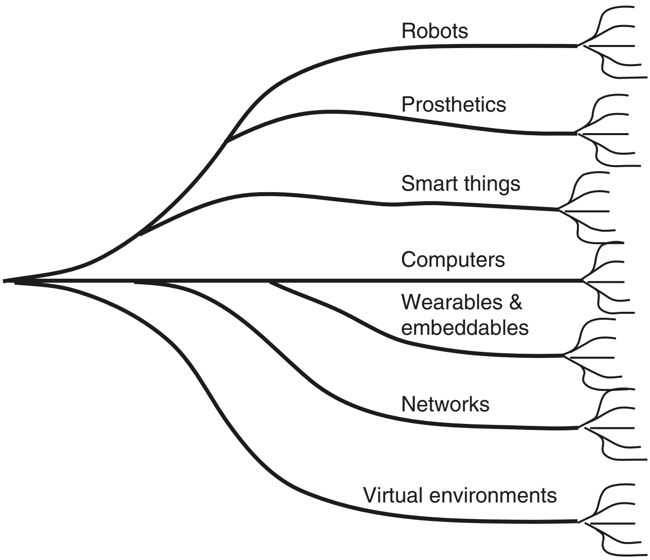


FIGURE 12.1

These include capabilities that we see throughout the animal kingdom and that we would like to both understand in their natural state and to give to artificial systems. These capabilities also include some that are specific to humans, such as language and speech.

1.3.1 Sense and Perceive

If things are going to behave usefully, then the first step is that they should connect with what is going on around them. We already take this for granted in many simple things, such as the thermostat that controls our central heating, or traffic lights that detect arriving traffic. Sensors are developing rapidly in every modality – sight, hearing, touch, pressure, balance, orientation. At higher levels we are learning how to realize sophisticated perception of scenes and objects within them, and to link these to different sorts of action. At low levels we are understanding how insects and simple creatures use algorithms that we can copy for such tasks as landing, and avoiding looming objects. At high levels we can start to endow machines with sophisticated object-level recognition of their environment.

Where the task of sensation and perception gets difficult is when we seek to fuse information from different senses, or to categorize sensory information in a knowledge-based way. How do we move from systems that see that an object is round, can tell that it can be held in a hand, realize that it is not very heavy through to classifying such a perceived object as a 'cup'?

1.3.2 Act

Automation is almost the defining word of the industrial revolution and a dominant theme of IT application. Yet we are still only at the early stages of giving machines the flexibility of action that we see in the natural world:

- balance on two legs, walk, run and jump
- pick things up, handle soft and flexible things
- play the piano
- climb a tree.

There is a growing understanding in the life sciences of the close relationship between perception and action, and how understanding one can inform the other.

1.3.3 *Think: Recognize, Recall, Compare, Reason, Decide, Plan, …*

Where computers excel today is in handling symbols, media and anything that we can give a formal symbolic description. What they find hard are the everyday sorts of reasoning that we think of as 'common sense', and such everyday skills as getting about without bumping into things or getting lost.

- The hard things for people are easy for computers …
  - Calculating, sorting, storing, searching and things that can be done that way, like designing complicated things, chess, logistic planning, web searching …
- The easy things are hard …
  - scene understanding
- understanding language
- picking things up and putting them together
- planning a path across the room
- speech.

The emergence of robotics as a major field of computational endeavour has created a strong convergence of interests in how viable agents can solve everyday problems.

1.3.4 *Feel: Have Emotions, Motivations and Relate to Others*

This is an area where we quickly run into trouble for lack of language. Feelings and emotions are so typically human, and so much a part of our internal experience, that ascribing them to machines runs straight into the 'big' questions of what machines can and cannot do, and their internal states. But there is a growing understanding between physical and life sciences of how motivational states underlie our shifting pattern of actions and thoughts, and are essential to the way we maintain many competing priorities for action over multiple timescales. Emotions are very practical – they create conditions for orchestrating whole classes of action and giving preference to one form of response over another, such as 'fight or flight', 'rest and digest', 'lust and trust'. Artificial systems will need similar repertoires to manage the range and diversity of responses available to them.

When artificial systems relate to humans, then issues of ascribing emotions to them immediately come to the fore. The past few years have seen people readily respond to such things as Tamagotchis, AIBO dogs and other toys that set out not to be 'intelligent' but to create a relational experience with humans through an explicitly designed set of 'emotions'. People, it seems, are ready to take things at 'interface value'.

1.3.5 *Communicate with Each Other*

Bacteria, ants, bees, lions, dolphins, chimps, humans – all communicate with each other over a huge range of sophistication, fulfilling individual and collective goals. The interaction of artificial agents is still in its infancy and has hardly been explored, but we are already encountering the first unexpected effects in such areas as financial trading systems and electricity networks. Emergent effects of many interacting intelligent agents will be both a source of new possibilities and new headaches. The term 'swarm computing' is being used for the construction of systems from large numbers of very simple agents in either the physical or virtual world.

While we struggle to understand how to predict and use such behaviour even at its simplest, there is no reason why future artificial systems should be limited to the complexity level of natural languages. There are suggestions that the size of our working memory relates closely to the rolling requirements of language processing, and that brain capacity and language skills influence the size of primate animal communities. Artificial systems will be able to enjoy the possibilities of much expanded language complexity in their communities.

1.3.6 *Learn*

Adapting and improving performance is part of the repertoire of natural agents for dealing with changing and unpredictable environments, expanding repertoires of behaviour based on experience, and for the development of more sophisticated behaviours such as language, which are maintained dynamically by a community.

Artificial systems have the dramatic advantage in that they can share code with other instances of themselves much more readily than natural systems. They thereby short-circuit many forms of learning, even observational learning. So, the evolution of artificial systems will be a web rather than a tree, in which advances spread between individuals, generations and across domains of application. This allows machine culture to accumulate experience at a much higher rate than the culture of humans or natural evolving organisms.

1.3.7 *Evolve*

We have already discussed the evolution of cognitive systems from a societal perspective – basic patterns of technology get established and then develop, expand and refine, sometimes over many generations. The evolution of artificial systems benefits from two key properties that are not available to natural agents. First, there is a network effect: advances in one field can apply immediately to all the others, in the metaphor of evolution, genetic advances jump across the species. And the wider the field of endeavour, the more cross-contributions occur. Secondly, completely new lines of technology can be started at any time, without any particular linear dependence on previous ones – new species can emerge discontinuously from previous ones.

Already genetic algorithms are used in practical situations to generate better algorithms and designs for engineering problems. We can expect evolutionary approaches to become an established part of progress from individual applications to whole new product areas.

1.4 Motivations

The field of cognitive systems can be pursued from the different perspectives of the life sciences, physical sciences and real-world engineering. At any one time, people in these different fields are studying substantially the 'same' problem from any or all of the following perspectives:

- *The 'pure' engineering approach* Building new, challenging, artificial systems with all means that come to hand, because we need them and the only important measure of success is how the system performs. Sometimes this will be done using *inspiration* from biological systems but without a specific commitment to understanding how the biological system does it.
- *The 'pure' life-science approach* Studies to understand biological systems, specifically how they do what they do, using any explanatory tools available, with the agenda driven purely by the natural evolution of the scientific agenda (available tools, current state of knowledge and so on).
- *Natural analysis and simulation* Building systems as a way of exploring and understanding through simulation how biological systems do what they do – using successful artificial simulation as a high standard of proof of adequacy of an explanatory model.
- *Engineering analysis and simulation* Falling somewhere between the first three perspectives, is the process of building simulations of natural processes and agents (fire, water, plants, animals, humans) because we need the results of the simulations, rather than through a concern with the authenticity of the underlying model. This may be for fun and games such as special effects and crowd scenes in a film, or 'serious' purposes such as modelling crowd behaviour in a burning building.
- *Architecture* Building low-level system architectures (hardware) that get closer to the fundamental architecture of neural systems. Studying capabilities of biological systems, with the agenda chosen to illuminate specifically how biological systems achieve computational tasks with architectures so radically different from

current artificial architectures (massive numbers of slow, highly connected, failure-prone, un-clocked … components).

- *Prosthetics* Building prostheses that achieve capabilities closely matched to, and interfaced to, humans.
- *Theories* Trying to build theories of any of the preceding sorts of artificial system, because we want to realize them in predictable ways, and have guarantees of performance etc.
- *Tools* Better tools for probing, measuring, intervening, into biological systems for the purposes of understanding them.

For example, consider the problems and the different sorts of projects that might come from solving them:

- *Identify an individual object from a cluttered scene, pick it up without damaging it, and perform some simple action on it:*
  - industrial robot on a next generation flexible assembly line
  - manipulator for the next Mars probe
  - artificial hand for prosthesis
  - artificial hand for next generation humanoid robot that can pick up and fold clothes.
- Observe natural scenes, pick out and identify important things:
  - identify pedestrians for car safety systems
  - smart binoculars for birdwatchers
  - shopping mall child tracker
  - 'Who is that?' memory jogger in your camera-phone.
- Get around complex environments without falling over or getting trapped:
  - autonomous robots for hazardous environments, military applications, toys, home robots.

In addition to the variety of motivations, there are six different application foci that correspond to different types of relationship between people and things, and therefore different types of interest and outcome for cognitive systems:

- person to thing: 'Computer – what's the weather forecast?'
- amongst things: collectives and networks of intelligent agents
- person-and-thing viewed together: cars, exoskeletons, cognitive implants
- lots of person-and-thing: traffic, combat, games
- things alone: robot explorers
- person-to-person mediation: communicative and collaborative environments.

1.5 Summary

From this section we can see the breadth of the field of cognitive systems, and appreciate why there is no single research community addressing it or articulating an overall agenda.

In the following sections I consider *all* these motivations and application foci within the scope of cognitive systems. A comprehensive overview is impossible, but I have attempted to bring together representative cases of many of these different possibilities, particularly looking for those where there is the likelihood of a strong shared agenda across the life sciences and physical sciences.

2 APPLICATIONS AND SOCIETAL IMPACT

The criteria for picking cognitive systems for discussion are:

- There is no dispute that systems of this class will emerge into the mainstream over this period.
- There will be widespread public policy implications.
- The impact will provoke public interest and debate.

The treatment of each area is necessarily superficial and, as I noted above, this is an individual perspective. While I have had valuable inputs from a number of specialists, there has been no opportunity to consult experts in every area, so nothing here should be regarded as in any sense a consensual view of the people involved in each field. My goal has been to generate a sense of the issues that are worthy of further discussion and debate.

The report makes frequent use of extended quotations, especially from on-line sources, because evidence of specific activities is the best indicator of ambitions and intent.

2.1 Business

2.1.1 The Ambient Web

Even after the dot.com bust, the most prosaic descriptions of the progress of computing in reshaping business recognize profound effects from e-commerce and the many ways that computing and communications restructure the underlying operations of business. Through the 1980s and 1990s the personal computer was at the heart of that revolution as the tool through which computing reached into everyday use. The PC provided a *de facto* standard around which every branch of business could make massive investments in IT.

The emergence of the World Wide Web (WWW) was a defining transition in computing. It shifted the model of applications away from individually crafted islands of computing to a worldwide utility that any computer anywhere can access by common standards. This is accelerating the shift to a new era when the PC, and the notion of personal computing it embodies, loses that central role in the evolution of our global infrastructure of information and communications technologies (ICT). The new era, variously called pervasive computing, ubiquitous computing or ambient intelligence, is built around core developments in three areas:

- The core computing and communications components become cheap enough for 'anything' to connect to the Web. Today and in the near future that means smart phones, TVs, cameras, hi-fi, car navigation. In the next two decades it is *any* artefact for which there is a reason to connect – if it makes sense for my coffee cup to be online, it will be.
- The WWW, with which we are all familiar, was built primarily for people to access information. The technical community is now busy building open standards under the heading of web services and the Semantic Web that will allow the exploding population of smart devices, information services and applications to interact directly with each other over the WWW in dynamic and flexible ways.
- The shift towards massive numbers of things interacting together in ways that nobody can completely design or control in advance has prompted a new approach to system design, known as agent based computing (http://www.agentlink.org). Agents are autonomous software entities that can act and react with each other in a distributed environment.

Taken together, these developments, which we can dub 'the Ambient Web', will create a profoundly different IT landscape from the one with which we are familiar. It will provide the setting for all the other areas of societal impact. It is not realistic to attempt here to anticipate the 20-year future of ICT, but keeping within the scope of the Foresight Cognitive Systems Project, we can highlight some dimensions of this pervasive change:

- We will be using a language of autonomy and intelligence with respect to wide classes of everyday things which operate under human supervision, with major ramifications for definitions of such things as product liability, security, service definition and so on. Pinning down responsibility will be particularly difficult as performance is the outcome of interactions between multiple agents, combining embedded and online capabilities.
- Intelligent behaviour of individual things will be achieved through an interaction between the thing itself and the capabilities of the Ambient Web 'behind the wall'. A simple rule of thumb is that if today a thing relies on power to operate, in the future it will rely on the intelligence of the Ambient Web. This will create

new levels of dependence of everyday life on pervasive infrastructure – thinking about this while trying to get to London by rail is a sobering experience. Questions of ownership and governance of the Ambient Web will loom very large.

- There will be massive economic activity in the trading of 'smarts' – pieces of software that improve the performance of things. The WWW has already fuelled completely new areas of software swapping, such as music, trading of virtual characters for popular games, software for intelligent toys. The precedents are that users will create these possibilities well ahead of the legal structures for them.
- It will be impossible to prevent this world of ambient intelligence from suffering all sorts of undesirable attacks such as viruses, spam, cyberterrorism etc. Ironically, the technical work needed to create robust, reliable and defensible networks may be a major driver of understanding of certain classes of cognitive systems (see, for example, IBM initiative on autonomic computing). The possibilities of new vulnerabilities may prove to be one of the biggest brakes on deployment of new capabilities.

As digital technology and software become an intrinsic part of previously disconnected products and services, the structures of governance that we have in place will almost certainly prove inadequate.

2.1.2 Commercialization Concerns

A major area of importance for the future impact of cognitive systems will be the incursion of commercial concerns, and the power they can wield, into new areas of our lives. To understand this better we can look at the recent history of computing.

The massive commercial and societal phenomenon of personal computing over the 1980s and 1990s was so dominated by two companies – Microsoft and Intel – that it came to be known in the trade as the 'Wintel' model. It is the nature of computing technology that it is built on layers of abstraction that allow investments at one level to be made independent of other layers.

Standards are essential for these layers. These may come from standards bodies, but in fast-moving areas they are as likely to come from commercial players who can use the momentum of change to maintain a pre-eminent position.

The prize for achieving architectural leadership over a layer is huge, with extensive consequences for other vendors and users, so extensive that protracted legal struggles emerge as positions verge towards monopoly. IBM was the target of such action in an earlier computing generation, and Microsoft in this.

Another example comes from the very different area of genetically modified foods, where the arguments over desirability of widespread adoption are inextricably bound up with concerns about the degree of control over the food chain that may accrue to commercial enterprises. For many people these concerns are more important than questions of safety, because of the way they threaten many aspects of choice and diversity of food supply, and have the potential to change fundamentally the structure of the industry.

Looking at each area of cognitive systems, there is no particular reason to suppose that any one of them will lead to such high levels of commercial dominance in any area of application, as we have seen in previous eras of computing, or concerns like those around GM food – but no reason either to suppose that they will not. Where there is a commercial logic then there will be attempts to achieve overwhelming market dominance. As each of the following discussions of impact shows, we are bringing computing power into a very close relationship with intimate aspects of human nature, and so the sensitivity to commercial influence over these aspects of life will become a major concern.

Developments in cognitive systems are likely to provoke very strong reactions precisely because we use language that borrows so heavily from words that we associate with human and biological capabilities, and which are therefore redolent with everyday meanings that overlap with issues of central concern to people's lives. It will take only a few sensational news items of the 'roborat' variety (see Assisted cognition in section 2.3.2) to create defining shifts in the terms of debate. It is therefore likely that a societal backlash to whole categories of cognitive systems could prevent the deployment of technologies that might otherwise seem broadly beneficial. The growing debate over nanotechnology provides a foretaste of the sort of reaction that we can expect, and illustrates how important and difficult it may to be to carry the debate forward in well-balanced ways.

2.2 Two Perspectives on the Ambient Web

We can illustrate the impact of the Ambient Web by considering two areas where the broad momentum of ICT in the business environment will provide a horizontal context for all the other areas of impact, in the way that each generation of computing has done up to now. We can look at this through the two complementary perspectives of the network and the personal interface through which individuals interact with it.

2.2.1 Multi-Agent Network Systems

We can already see the impact of systems based on autonomous agents in two major areas: the underlying infrastructure that will create the Ambient Web; and applications that involve huge numbers of measurements, decisions or trades, such as scheduling, process management and on-line trading.

As telecoms research works to build ever bigger, more capable and more reliable networks, it has found a rich vein of inspiration in the study of 'swarm intelligence' – the way that a colony of simple creatures such as ants can exhibit high-level adaptive behaviour. There is a natural and appealing congruence between the millions of network elements that operate under decentralized control rules and the nature of insect colonies. For instance, studying how a colony of ants can quickly discover the shortest route to food, can lead to new ideas for how mobile agents in a network can adaptively create 'shortest path' routing for network traffic.

The success of the Internet as a world-wide network is due in large part to the highly decentralized nature of the underlying networking protocols. These limit the effect of failure in any one part, and allow the network to adapt extremely well to the loads placed upon it. As the emerging Ambient Web drives the reach of the network down to billions of individual devices we will need much deeper understanding of how to achieve desirable emergent properties.

At the application level, on-line trading is set to drive forward the technology of multi-agent systems. Business-to-business (B2B) online auctions are already a major feature of e-commerce, with many billions of pounds of goods already traded.

Agent approaches, built over the standards of the Semantic Web, will extend program trading significantly to these on-line marketplaces. This will drive the emergence of ever-more fine-grained and liquid trading in all sorts of areas. Governments will need to develop extensive standardization and regulatory capabilities well beyond those they currently have to ensure that these markets develop in an orderly way.

In realizing these new, multi-agent, open systems, system designers will need to import much of the social and political language of human organization design, for example to govern how rights are delegated to a transferable preference voting system between agents representing multiple parties. Inter-regional issues of legislation will present a particular problem as the net links many human and automated agents distributed around the globe.

As we have already noted, we are travelling towards a future of systems that will manifest emergent properties as many agents interact. It is important to realize that there is a serious lack of theoretical tools to underpin our understanding of these emergent effects. We do not even have a full understanding of the simple nine-cell John Conway Game of Life program, an example of a cellular automaton much studied by mathematicians. We can therefore expect that the growing pains of the Ambient Web will be associated with unwelcome surprises as we encounter unforeseen emergent effects.

The early difficulties with program trading in stock markets when automated trading programs created a cascade of transactions are an example of the problems ahead and the type of impact they may have on everyday life. As in that case, the solutions will come not just from the technology, but from adapting the social and legislative framework that governs their use. (See Chapter 2 on Large-Scale, Small-Scale Systems and http://www.agentlink.org).

2.2.2 From PDA to PDE: The Personal Digital Environment

The move to the post-PC world of the Ambient Web is bringing an explosion in the number of connected devices available to an individual. There is a thriving research and commercial community that is exploring this new world of wearable and ultra-portable computing. The defining market transition is the current move towards 2.5 and 3G phones, giving individuals constant access to the WWW.

The cost of computing means that there is now a forced bundling of all sorts of capabilities into smart phones and WDAs (Wireless Digital Assistants). This is a passing phase that will gave way to what we might call the Personal Digital Environment. This will entail fully connected technology at the scale of credit cards, jewellery, glasses, coins, wristwatch and so on, that can be embedded into whatever form is most convenient in our personal environment. Individuals will take many different paths through the space of possibility this opens up, from eager adoption to total rejection of the 'wired world'. The significance of this for cognitive systems lies in a number of possibilities:

- People will be able to participate continuously in on-line worlds if they choose. The fascination with personal entertainment and communication devices suggests that they will significantly change how people go about their lives.
- Many people will adopt a range of online persona or 'avatars' to provide interfaces and buffers to the otherwise overwhelming intrusion of the (on-line) world. We may see surprising crossovers of media with people using 'artificial' communication in face-to-face situations.
- We are coming to terms with the huge impact of the e-mail 'trail' left behind from our daily interactions, with the loss of privacy and control that implies. This is set to change by orders of magnitude, with many consequent problems: 'Your brain trace proves you weren't paying attention when you crossed the road and caused that accident.'
- Assistive technologies can be fully on line: continuous telemonitoring of physical and cognitive sensors and feedback mechanisms will be available. (See more in the discussion of assisted cognition in section 2.3.2.)

The evidence of technology adoption so far suggests that we will move rapidly into this world, well ahead of our understanding of the issues it raises.

2.3 Embodied Cognition: Robots and Smart Things

The natural cognitive systems that we can study inhabit a living body of some sort. So there is a deep and productive relationship between the study of natural cognition and the construction of robots and intelligent machines that must exhibit functional ability in natural settings. While this is a statement of the obvious, it is important to recognize that embodied cognition has not been the driving force for the evolution of computing over the past few decades.

Business, scientific and personal computing all evolved around a central model of symbolic computing that is disconnected from the demands of the natural environment.

It has turned out that embodied cognitive tasks – like walking on two legs, sorting out objects in a cluttered scene, or building up a map by moving around – are astonishingly difficult. However, decades of work on industrial robotics, and the emergence of a wide range of artificial sensory-motor technologies around the core of embedded computing, mean that embodied artificial cognition is set to become a mainstream branch of computing's evolutionary tree. These systems will cover a spectrum from robots designed for a wide range of autonomous functions to 'smarts' embedded in everyday things. Interestingly, the concept of 'embodied cognition' is also beginning to have an impact on neuropsychology.

We are still at the early stages of understanding how embodied cognition works, as we are for cognitive systems in general. So, while they may be inspired by solutions adopted by natural organisms, our artificial ones will be developed within existing engineering models. As a consequence, we do not know where the real boundaries of performance lie.

Just as IBM's Deep Blue did not need to play chess the same way as Garry Kasparov to beat him, so artificial embodied cognition solves problems in its own way. We do not know which problems are deceptively easy, and which are deceptively difficult.

The US Department of Energy provides a good benchmark, and an associated technology roadmap, for assessing the impact of robotics. In its vision for 2020 it states:

> Over the next few decades, advanced RIM [Robotics and Intelligent Machines] technologies will fundamentally change the manner in which people use machines, and by extension, the way DOE accomplishes its missions. New robotic systems, fuelled by improvements in computing, communication and micro-engineered technologies, will transform many of our most difficult tasks. It is expected, for example, that:
>
> micro-scale robots with the ability to crawl, fly and swim will be able to work together to perform monitoring, surveillance and intelligence operations;
>
> environmental facility remediation, monitoring and inspection, as well as resource exploration, will be performed with high efficiency and low risk through autonomous teams of robots; and
>
> automated methods closely coupling design and manufacturing will allow cost-effective, totally automated production of both large- and small-lot manufacturing products.
>
> … By the year 2020, RIM will both duplicate and extend human dexterity, perception, and work efficiencies in a broad range of tasks – these technologies will be as pervasive and indispensable in DOE operations and the National economy as the personal computer is today. (Robotics and Intelligent Machines in the US Department of Energy. A Critical Technology Roadmap. Sandia Report SAND98-2401. October 1998).

As an example of current commercial activity, Honda is making very public commitments to robotic technology with its high profile ASIMO programme for humanoid robots. The company see these as a major focus for technology development over the coming decades (http://world.honda.com/ASIMO/technology/).

Honda's main business, of course, is engines and vehicles. This illustrates the point that these robotic and smart technologies need to be viewed as new set of capabilities to be deployed across all industries. Just as personal computing has generated components for deployment across the many fields of embedded applications, so the notion of things that sense and act with a degree of autonomous decision-making will become more and more commonplace across every area of application. For our purposes here, we consider robotics as a 'horizontal' technology to be included in the analysis of each of the following domains of impact.

2.4 Health, Well-being and Performance

2.4.1 *Drivers of Change*

There is nothing new with using technology to repair defects, overcome disabilities, or enhance our performance – whether it is a pair of glasses, laser correction of sight, a heart pacemaker, a pocket calculator, or a PDA with automatic reminders of things to do. What is changing is the depth of knowledge and technology that we can bring to bear on these issues, and in consequence the intimacy of relationship possible between ourselves and our personal technologies.

There are five core areas of research and development that, taken together, will radically change our personal relationship to technology in the area of cognitive systems.

Macro Mapping of Brain Function: Architecture of Brain and Mind

The use of a wide range of imaging techniques to map brain activity while people undertake controlled tasks is leading to a rapid growth in knowledge of which parts and networks of the brain do what, and how they combine in overall performance. Targeted approaches to intervention will follow.

External Brain Interfaces

The same technologies used to monitor and model function externally can:

- direct brain communication with other systems
- give the individual direct feedback on their performance (training brain states for optimal performance)
- or connect back to real-time intervention through implants.

Micro Mapping of Brain Function: Cognitive Components

As well as macro understanding of the function of areas of brain, research will progressively unravel specific micro aspects of performance, understanding what happens at the level of neurons, how it processes inputs and outputs, allowing the development of specific enhancement at the component level and replacement technologies for particular functions. Early stage sensory processing, such as the cochlea and retina, are examples today.

Internal Neuronal Interfaces

Technologies are developing rapidly for directly interfacing living neurons and silicon-based electronics, making ever more sophisticated implants feasible. For the present, however, this kind of technology is more at the level of blue-skies research than marketable products.

Personal Sensing

It is already commonplace in the gym to wear a heart monitor that interacts with the equipment to produce the best workout, or for individuals to wear sensors that collect data for the management of a medical condition. All sorts of technologies for health and well-being will rapidly colonise this space of opportunity, building on the general PDE (Personal Digital Environment).

2.4.2 *Applications*

There is no one term that the physical and life sciences communities use to cover the range of approaches to repairing, supporting and enhancing performance through interfacing with the human nervous system that derive from these core developments. However, the variety, scale and interactions between all these efforts will have dramatic effects over the next 20 years. There are some common themes that we illustrate in the following discussion:

- Implanted chips will become available for new and experimental treatments of a wide range of cognitive dysfunctions. Researchers will be keen to push the boundaries, challenging existing regulatory regimes.

- The distinctions between treatment, enhancement, recreation and so on will become ever more blurred. Technologies developed for the disabled may become more widely used – such as software for predictive typing by people with motor problems extending into general use. Enterprises are springing up to offer ever more enhancements as soon as there is any indication, however dubious, that we can 'switch on' better memory, faster thinking and so on.
- Many forms of technology and cognitive intervention will thrive in the world of alternative medicine. There will be major challenges over issues of regulation of practice.
- Many more people will depend on cognitive assist technologies in everyday situations. What limitations will there be? What will be the consequences for insurance?

The rest of this section brings out these and other concerns from areas where there is current research and application.

Neuroprosthetics

Neuroprosthetics is the use of direct electrical stimulation of the nervous system for functional performance. It is already an area of high clinical value. We can expect to see research investment and major progress. Cardiac pacemakers have been around the 1930s. Since then there has been progress in a wide variety of areas:

- cochlear implants and early versions of retinal prosthesis
- autonomic functions such as control of bladder and bowel
- movement, posture, spasticity.

As an example of the research frontier for cognitive systems, there is the recent widely reported work on a hippocampus brain chip of Theodore Berger at the University of Southern California in Los Angeles. The hippocampus plays a key role in laying down memories, being able to replace it when damaged could be a fundamental contribution. The researchers, whose work is based on modelling a rat hippocampus, report (http://www.usc.edu):

> 'Our current chip has 18 dynamic neuron synapses, and it behaves just like a network of real biological neurons in the hippocampus,' Granacki said. 'When the chip receives real electrical signals as inputs, it processes them and sends out exactly the same signals that a real neuron would send.'
>
> Berger's ultimate goal is to make a computer chip that can be connected to human brain tissue and take over a cognitive function that has been destroyed by epilepsy, Alzheimer's disease or some other brain problem.
>
> 'We'll need at least 10 000 neuron models to do anything useful in the human brain, and these will have to be on a chip small enough to be surgically and strategically placed in a particular location of the brain,' said Berger.

Given the limitations in our understanding of the fundamentals of brain computation, there will be major issues when deciding what level of equivalence the technology has to achieve before researchers will be allowed to test implants such as these in humans.

Neurofeedback

Neurofeedback is a learning strategy or procedure in which a person watches their own brain activity (via monitoring of EEG signals) and learns ways to alter it for some purpose such as improving performance or stabilizing their mood. Think of it as exercises for the brain. It is already widely used for conditions such as attention deficit hyperactivity disorder (ADHD), depression, epilepsy, alcoholism, sleep disorders and many more (see for example http://www.eegspectrum.com).

Recent work at Imperial College, London, illustrates the sort of applications that will bring this into everyday mainstream interest (http://www.ic.ac.uk/p4330.htm):

> Researchers from Imperial College London and Charing Cross Hospital have

> discovered a way to help musicians improve their musical performances by an average of up to 17 per cent, equivalent to an improvement of one grade of class of honours …
>
> Professor John Gruzelier, from Imperial College London at Charing Cross Hospital, and senior author of the study, adds: 'These results show that neurofeedback can have a marked effect on musical performance. The alpha/theta training protocol has found promising applications as a complementary therapeutic tool in post-traumatic stress disorder and alcoholism. While it has a role in stress reduction by reducing the level of stage fright, the magnitude and range of beneficial effects on artistic aspects of performance have wider implications than alleviating stress.'

Up to now, the technologies and regimes for monitoring and feedback have resided primarily in the research and therapeutic communities. However, under the influence of the core technology trends the scene is set for these to become available as mass consumer technologies.

Direct Brain Interfaces

There is a high level of interest in helping paralysed people to use conscious control of their brain states to drive assistive technologies, such as entering commands to a computer or controlling a wheelchair. This is proving very hard to do, but we might reasonably expect that, as we gain a much more precise understanding of the brain's architecture, we will be able to home in on correlates of particular brain functions that are under conscious control.

Closing the Loop

As neuroprosthetics, neurofeedback and direct brain control develop, there will be increasing opportunities to 'close the loop'. A new generation of cognitive prosthetics will then emerge in which individuals engage with sophisticated real-time feedback to influence their own body and brain performance through assistive technology. See for example, http://oea.larc.nasa.gov/news_rels/1999/Sep99/99-065.html:

> Some of the more than 15 million Americans who have diabetes may soon use NASA virtual reality technology as a new treatment in the self-management of the disease.
>
> Preliminary observations show that NASA's artificial-vision technology can help patients at risk for nerve damage associated with diabetes to visualize and control blood flow to their arms and legs. This application, which comes from several years of research aimed at enhancing aviation safety, combines two technologies: sensors to measure the body's reactions and powerful computer graphics to turn those measurements into a 3D virtual environment.
>
> The graphics technologies are used in research with cockpit artificial-vision systems to help pilots see in low- or no-visibility situations, and as data-visualization tools to help designers study air-flow patterns around new aircraft shapes. In this fall's studies, diabetes patients will wear a 3D virtual-reality headset to visualize the contraction and expansion of their own blood vessels.
>
> Using self-management, or biofeedback methods – including changes in breathing and muscle flexing – the patients will increase blood flow, which will be measured through sensors attached to their fingertips. The system uses skin-surface pulse and temperature measurements to create a computer-generated image of what is actually happening to blood vessels under the skin. Just as pilots use artificial vision to 'see' into bad weather, patients will use this virtual reality device to see beneath their skin.

Assisted Cognition

In the previous sections we considered cognitive systems that involve the engagement of technology in new ways with our nervous systems. However, just as much might come from the use of computing in what we might regard as the culturally more conventional form of 'things that make us smart' (Norman, 1993). This field does not have a single name or focus, but is variously called cognitive prosthetics, interactive cognition, assisted cognition or distributed cognition. The common research foundation is a concern to use evidence-based approaches to understand how humans use the things around them as part of their cognitive processes.

This is a very important difference in perspective from the sort of laboratory studies of brain performance that dominate the agenda for brain science. It studies directly how the things in our environment support our ability to perform specific functions. The big expansion of possibilities that is on the horizon comes from the trends covered in the section of this review on the Ambient Web and personal digital environment. The effect of these for assisted cognition arise from capabilities such as:

- Assistive systems will be able to have very fine grain, real time information about an individual – where they are, how they are moving, their physical and brain states etc.
- The environment can be 'smart' in all sorts of ways to offer the right information or intervention at the right moment, in the right place.
- The assistive system can learn patterns of behaviour and use these to support routines, spot potential problems and raise alarms under unexpected conditions.
- An individual can be connected to the whole on-line world, to offer both augmented processing to their mobile prosthetics, and to other people who can be involved in a wide variety of assistive roles.

At the moment, the main investment in pervasive computing is driven more by commercial interests in such areas as entertainment and on-line commerce, but it would be of great benefit to society to direct these same resources to the rapidly growing care requirements of the ageing population. A number of researchers are beginning to look specifically at how assisted cognition might help to tackle the problems of Alzheimer's patients, see for example http://www.cs.washington.edu/assistcog/.

Recently, researchers at Intel in the USA have taken the lead in establishing a broad initiative that includes such work (http://www.agingtech.org/). In the keynote address at the conference that launched the initiative, the following example is given of the sort of system that is being researched:

> we built a prototype system in our lab to prompt and assist someone to fix a cup of tea and to monitor her or his progress of that activity over time. Using 'mote' technology —-a small plug-and-play processor and wireless transmitter from our Intel Research Berkeley lab – we have plugged in five kinds of sensors: (1) motion sensors for activity detection; (2) pressure sensors in chairs to know whether someone is sitting; (3) switches to know when drawers, cabinets, or objects in the kitchen have been moved; (4) RFID antennas situated between the family room and the kitchen to identify small tags placed in peoples' shoes; and (5) an IR-tracking camera that detects whether a badge-wearing 'patient' has fallen. All the real-time data travels through the motes' wireless network back into a host PC for processing, prioritization, and communication. (http://www.broadbandhomecentral.com/report/backissues/Report0307_3.html).

It is reasonable to assume that there will be a very rapid growth in this segment, and that there will be a particularly important opportunity in supporting independence via sophisticated 'tele-care'. Just as people carry emergency alarms, we can expect them to sign up for systems that allow them to be monitored and assisted in all sorts of ways. Many issues of privacy and control will naturally arise.

Assistive Robotics

Robotics is set to become a key driver of prosthetic technology because there is intrinsic scientific, technological and commercial interest in giving autonomous machines the sensory, movement and control capabilities of humans and animals. As these technologies progress to within the size and weight range that matches humans then they become available for assistive functions. See, for example, the MIT Leg Lab and its spin out commercial ventures:

http://www.ai.mit.edu/projects/leglab/home.html.

However, the very high costs of these technologies means that for the immediate future, the drive to development is from the military potential of exoskeletons. DARPA has promoted research in this area for several years (see section 2.8 on military capabilities).

It is reasonable to suppose that over the next 20 years these technologies will come within range of mainstream use with consequent issues:

- How will national health systems respond as costs spiral for new mechanical assistive technologies for the disabled? What will happen when, instead of crutches, those with injuries need an expensive exoskeleton? Will it be a two-tier world where those who can afford the equivalent of a new car enjoy functional mobility while others are literally left behind?
- Amplification technology, deployed first for the military to increase survivability and power, will be quickly adopted for industrial applications, sports and crime. How will we control their use?
- Neural interfaces will bring these technologies into increasingly natural relationship with their owners.
- Will some people choose permanent augmentation as a lifestyle choice?
- The cost and availability of these technologies is bound to emerge as a major concern for public policy. We will need new cost/value assessments as it becomes possible to make radical shifts in care patterns. Will the rich provide the leading edge of adoption and the market take care of the rest? The UK lacks the diversity of funding sources of the US which speeds the diffusion of such new technologies. Should the UK try to match the USA?

2.5 Transport

Humans did not evolve to drive vehicles safely, at high speed, in conditions of high congestion and poor visibility on motorways or in busy city streets. So it is not surprising that a lot of accidents happen. Nearly all the intelligence involved in controlling a vehicle is currently within the human behind the wheel. This will change rapidly. Today we are accustomed to the idea that an anti-lock braking system (ABS) can help us to operate our car more safely in a crisis. In future we are likely to hand over much more control. Car manufacturers are already researching advanced forms of vehicle control and driver assistance that will radically change how we drive:

> … Video cameras record the surroundings from the perspective of the driver, while clever and extremely fast image-processing software makes it possible to identify vehicles, pedestrians, road signs, traffic lights and road markings reliably and also to monitor moving objects. All of this is necessary if drivers are to be warned of dangers before they actually recognize them.
>
> The Distronic automatic distance cruise control system is already available in Mercedes-Benz models. It relieves the driver of routine tasks in dense lines of traffic by automatically controlling the distance to the vehicle ahead – within the limits of the system's capabilities. The functionality of this assistance system can be extended in a variety of ways. For example, researchers at DaimlerChrysler are working on improving Distronic by means of data exchange between vehicles.
>
> This is how it will work. If a car traveling far ahead on a highway has to slow

> down, it will forward the data on the magnitude of its braking deceleration to the vehicles behind it. The cruise control system will use this signal to reduce the speed of its vehicle and increase the distance to the car directly ahead as a precaution – even if the latter isn't braking yet. The assistance system will thus permit 'foresighted driving' and ensure that a dangerous situation never arises in the first place. (http://www.daimlerchrysler.com).

Even experienced drivers have little opportunity to learn how to control a vehicle under demanding crash conditions. So they do not get the benefit of the learning that is available to artificial systems. As experience with artificial control systems accumulates, they will have the advantage over humans. They can learn and tune control algorithms over millions of hours of simulated and real driving. It can surely only be a matter of time before we start to prefer automated control to the human variety.

Consider the sort of issue this will raise in the following scenario. Cars can be conceived as (semi)-autonomous intelligent agents that need to create emergent safe behaviour between them. Imagine a motorway in 20 years time, when an accident is emerging. All the approaching cars are immediately aware of it and start to communicate with each other to bring themselves collectively to a halt in safety. They have a much better chance of achieving this than the drivers could unaided. However, in optimizing the overall outcome the collective behaviour of the system may produce a worse outcome for one of the drivers than that individual might have experienced unaided. This is a familiar moral dilemma in crisis management. Does one suffer avoidably for the sake of safety for the many?

Even without having to tackle such dilemmas, we will face many issues of investment and risk management. Once it becomes obvious that more capable control systems for identifying hazards and controlling vehicles can achieve major reductions in accidents, there will be a major issue of the rate of deployment of such technologies, just as there is on the railways today.

Testing, responsibility and accountability will all have to be radically re-thought. What will it mean to take your driving test on a vehicle that is more competent than you are under many circumstances? Will cars limit your ability to control them to the expertise level you have built up through actual driving? How will this personal level of expertise be encoded and communicated to the vehicle? How will insurers respond? Who is responsible for the assuring the collective behaviour of multiple vehicles?

Looking further into the future, are we heading towards the time when human driving will become a form of extreme sport to be allowed only within controlled areas? At the very least, we will surely insist on drastically reducing the scope of human control under many conditions.

2.6 Sociable Technologies: Arts, Entertainment and Companions

> I have come to the conclusion that if we want computers to be genuinely intelligent, to adapt to us, and to interact naturally with us, then they will need the ability to recognize and express emotions, to have emotions, to have what has come to be called 'emotional intelligence'. (Picard, 1997).
>
> Replacing human contact [with a machine] is an awful idea. But some people have no contact [with caregivers] at all. If the choice is going to a nursing home or staying at home with a robot, we think people will choose the robot. (Sebastian Thrun, Assistant Professor of Computer Science, Carnegie Mellon University).
>
> AIBO [Sony's household entertainment robot] is better than a real dog. It won't do dangerous things, and it won't betray you. … Also, it won't die suddenly and make you feel very sad. (A 32-year-old woman on the experience of playing with AIBO).
>
> Consider the response to the question 'Is the Furby alive?'. Jen (age 9): 'I really

> like to take care of it. So, I guess it is alive, but it doesn't need to really eat, so it is as alive as you can be if you don't eat. A Furby is like an owl. But it is more alive than owl because it knows more and you can talk to it. But it needs batteries so it is not an animal. It's not like an animal kind of alive.
>
> … my daughter, upon seeing a jellyfish in the Mediterranean, said, 'Look Mommy, a jellyfish, it looks so realistic!' Likewise, visitors to Disney's Animal Kingdom in Orlando have complained that the biological animals that populated the theme park were not 'realistic' compared to the animatronic creatures across the way at Disneyworld. (Turkle, 2002).
>
> Matsushita Electric announced its entry into the 'pet' robot market … with Tama, a robotic cat designed to be a conversation partner for elderly people.
>
> Unlike other robotic pets, like Tiger Electronic's Furby or Sony's Entertainment Robot, the catlike Tama will have more than just entertainment value, offering companionship and a variety of other services to the aged, said Matsushita.
>
> 'The idea [behind Tama] is animal therapy,' said Kuniichi Ozawa, General Manager of Matsushita Electric's Health and Medical Business Promotion Office. 'A network system will enable the pets to speak to the elderly in a natural way, especially to people who are living alone, and this will make them more comfortable.'
>
> Tama can be connected via cell phone or ISDN line to a network system center, allowing health or social workers to send local news, medical information, and encouraging messages to elderly people. (Michael Drexler, IDG News Service Wednesday, 24 March 1999).

People have a natural tendency to attribute human 'personality' to things. The explicit design of such characteristics is now a mainstream activity of games, interactive media and in the broader field known as affective computing. Toys and entertainment provide a strong driver because questions of scientific authenticity of machine emotions are put on one side in preference for creating a particular experience. How it 'works' is less important than whether it achieves the 'right' effect.

This understanding of affective computing is becoming particularly powerful because it is happening alongside rapid progress in the related areas of:

- modelling virtual reality
- physically immersive interfaces in virtual reality
- multi-agent simulation
- humanoid and animatronic robotics.

Together these bring technology into a rich variety of everyday situations and will vastly expand the ways in which we might find ourselves spending time relating to technology rather than to people. On the positive side, there is certainly a case to be made that in many situations the patience and responsiveness of artificial systems used alongside human relations will be a big step forward in managing the complexity and stress of life. However, looking for issues, there are plenty of areas that will be controversial:

- Debates about the influence of media over behaviour will surely become much more intense when the media have properties of profound emotional engagement which further blur the boundary between real and simulated reality.
- As commercial concerns play an ever-bigger role in delivering the experiences that make up everyday life for many people, there will be many issues of how 'normal' human responses are being manipulated for commercial gain.
- There have already been reports of schools having to deal with children who have spent so much time watching television rather than having real interactions that they lack important social and conversational skills for learning. Will the next generation of relational media make this situation better or worse?

- How might the trade in affective computing develop? People already hack their Furby and Sony's AIBO toys, trade with one another for expert levels of game agents etc. Presumably this will grow explosively, and there will be all sorts of unusual virtual agents out there that you might not want your children to play with.

2.7 Education

Even more than other domains treated here, education requires the caveat that it has not had the input and treatment it deserves. The notes here are intended to provoke such a discussion.

There are two broad classes of impact to consider. The first will come from our understanding of the *process* of learning and the role it plays in developing the performance of natural and artificial cognitive systems: just why it is that certain things should be 'wired in' and others learned through exposure to the environment, how learning actually happens and how to improve it.

The second area will be in the way that many tasks come to be conceived as a symbiosis between a human and machine, with consequent changes in how we train and assess people. A few examples:

- Tutoring systems will become much more powerful assistants in the teaching and testing process as they are built around detailed models of learning, with abilities to diagnose specific barriers and to develop coaching routines around them. Perhaps we will have to replace many exams with the reverse Turing test. You are only considered to know a topic if you can convince the computer that you do.
- As discussed in the earlier section on neurofeedback, we will have much more sophisticated ways to understand our own performance and how to bring our minds into optimal performance for different types of task. When the difference in performance that can be achieved is measured in whole grades, there will be demands for pervasive adoption, with consequent major issues of safety, validation of techniques etc.

  As cognitive systems and prosthetics grow in capability and are available 'always on', we will have to change our notion of human performance. Just as driving performance does not exist outside the context of cars and roads, just as you need different licences to drive cars with manual and automatic gear changes, perhaps cognitive performance on many tasks will be inseparable from the cognitive prosthetics and systems we use to undertake them. For example, we are likely to get used to 'seeing' all the relevant information for the task at hand through augmented reality, and drawing on many forms of assistance. Part of our education will consist in building up our own systems: we will need radically different notions of assessment to cope with this symbiotic world.

There is, of course, a third type of impact, which is the need to educate people to understand these emerging technologies of cognitive systems. The challenge of undertaking the Foresight Cognitive Systems Project is a good indicator of how thinly the expertise is spread.

2.8 Military

It is not my intent to survey the impact of cognitive systems on military capabilities, but rather to highlight the strong relationship there is likely to be between military research and the field of cognitive systems. The military are interested in new, advanced capabilities for artificial systems, and in augmenting human capabilities in a range of demanding situations. In many instances this leads them to pursue the same or closely related agenda to that described in other sections of this chapter. The following topics illustrate this relationship:

Exoskeletons

> The Defense Advanced Research Projects Agency (DARPA) is soliciting innovative

> research proposals on Exoskeletons for Human Performance Augmentation (EHPA). The overall goal of this program is to develop devices and machines that will increase the speed, strength, and endurance of soldiers in combat environments. Projects will lead to self-powered, controllable, wearable exoskeletal devices and/or machines … To meet the challenges set forth, DARPA is soliciting devices and machines that accomplish one or more of the following: 1) assist pack-loaded locomotion, 2) prolong locomotive endurance, 3) increase locomotive speed, 4) augment human strength, and 5) leap extraordinary heights and/or distances. These machines should be anthropomorphic. (http://www.darpa.mil/baa/baa00-34.htm).

Autonomous Vehicles

> DARPA intends to conduct a challenge of autonomous ground vehicles between Los Angeles and Las Vegas in March of 2004. A cash award of $1 million will be granted to the team that fields the first vehicle to complete the designated route within a specified time limit. The purpose of the challenge is to leverage American ingenuity to accelerate the development of autonomous vehicle technologies that can be applied to military requirements. (http://www.darpa.mil/grandchallenge/)

Micro Air Vehicles

> The small speck in the sky approaches in virtual silence, unnoticed by the large gathering of soldiers below. In flight, its tiny size and considerable agility evade all but happenstance recognition. After hovering for a few short seconds, it perches on a fifth floor window sill, observing the flow of men and machines on the streets below. Several kilometers away, the platoon leader watches the action on his wrist monitor. He sees his target and sends the signal. The tiny craft swoops down on the vehicle, alighting momentarily on the roof. It senses the trace of a suspected chemical agent and deploys a small tagging device, attaching it to the vehicle. Just seconds later it is back in the sky, vanishing down a narrow alley. Mission accomplished …
>
> Sound like science fiction? This scenario may be closer than you think if success is achieved in the development of a new class of flight vehicles, the Micro Air Vehicles (MAVs), by the Defense Advanced Research Projects Agency (DARPA). The high level of current interest in developing a class of very small flight vehicles is the result of the nearly simultaneous emergence of their technological feasibility and an array of compelling new military needs, especially in urban environments. (http://www.darpa.mil/tto/MAV/mav_auvsi.html)

3 WIDER VIEW

Each area that has been considered here is in the early stages of development. In each there is a sense of gathering pace, as core areas of research and development bring into view whole classes of cognitive system that could not be realistically tackled a few years ago. As a consequence, it is hard to see the overall impact, but we should expect the combined effect to be quite dramatic.

A major resource for discussing the research frontier is the collection of papers in the Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology and Cognitive Science (NSF/DOC-sponsored Report, 2001: http://www.wtec.org/Converging Technologies/Report/nbic-complete-screen.pdf).

A key difference between this US initiative and the OST's Cognitive Systems Project is the former's consideration of cognitive science in combination with nanotechnology and biotechnology as well as information technology. This creates a field which makes much more ambitious predictions about technologies being embedded and pervasive throughout both the man-made world, but also embedded inside humans.

For all the seemingly futuristic ideas I have described, this survey has kept to the conservative side of prediction, firmly rooted in developments that are already visible. Anyone wanting to enter further into the debate should look at the more visionary end of the spectrum, particularly because the ambitious claims and visions are likely to capture public imagination and provoke reactions. As a single, but representative example, of such visions, see the set of essays by Ray Kurzweil, who is known for his ambitious predictions (http://www.kurzweilai.net):

> 'Experience beamers' will beam their entire flow of sensory experiences as well as the neurological correlates of their emotional reactions out on the Web just as people today beam their bedroom images from their web cams. A popular pastime will be to plug in to someone else's sensory-emotional beam and experience what it's like to be someone else, *à la* the plot concept of the movie *Being John Malkovich.*

As this quote and earlier examples in this chapter indicate, we must accept that the pursuit of research in cognitive systems will be surrounded by debate in which fact, fiction and ambition will be closely intertwined. The issues will touch on the very heart of what it means to be human. This report has attempted to show that, even staying within the conservative zone of prediction, cognitive systems will have a profound impact over the next two decades.

- The shift to the Ambient Web, personal digital environment and pervasive 'smarts' is firmly underway. It means that our lives will become as entwined with digital technology as they are already with electricity.
- Many of the effects will be slow, but cumulative. It is a commonplace of technology forecasting that predictions often overestimate short-term effects and underestimate long-term ones. No one set out to make cars change the way we live, or energy consumption change the weather.
- Cognitive systems are not a single identifiable class of 'thing', like cars, mobile phones or the Web. They are the future of systems and will be increasingly what we think pervasive digital technology 'is'.

At some point in our evolutionary past, humans entered the 'cognitive niche', fundamentally changing their relationship to the world around them and the rate at which they could develop their culture. The artificial world is about to do the same, with humans as partners in the process. The effects will be equally far reaching. There is no reason to be either utopian or dystopian about the outcome since human nature will evolve no faster than it has. But there is surely a duty on all those working in the field to help to inform and guide the widest debate on the choices we will have in every domain of human activity.

Acknowledgements

I am grateful for the guidance, input, and comments from Dr John Taylor, Prue Backway, Professors Dave Cliff, Richard Harper, David May, Richard Morris, Bob Rowthorn, Lionel Tarassenko and Ian Page. Special thanks are due to Michael Kenward for doing his best to clean up my wayward prose. All errors and shortcomings remain my own.

References

Hofstadter, Douglas (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Banic Books.

Merkle, Ralph (1989) *Foresight Update No. 6.* Foresight Institute, 1989; http://www. merkle.com/humanMemory.html; http://www.merkle.com/brainLimits.html.

Norman, Donald A. (1993) *Things That Make Us Smart*. Cambridge, MA: Perseus Publishing.

Picard, Rosalind (1997) *Affective Computing*. Boston, MA: MIT Press.

Turkle, Sherry (2002) Sociable technologies: enhancing human performance when the computer is not a tool but a companion. In *Converging Technologies for Improving Human Performance*. Boston, MA: MIT Press.

Index