168

PROGRESS IN
BRAIN RESEARCH

Models of Brain and Mind

Physical, Computational and Psychological Approaches

EDITED BY
Rahul Banerjee
Bikas K. Chakrabarti

PROGRESS IN BRAIN RESEARCH

VOLUME 168

MODELS OF BRAIN AND MIND

PHYSICAL, COMPUTATIONAL AND PSYCHOLOGICAL APPROACHES

Other volumes in PROGRESS IN BRAIN RESEARCH

Volume 131: Concepts and Challenges in Retinal Biology, by H. Kolb, H. Ripps and S. Wu (Eds.) – 2001, ISBN 0-444-50677-2.
Volume 132: Glial Cell Function, by B. Castellano López and M. Nieto-Sampedro (Eds.) – 2001, ISBN 0-444-50508-3.
Volume 133: The Maternal Brain. Neurobiological and Neuroendocrine Adaptation and Disorders in Pregnancy and Post Partum, by J.A. Russell, A.J. Douglas, R.J. Windle and C.D. Ingram (Eds.) – 2001, ISBN 0-444-50548-2.
Volume 134: Vision: From Neurons to Cognition, by C. Casanova and M. Ptito (Eds.) – 2001, ISBN 0-444-50586-5.
Volume 135: Do Seizures Damage the Brain, by A. Pitkänen and T. Sutula (Eds.) – 2002, ISBN 0-444-50814-7.
Volume 136: Changing Views of Cajal's Neuron, by E.C. Azmitia, J. DeFelipe, E.G. Jones, P. Rakic and C.E. Ribak (Eds.) – 2002, ISBN 0-444-50815-5.
Volume 137: Spinal Cord Trauma: Regeneration, Neural Repair and Functional Recovery, by L. McKerracher, G. Doucet and S. Rossignol (Eds.) – 2002, ISBN 0-444-50817-1.
Volume 138: Plasticity in the Adult Brain: From Genes to Neurotherapy, by M.A. Hofman, G.J. Boer, A.J.G.D. Holtmaat, E.J.W. Van Someren, J. Verhaagen and D.F. Swaab (Eds.) – 2002, ISBN 0-444-50981-X.
Volume 139: Vasopressin and Oxytocin: From Genes to Clinical Applications, by D. Poulain, S. Oliet and D. Theodosis (Eds.) – 2002, ISBN 0-444-50982-8.
Volume 140: The Brain's Eye, by J. Hyönä, D.P. Munoz, W. Heide and R. Radach (Eds.) – 2002, ISBN 0-444-51097-4.
Volume 141: Gonadotropin-Releasing Hormone: Molecules and Receptors, by I.S. Parhar (Ed.) – 2002, ISBN 0-444-50979-8.
Volume 142: Neural Control of Space Coding, and Action Production, by C. Prablanc, D. Pélisson and Y. Rossetti (Eds.) – 2003, ISBN 0-444-509771.
Volume 143: Brain Mechanisms for the Integration of Posture and Movement, by S. Mori, D.G. Stuart and M. Wiesendanger (Eds.) – 2004, ISBN 0-444-513892.
Volume 144: The Roots of Visual Awareness, by C.A. Heywood, A.D. Milner and C. Blakemore (Eds.) – 2004, ISBN 0-444-50978-X.
Volume 145: Acetylcholine in the Cerebral Cortex, by L. Descarries, K. Krnjević and M. Steriade (Eds.) – 2004, ISBN 0-444-51125-3.
Volume 146: NGF and Related Molecules in Health and Disease, by L. Aloe and L. Calzà (Eds.) – 2004, ISBN 0-444-51472-4.
Volume 147: Development, Dynamics and Pathology of Neuronal Networks: From Molecules to Functional Circuits, by J. Van Pelt, M. Kamermans, C.N. Levelt, A. Van Ooyen, G.J.A. Ramakers and P.R. Roelfsema (Eds.) – 2005, ISBN 0-444-51663-8.
Volume 148: Creating Coordination in the Cerebellum, by C.I. De Zeeuw and F. Cicirata (Eds.) – 2005, ISBN 0-444-51754-5.
Volume 149: Cortical Function: A View from the Thalamus, by V.A. Casagrande, R.W. Guillery and S.M. Sherman (Eds.) – 2005, ISBN 0-444-51679-4.
Volume 150: The Boundaries of Consciousness: Neurobiology and Neuropathology, by Steven Laureys (Ed.) – 2005, ISBN 0-444-51851-7.
Volume 151: Neuroanatomy of the Oculomotor System, by J.A. Büttner-Ennever (Ed.) – 2006, ISBN 0-444-51696-4.
Volume 152: Autonomic Dysfunction after Spinal Cord Injury, by L.C. Weaver and C. Polosa (Eds.) – 2006, ISBN 0-444-51925-4.
Volume 153: Hypothalamic Integration of Energy Metabolism, by A. Kalsbeek, E. Fliers, M.A. Hofman, D.F. Swaab, E.J.W. Van Someren and R.M. Buijs (Eds.) – 2006, ISBN 978-0-444-52261-0.
Volume 154: Visual Perception, Part 1, Fundamentals of Vision: Low and Mid-Level Processes in Perception, by S. Martinez-Conde, S.L. Macknik, L.M. Martinez, J.M. Alonso and P.U. Tse (Eds.) – 2006, ISBN 978-0-444-52966-4.
Volume 155: Visual Perception, Part 2, Fundamentals of Awareness, Multi-Sensory Integration and High-Order Perception, by S. Martinez-Conde, S.L. Macknik, L.M. Martinez, J.M. Alonso and P.U. Tse (Eds.) – 2006, ISBN 978-0-444-51927-6.
Volume 156: Understanding Emotions, by S. Anders, G. Ende, M. Junghofer, J. Kissler and D. Wildgruber (Eds.) – 2006, ISBN 978-0-444-52182-8.
Volume 157: Reprogramming of the Brain, by A.R. Møller (Ed.) – 2006, ISBN 978-0-444-51602-2.
Volume 158: Functional Genomics and Proteomics in the Clinical Neurosciences, by S.E. Hemby and S. Bahn (Eds.) – 2006, ISBN 978-0-444-51853-8.
Volume 159: Event-Related Dynamics of Brain Oscillations, by C. Neuper and W. Klimesch (Eds.) – 2006, ISBN 978-0-444-52183-5.
Volume 160: GABA and the Basal Ganglia: From Molecules to Systems, by J.M. Tepper, E.D. Abercrombie and J.P. Bolam (Eds.) – 2007, ISBN 978-0-444-52184-2.
Volume 161: Neurotrauma: New Insights into Pathology and Treatment, by J.T. Weber and A.I.R. Maas (Eds.) – 2007, ISBN 978-0-444-53017-2.
Volume 162: Neurobiology of Hyperthermia, by H.S. Sharma (Ed.) – 2007, ISBN 978-0-444-51926-9.
Volume 163: The Dentate Gyrus: A Comprehensive Guide to Structure, Function, and Clinical Implications, by H.E. Scharfman (Ed.) – 2007, ISBN 978-0-444-53015-8.
Volume 164: From Action to Cognition, by C. von Hofsten and K. Rosander (Eds.) – 2007, ISBN 978-0-444-53016-5.
Volume 165: Computational Neuroscience: Theoretical Insights into Brain Function, by P. Cisek, T. Drew and J.F. Kalaska (Eds.) – 2007, ISBN 978-0-444-52823-0.
Volume 166: Tinnitus: Pathophysiology and Treatment, by B. Langguth, G. Hajak, T. Kleinjung, A. Cacace and A.R. Møller (Eds.) – 2007, ISBN 978-0-444-53167-4.
Volume 167: Stress Hormones and Post Traumatic Stress Disorder: Basic Studies and Clinical Perspectives, by E.R. de Kloet, M.S. Oitzl and E. Vermetten (Eds.) – 2008, ISBN 978-0-444-53140-7.z

PROGRESS IN BRAIN RESEARCH

VOLUME 168

MODELS OF BRAIN AND MIND

PHYSICAL, COMPUTATIONAL AND PSYCHOLOGICAL APPROACHES

EDITED BY

RAHUL BANERJEE
BIKAS K. CHAKRABARTI
Centre for Applied Mathematics and Computational Science
Saha Institute of Nuclear Physics, Calcutta, India

ELSEVIER

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

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

First edition 2008

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data

Models of brain and mind : physical, computational and psychological approaches. - (Progress in brain research; v. 168)
1. Consciousness - Congresses 2. Brain - Computer simulation - Congresses
I. Banerjee, Rahul II. Chakrabarti, B. K. (Bikas K.), 1952-612.8'2
ISBN-13: 9780444530509

ISBN: 978-0-444-53050-9 (this volume)
ISSN: 0079-6123 (Series)

For information on all Elsevier publications
visit our website at books.elsevier.com

Printed and bound in The Netherlands

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

List of contributors

I. Aleksander, Department of Electrical and Electronic Engineering, Imperial College, Room 615, South Kensington Campus, London SW7 2BT, UK
R. Banerjee, Saha Institute of Nuclear Physics, Sector 1, Block AF, Bidhan Nagar, Calcutta 700064, India
R.S. Bapi, Department of Computer and Information Sciences, University of Hyderabad, P.O. Central University, Gachibowli, Hyderabad 500046, India
Abhik Basu, Theoretical Condensed Matter Physics Division, Saha Institute of Nuclear Physics, Calcutta 700064, India
Arindam Basu, Department of ENT, Guru Teg Bahadur Medical Centre, 11 D.L. Khan Road, Calcutta 700027, India
K. Bhaumik, West Bengal University of Technology, BF-142 Sector 1, Salt Lake, Calcutta 700064, India
B.K. Chakrabarti, Centre for Applied Mathematics and Computational Science, Saha Institute of Nuclear Physics, Sector 1, Block AF, Bithannagar, Calcutta 700064, India
N. Chatterjee, AU-KBC Research Centre, Anna University, Chromepet, Chennai 600044, India
D.V. Chavan, Vipassana Research Institute, Dhammagiri, Igatpuri 422403, India
A. Cleeremans, Cognitive Science Research Unit, Université Libre de Bruxelles CP 191, 50 Ave. F.-D. Roosevelt, B1050 Brussels, Belgium
S.K. Das, Department of Neuromedicine, Bangur Institute of Neuroscience and Psychiatry, Calcutta 700025, India
T. Das, National Brain Research Centre, NH-8, Nainwal Mode, Manesar 122050, India
T. Dhibar, Department of Neuroradiology, Bangur Institute of Neuroscience and Psychiatry, Calcutta, India
Z. Dienes, Department of Psychology, School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK
A. Dutt, Department of Psychology, University College of Science and Technology, 92 APC Road, Calcutta 700009, India
K. Ghosh, Centre for Soft Computing Research, Indian Statistical Institute, 203 B.T. Road, Calcutta 700108, India
A. Hazra, Department of Pharmacology, Institute of Postgraduate Medical Education and Research, 244 B A.J.C. Bose Road, Calcutta 700020, India
J.-i. Inoue, Complex Systems Engineering, Graduate School of Information of Science and Technology, Hokkaido University N14-W9, Kita-Ku, Sapporo 060-0814, Japan
R. Kasturirangan, National Institute of Advanced Study, Indian Institute of Science Campus, Bangalore 560012, India
H.C. Lau, Wellcome Trust Functional Imaging Laboratory, University College London, 12 Queen Square, London WC1N 3BG, UK; *and* Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK
R. Llinás, Department of Neuroscience and Physiology, New York University School of Medicine, MSB 4 448, 550 First Avenue, New York, NY 10016, USA
K. Mainzer, Institute of Interdisciplinary Informatics, University of Augsburg, D-86135 Augsburg, Germany

T. Metzinger, Philosophisches Seminar, Johannes Gutenberg Universität, D-55099 Mainz, Germany
H. Morton, School of Human Science and Law, Brunel University, Uxbridge, UB8 3PH, UK
P. Mukhopadhyay, Department of Psychology, University College of Science and Technology, 92 APC Road, Calcutta 700009, India
S. Roy, Physics and Applied Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700108, India
T. Roy, Department of Neuromedicine, Bangur Institute of Neuroscience and Psychiatry, Calcutta 700025, India
S. Sarkar, Microelectronics Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Calcutta 700064, India
L. Singh, National Brain Research Centre, NH-8, Nainwal Mode, Manesar 122050, India
N.C. Singh, National Brain Research Centre, NH-8, Nainwal Mode, Manesar 122050, India
S. Sinha, The Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600113, India
N. Srinivasan, Centre for Behavioural and Cognitive Sciences, University of Allahabad, Allahabad 211002, India
M. Velmans, Department of Psychology, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
S.M. Wanjerkhede, Department of Computer and Information Sciences, University of Hyderabad, P.O. Central University, Gachibowli, Hyderabad 500046, India
S. Zeki, Wellcome Laboratory of Neurobiology, University College London, Gower Street, London WC1E 6BT, UK

Guy B. Marin

Ghent, Belgium, July 2007

Preface

The last decade has seen the reemergence of consciousness as a subject suitable for scientific investigation and research. To introduce young researchers to the exciting developments in this field, a Workshop was organized under the auspices of the Centre for Applied Mathematics and Computational Science, at the Saha Institute of Nuclear Physics, Calcutta from 21st to 24th November, 2006. Judged by the enormous enthusiasm and response, from the (hundred odd) participants including the distinguished scientists and philosophers (who came as invited speakers), the Workshop must have been a grand success.

The (review) papers contained in this volume are essentially based on the proceedings of the Workshop. A significant portion of the book thus deals with the study of consciousness from a neurobiological and psychological perspective. Related philosophical issues are also extensively discussed. Although consciousness is the dominant theme, not all the chapters deal directly with consciousness. About half the volume deals with the physical modeling of brains (neural networks), modes of perception, signal transduction pathways etc. and mathematical modeling of brain functions. In addition to being important subjects in their own right, their contribution in understanding the phenomenon of consciousness will probably not be disputed. Finally, we have also included two papers dealing with Indian "first person" methods and a classical model of conscious experience derived from them. We expect that such studies might generate fruitful ideas in the near future.

We sincerely believe that the papers contained in this volume will inspire and motivate a wider cross-section of younger researchers to take up studies on the profound problems related to consciousness.

Rahul Banerjee
Bikas K. Chakrabarti
Calcutta

Contents

See Color Plate Section at the end of this book

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 1

How to separate conceptual issues from empirical ones in the study of consciousness☆

Max Velmans*

Department of Psychology, Goldsmiths, University of London, New Cross, London SE14 6NW, UK

Abstract: Modern consciousness studies are in a healthy state, with many progressive empirical programmes in cognitive science, neuroscience, and related sciences, using relatively conventional third-person research methods. However not all the problems of consciousness can be resolved in this way. These problems may be grouped into problems that require empirical advance, those that require theoretical advance, and those that require a re-examination of some of our pre-theoretical assumptions. I give examples of these, and focus on two problems — what consciousness *is*, and what consciousness *does* — that requires all three. In this, careful attention to conscious phenomenology and finding an appropriate way to relate first-person evidence to third-person evidence appears to be central to progress. But we may also need to re-examine what we take to be "natural facts" about the world, and how we can know them. The same appears to be true for a trans-cultural understanding of consciousness that combines classical Indian phenomenological methods with the third-person methods of Western science.

Keywords: consciousness; mind; brain; cognitive science; neuroscience; hard problem; easy problem; first-person; third-person; phenomenology; Indian philosophy; dualism; materialism; reductionism; reflexive monism; causation; causal problem; natural fact

What are the problems of consciousness?

Traditionally, the puzzles surrounding consciousness have been known as the "mind-body" problem. However, it is now clear that "mind" is not quite the same thing as "consciousness", and that the aspect of body most closely involved with consciousness is the brain. It is also clear that there is not one consciousness–brain problem, but many.

As a first approximation, these can be divided into five groups, each focused on a few, central questions:

Problem 1. What and where is consciousness?

Problem 2. How are we to understand the *causal relationships* between consciousness and matter and, in particular, the causal relationships between consciousness and the brain?

Problem 3. What is the function of consciousness? How, for example, does it relate to human information processing?

☆This chapter is based on a paper given at an International Workshop on "Models of Brain and Mind: Physical, Computational and Psychological Approaches" hosted by the Saha Institute of Nuclear Physics, Kolkata, 21st–24th November, 2006, and was partly supported by British Academy Overseas Conference Travel Grant: OCG42502.

*Corresponding author. Tel.: +44 (0)20 7919 7870; Fax: +44 (0)20 7919 7873; E-mail: m.velmans@gold.ac.uk

DOI: 10.1016/S0079-6123(07)68001-7

Problem 4. What *forms of matter* are associated with consciousness — in particular, what are the neural substrates of consciousness in the human brain?

Problem 5. What are the appropriate ways to *examine* consciousness, to discover its nature? Which features can we examine with first-person methods, which features require third-person methods, and how do first- and third-person findings relate to each other?

According to Chalmers (1995) the problems of consciousness may be divided into the "easy" problems and the "hard" problem. "Easy" problems are ones that can be researched by conventional third-person methods of the kind used in cognitive science, for example, investigations of the information processing that accompanies subjective experience. The "hard" problem is posed by subjective experience itself. As he notes:

> It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does. If any problem qualifies as *the* problem of consciousness, it is this one (Chalmers, 1995, p. 201).

Following the strenuous efforts in the late 20th century to demonstrate subjective experience to be nothing more that a state or function of the brain (see review in Velmans, 2000, Chapters 3, 4, and 5), Chalmers' "easy" versus "hard" problem distinction provided a useful reminder that a purely third-person functional analysis of human information processing cannot reveal what it is like to have a subjective experience or explain why it arises (see also Velmans, 1991a). However, this division of the problems of consciousness into "easy" and "hard" ones was, in turn, an oversimplification. As Chalmers himself accepted, even so-called "easy" (empirically researchable) problems can in practice be very difficult to solve. It may also be that the "hard" problem only seems unusually hard because we have been thinking about it in the wrong way. If so, changing some of our unexamined assumptions might be all we need to make the problem "easy".

For example, in contrast to consciousness, we usually take the existence of matter for granted, and we assume that physics does not present similarly "hard" problems. But there are many. Why, for example, should electricity flowing down a wire be accompanied by a magnetic field around the wire, why should photons sometimes behave as waves and at other times as particles, and why there should be any matter in the universe at all? We simply assume these to be natural facts that we can observe in the world. We can try to explain them by incorporating them into some body of theory, but we do not agonize over their *existence*. It might equally be a natural and irreducible fact about the world that certain forms of brain functioning are accompanied by certain forms of first-person experience. That would require us to change a few of our pre-theoretical assumptions about the nature of matter and its relationship to consciousness, and we would still have to investigate the principles that govern the consciousness–brain relationship in great detail. But the fact that given conscious states accompany certain forms of brain functioning would then be "hard" to understand in the same sense as many facts in physics.

Given this, it seems more useful to sort the problems of consciousness into those that require empirical advance, those that require theoretical advance, and those that require a re-examination of some of our pre-theoretical assumptions.[1] If, for

[1]Some problems require a combination of these approaches.

example, the problem is ''What are the neural substrates of consciousness?'' or ''What forms of information processing are most closely associated with consciousness?'', then conventional cognitive and neuropsychological techniques look as if they are likely to yield useful results. There are many questions of this empirical kind and, consequently, the new ''science of consciousness'' is already very large (see the extensive reviews and readings in Velmans and Schneider, 2007).

Examples of empirical questions and investigations within neuropsychology include:

- The search for the neural causes and correlates of major changes in normal, global conscious states such as deep sleep, rapid eye movement dreaming, and the awake state.
- The search for added neural conditions that support variations in conscious experience within normal, global states, such as visual, auditory and other sensory experiences, experiences of cognitive functioning (the phonemic and other imagery accompanying thinking, metacognition, etc.), and affective experience.
- The search for neural conditions that support altered states of consciousness in psychopathology and in non-pathological altered states, such as the hypnotic state, some drug-induced states, meditation, and mystical states.

Examples of empirical questions and investigations within cognitive psychology include:

- Examination of the timing of conscious experience. When in the course of human information processing (e.g., in input analysis) does a conscious experience arise?
- The determination of functional conditions that suffice to make a stimulus conscious. For example, does material that enters consciousness first have to be selected, attended to and entered into working memory or a ''global workspace''?
- The investigation of functional differences between preconscious, unconscious, and conscious processing (e.g., in studies of non-attended vs. attended material).

Given that we can get an empirical handle on such investigations, it is sometimes assumed in the consciousness studies literature that these problems are *entirely* empirical — and even that *all* the problems of consciousness will eventually be resolved in this way.

But it is easy to show that this is not so. One might think, for example, that Problem 1, the nature and location of consciousness, should be easy to resolve, as we all have access to and information about our own consciousness. However both its nature and location are much disputed in the literature — and the same may be said about the enduring puzzles and disputes surrounding the causal relationships between consciousness and the brain (Problem 2). Although empirical progress can be made with many questions (of the kind listed above) without first settling such disputes, we cannot in the end ignore them, for the simple reason that pre-theoretical assumptions, theories, and empirical problems interconnect. How, for example, could one arrive at an agreed understanding of the neural correlates of consciousness without an agreed understanding of what consciousness is — and without an understanding of how consciousness *could* have causal effects on matter, how can one determine its function in the workings of the brain?

Problems of definition

According to Nagel (1974) consciousness is ''what it is like to be something.'' Without it, after all, it would not be like anything to exist. While it is generally accepted in modern philosophy of mind that this does capture something of the essence of the term, no universally agreed ''core meaning'' exists. This is odd, as we each have ''psychological data'' about what it is like to *be conscious* or to *have consciousness* to serve as the basis for an agreed definition.

This uncertainty about how to define consciousness is partly brought about by the way global theories about consciousness (or even about the nature of the universe) have intruded into definitions. In the classical Indian tradition, for example, in the Upanishads, consciousness is thought to

be, in essence, transcendental, for example, as Ātman (sometimes identified with Brāhman) — a pure, subject-object-less consciousness that underlies and provides the ground of being of both Man and Nature (Saksena, 1965). In the classical Western tradition, ''substance dualists'' such as Plato and Descartes bifurcated the universe, believing it to consist of two fundamental kinds of stuff, material stuff and the stuff of consciousness (a substance associated with soul or spirit). Following the success of the brain sciences and related sciences, 20th century theories of mind in the West became increasingly materialistic, assuming physical ''stuff'' to be basic and consciousness in some way ''supervenient'' or dependent on the existence of physical forms. For example, ''property dualists'' such as Sperry and Libet took consciousness to be a special kind of property that is itself non-physical, but which emerges from physical systems such as the brain once they attain a certain level of complexity. Taking materialism to its logical conclusion, ''reductionists'' such as Crick (1994) and Dennett (1991) argued consciousness to be nothing more than a state or function of the brain. Within cognitive psychology, there were many similar reductive proposals which identified consciousness with some aspect of human information processing, for example, with working memory, focal attention, a central executive, or a ''global workspace''.

I do not have space to examine the arguments for and against these and many other proposals here (see Velmans, 2000, Chapters 2–5). Suffice it to say that these differing claims about consciousness often start more from some pre-existing *theory* about the nature of the mind or world than from the *everyday phenomenology of consciousness itself*. In the modern literature, for example, Dennett provides a prominent example of the triumph of materialist theory over phenomenological evidence when he tries to deny the very existence of phenomenal qualities (as normally understood). He makes this perfectly clear when he writes:

> Philosophers have adopted various names for the things in the beholder (or properties of the beholder) that have been supposed to provide a safe home for the colors and the rest of the properties that have been banished from the external world by the triumphs of physics: raw feels, phenomenal qualities, intrinsic properties of conscious experiences, the qualitative content of mental states, and, of course, qualia, the term I use. There are subtle differences in how these terms have been defined, but I am going to ride roughshod over them. *I deny that there are any such properties. But I agree wholeheartedly that there seem to be* (Dennett, 1991, p. 372).

Dennett arrives at this view by presupposing that while information obtained from a third-person perspective is ''scientific'' and reliable, first-person information has no credence at all. Indian philosophy assumes the opposite to be true. There are, of course, also many Western views that take conscious phenomenology seriously — and I have given a detailed critique of Dennett's position that I do not have space to repeat here (see Velmans, 2000, Chapter 5, 2001, 2007a, b). It should be evident however that such oppositions neatly exemplify a situation where pre-theoretical assumptions (in this case about the nature of the world and how we can know it) motivate the argument.

There can be no point of convergence and certainly no consensus between researchers who take the existence of conscious phenomenology to be both self-evident and important, with those who give no credence to that phenomenology at all. Classical Indian investigations of consciousness, for example, have been *primarily* phenomenological. Note however that their conclusions about the nature of consciousness arise largely from altered conscious states consequent on prolonged periods of meditation, and this can be another, potential source of confusion in East–West discussions focused on the nature of *everyday* conscious phenomenology. The pure, contentless consciousness said to be experienced in such states is, in various writings, thought to underlie all of Nature, which makes this a claim about what in

the West is sometimes referred to as ''the ground of being'' or, in Kantian terms, ''the thing in itself,'' rather than a claim about the forms of ''phenomenal consciousness,'' that are more usually investigated in modern consciousness studies.

In short, to make progress towards some global understanding of consciousness, we may first have to take a step backwards, and re-examine our different points of departure and the presuppositions that support them. Empirical investigations on their own won't always do.

To what does the term ''consciousness'' refer?

To investigate any phenomenon, one first has to ''point to'' or pick out the phenomenon that one wishes to investigate sufficiently well for independent investigators to be confident that they are investigating the same thing. For many researchers, the phenomenology of everyday conscious experience (often referred to as ''phenomenal consciousness'') provides a natural place to start.

In some writings ''consciousness'' is synonymous with ''mind.'' However, given the extensive evidence for nonconscious mental processing this definition of consciousness is too broad. In Western psychology, ''mind'' typically refers to psychological states and processes that may or may not be ''conscious''.

In other writings ''consciousness'' is synonymous with ''self-consciousness''. As one can be conscious of many things other than oneself (other people, the external world, etc.), this definition is too narrow. Here, self-consciousness is taken to be a special form of *reflexive* consciousness in which the object of consciousness is the self or some aspect of the self.

The term ''consciousness'' is also commonly used to refer to a state of wakefulness. Being awake or asleep or in some other state such as coma clearly influences what one can be conscious of, but it is not the same as being conscious in the sense of having ''phenomenal contents''. When sleeping, for example, one can still have visual and auditory experiences in the form of dreams. Conversely, when awake there are many things at any given moment that one does *not* experience. So in a variety of contexts it is necessary to distinguish ''consciousness'' in the sense of ''phenomenal consciousness'' from wakefulness and other states of arousal, such as dream sleep, deep sleep, and coma.

''Consciousness'' is also sometimes used to mean ''knowledge'', in the sense that if one is conscious of something one also has knowledge of it. However, at any moment, much knowledge is nonconscious, or implicit (e.g., the knowledge gained over a lifetime, stored in long-term memory). So consciousness and knowledge cannot be co-extensive.

The above distinctions are quite widely accepted in the contemporary scientific literature (see, Farthing, 1992; readings in Velmans and Schneider, 2007) although it is unfortunate that various writers, both East and West, adopting different philosophical positions and pre-theoretical assumptions, continue to use the term ''consciousness'' in very different ways. One might think, for example, that confining the use of the term ''consciousness'' to ''phenomenal consciousness'' (its everyday phenomenology), might be enough to reach at least an initial agreement about what it is that we are investigating. Unfortunately, nothing could be further from the truth — as differing philosophical positions make very different claims even about the *global* nature of conscious phenomenology. Following Descartes, for example, Dualists maintain that conscious experiences are composed of ''thinking stuff'' (res cogitans) that, unlike material stuff (res extensa), has no location or extension in space. Materialist Reductionists claim something very different. If consciousness is nothing more than a state or function of the brain then conscious experiences must also have a location and extension in the brain. Reflexive Monism argues against both of these positions on the grounds that both give a false account of our everyday experience. Rather than the phenomenology of our conscious experiences being either ''nowhere'' or ''in the brain'' the contents of consciousness include the entire phenomenal world that has phenomenal location and extension beyond our experienced bodies up to the dome of the sky. Experienced objects and events also

have an experienced location and extension within that phenomenal space. Such disputes about the location and extension of conscious phenomenology are not merely ''philosophical,'' for the reason that they invite different avenues of scientific investigation. Reflexive Monism, for example, invites an investigation of how preconscious processes within the brain, interacting with the surrounding world, support experiences of objects and events that appear to be outside the brain (''perceptual projection'') and of how visual processing supports an entire, experienced, three-dimensional, phenomenal world.

Again, I do not have space to elaborate on this here (c.f. Velmans, 2000, Chapters 6–12, 2007a). I simply want to stress, once again, that agreeing on a common point of departure is important. Once a given reference for the term ''consciousness'' is fixed in its *phenomenology*, the investigation of its nature can begin, and this may in time transmute the meaning (or sense) of the term. As Dewey (1910) notes, to grasp the meaning of a thing, an event or situation is to see it in its relations to other things — to note how it operates or functions, what consequences follow from it, what causes it, and what uses it can be put to. Thus, to understand what consciousness is, we need to understand what causally affects it, what its function (s) may be, how it relates to nonconscious processing in the brain, and so on. As our scientific understanding of these matters deepens, our understanding of what consciousness *is* will also deepen. A similar transmutation of meaning (with growth of knowledge) occurs with basic terms in physics such as ''energy'', and ''time''.

Conceptual problems surrounding the function of consciousness and the causal interaction of consciousness and brain

Let us turn now to the question of how our presuppositions influence our theories about what consciousness *does* (rather than about what it *is*). Cognitive theories of consciousness often assume it to be some form of information processing. Consequently, cognitive studies of consciousness commonly assume that functional differences between ''conscious'' and ''preconscious'' or ''unconscious'' processing will reveal the function of consciousness (see review in Velmans, 2000, Chapter 4; Baars, 2007). However, researchers adopting this approach are seldom clear about the precise *sense* in which the process under investigation can be said to ''be conscious.'' As noted in Velmans (1991a) a process can be said to be ''conscious''

(a) in the sense that one is conscious *of* the process;
(b) in the sense that the operation of the process is *accompanied* by consciousness (of its *results*); and
(c) in the sense that consciousness *enters into* or *causally influences* the process.

Why does this matter? It is only sense (c) that is relevant to claims that consciousness has a third-person causal or functional role — and, crucially, one cannot assume a process to be conscious in sense (c) on the grounds that it is conscious in senses (a) or (b). Sense (a) is also very different to sense (b). Sense (a) has to do with what experiences *represent*. Normal conscious states are always *about something*, that is they provide information to those who have them about the external world, body or mind/brain itself. Some mental processes (problem solving, thinking, planning, etc.) can be said to be partially ''conscious'' (in this sense) in so far as their detailed operations are accessible to introspection. Sense (b) contrasts different *forms* of mental processing. Some forms of mental processing result in conscious experiences, while others do not. For example, analysis of stimuli in attended channels usually results in a conscious experience of those stimuli, but not in non-attended channels.

Theories that attribute a third-person causal role to consciousness solely on the basis of functional contrasts between ''conscious'' and preconscious or unconscious processes invariably conflate these distinctions. They either take it for granted that if a process is conscious in sense (a) or sense (b) then it must be conscious in sense (c). Or they simply *redefine* consciousness to be a form of processing, such as focal attention, information in a ''limited capacity channel,'' a ''central executive,'' a ''global workspace'' and so on, thereby

begging the question about the functional role of conscious *phenomenology* in the economy of the mind.

Further problems with conscious causation

While Dualist theories of consciousness do not make such question-begging assumptions about conscious phenomenology, they have problems with conscious causation that are equally serious. According to Plato and Descartes, the material body *interacts* with the soul. In the acquisition of knowledge, the body influences the soul through the operation of its senses. The soul is the source of consciousness and reason, and through the exercise of will, it manipulates the body. Eccles (1980), a modern Dualist, defended a similar view — although he replaced the term ''soul'' with ''the self-conscious mind.'' But how could such causal interactions between brain and soul or self-conscious mind take place?

Causal Problem 1. The physical world appears causally closed.

From an external, third-person perspective one can, in principle, trace the effects of input stimuli on the central nervous system from input to output, without finding any ''gaps'' in the chain of causation that consciousness might fill. Additionally, if one inspects the brain from the outside, no subjective experience can be observed at work. Nor does one need to appeal to the existence of subjective experience to account for the neural activity that one *can* observe. The same is true if one thinks of the brain in systems-theory terms. Once the processing within a system required to perform a given function is sufficiently well specified in procedural terms, for example, in terms of the information processing required, one does not have to add an ''inner conscious life'' to make the system work.

Causal Problem 2. The Conservation of Energy Principle.

If non-material conscious experiences are to influence physical events, physical energy must be created from some non-material source, and the total physical energy of the universe is thereby increased. Equally, for physical events to influence conscious ones, energy must be drawn from the physical universe. However, according to the Conservation of Energy Principle energy can neither be created nor destroyed.

Causal Problem 3. One is not conscious of one's own brain/body processing. So how could there be conscious control of such processing?

We normally think of speech production as being under ''conscious control''. But in what sense does one have conscious control of the articulatory system? In speech, the tongue may make as many as 12 adjustments of shape per second that need to be coordinated with other rapid, dynamic changes within the articulatory system. In 1 min of discourse as many as 10–15,000 neuromuscular events occur (Lenneberg, 1967). Yet only the results of this activity (the overt speech) normally enter consciousness. According to Eccles, the self-conscious mind controls activities in the motor cortex through the exercise of free will. But how could a consciously experienced wish to do something activate neurons or move muscles? The processes required to activate neurons are not even represented in consciousness! For example, the phenomenology of a ''wish'' includes no details of where our motor neurons are located, let alone how to activate them. Consequently, if some aspect of the mind does control the momentary activities of neurons, that aspect of the mind must be *nonconscious* — which involves a paradox.

Causal Problem 4. Conscious experiences appear to come too late to causally affect the processes to which they most obviously relate. Speech production, speech perception, and reading are among the most complex forms of ''conscious'' human information processing. Yet, as noted above, in conscious speech production there is a sense in which one is only conscious of what one wants to say *after one has said it*! The same is true of conscious reading. Try silently reading the following sentence: *The forest ranger did not permit us to enter the park without a permit.* Notice that, on its first occurrence in your phonemic imagery or

"covert speech", the word "permit" was (silently) pronounced with the stress on the second syllable (per*mit*) while on its second occurrence the stress was on the first syllable (*per*mit). But how and when did this allocation of stress patterns take place? Clearly, the syntactic and semantic analysis required to determine the appropriate meanings of the word "permit" must have taken place prior to the allocation of the stress patterns; and this, in turn, must have taken place *prior* to the phonemic images entering awareness. Note too, that while reading, one is not conscious of the analysis and identification of individual words or of any syntactic or semantic analysis being applied to the sentence. Nor is one aware of the processing responsible for the resulting covert speech (with the appropriate stress patterns on the word "permit"). The same may be said of the paragraph you are now reading, or of the entire text of this chapter.

Oddly, a similar sequence of events occurs with conscious verbal thoughts Once one *has* a conscious verbal thought, in the form of phonemic imagery, the complex cognitive processes required to generate that thought, including the processing required to encode it into phonemic imagery have already operated — and the same is true of conscious feelings (Panksepp, 2007) and volitions (Banks and Pockett, 2007).

How can one make sense of conscious causation?

In the cognitive literature consciousness has been thought by one or another author to enter into every main phase of human information processing, that is, in input, storage, transformation, and output. This includes stimulus analysis, selective attention, learning, memory, choosing, thinking, planning, language use, and the control and monitoring of complex overt response (see reviews in Velmans, 1991a, 2000, Chapter 4). Viewed from a first-person perspective, it seems intuitively plausible that many of these claims are true. It seems obvious, for example, that one cannot identify new or complex stimuli without first being conscious of them. Nor can one enter information into long-term memory, making it part of one's "psychological past," without it having been in one's conscious "psychological present." Nor can one transform information through problem solving, thinking or output the results in speech or writing unless one is conscious of that information.

Yet, viewed from a purely third-person perspective it is very difficult to see how *any* of these claims are true. To the best of our knowledge there are no "gaps" in the chain of neurophysiological events which require the intervention of consciousness to make the brain work (the physical world is causally closed). And, as George Miller pointed out in 1962, the same functions appear to be realisable in physical and electrical systems entirely without consciousness. In short, from a third-person perspective epiphenomenalism seems true, while from a first-person perspective it seems false. In Velmans (1991a, b, 1993, 1996), I called this the "causal paradox" — and I suggested that a good theory of how conscious phenomenology relates to brain processing needs to show, not that the view from either the first- or third-person perspective is false, but rather how it can be that both perspectives might yield true insights, for example, if they are complementary. Once again, I don't have space to go into the debates surrounding this complex issue here (see Velmans, 1991a, b, 1993, 1996, 2000, 2002a, b, 2003).

Conclusion

Modern consciousness studies are in a healthy state, with extensive, well-established research programmes in the cognitive sciences, neurosciences, and related sciences. However we should not lose sight of the fact that deep puzzles remain, and that resolving these may have more to do with the need to re-examine basic assumptions embedded in prevailing philosophical positions than with anything that can be resolved by empirical research. Such a re-examination becomes particularly important when the field becomes trans-cultural, for example, in some future blend of Indian philosophy, rooted in first-person meditative investigations, with Western-style third-person empirical research. As with the "causal paradox" discussed above, the challenge

for a global theory of consciousness may not be to show that the conclusions drawn from one or the other tradition are false, but rather how the perspectives adopted by both of these traditions might offer useful, complementary insights into the nature of consciousness.

References

Baars, B. (2007) The global workspace theory of consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell, New York.

Banks, W.P. and Pockett, S. (2007) Benjamin Libet's work on the neuroscience of free will. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell, New York.

Chalmers, D. (1995) Facing up to the problems of consciousness. J. Conscious. Stud., 2(3): 200–219.

Crick, F. (1994) The Astonishing Hypothesis: The Scientific Search for the Soul. Simon & Schuster, London.

Dennett, D.C. (1991) Consciousness Explained. Allen Lane, The Penguin Press, London.

Dewey, J. (1910) How We Think. Prometheus, Buffalo, NY.

Eccles, J.C. (1980) The Human Psyche. Springer, New York.

Farthing, J.W. (1992) The Psychology of Consciousness. Prentice-Hall, Englewood Cliffs, NJ.

Lenneberg, E.H. (1967) Biological Foundations of Language. Wiley, New York.

Nagel, T. (1974) What it is it like to be a bat? Philos. Rev., 83: 435–451.

Panksepp, J. (2007) Affective consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell, New York.

Saksena, S.K. (1965) The story of Indian philosophy. In: Ferm V. (Ed.), History of Philosophical Systems. Littlefield, Adams & Co., Paterson, NJ.

Velmans, M. (1991a) Is human information processing conscious?. Behav. Brain Sci., 14(4): 651–701.

Velmans, M. (1991b) Consciousness from a first-person perspective. Behav. Brain Sci., 14(4): 702–726.

Velmans, M. (1993) Consciousness, causality and complementarity. Behav. Brain Sci., 16(2): 404–416.

Velmans, M. (1996) Consciousness and the "causal paradox". Behav. Brain Sci., 19(3): 537–542.

Velmans, M. (2000) Understanding Consciousness. Routledge, Psychology Press, London.

Velmans, M. (2001). Heterophenomenology versus critical phenomenology: a dialogue with Dan Dennett. On-line debate at http://cogprints.soton.ac.uk/documents/disk0/00/00/17/95/index.html

Velmans, M. (2002a) How could conscious experiences affect brains? J. Conscious. Stud., 9(11): 3–29.

Velmans, M. (2002b) Making sense of causal interactions between consciousness and brain. J. Conscious. Stud., 9(11): 69–95.

Velmans, M. (2003) Preconscious free will. J. Conscious. Stud., 10(12): 42–61.

Velmans, M. (2007a) Dualism, reductionism, and reflexive monism. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell, New York.

Velmans, M. (2007b). Heterophenomenology versus critical phenomenology. Phenomenol. Cogn. Sci., 6: 221–230.

Velmans M. and Schneider S. (Eds.), (2007). The Blackwell Companion to Consciousness. Blackwell, New York.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 2

The disunity of consciousness

Semir Zeki*

Wellcome Laboratory of Neurobiology, University College London, Gower Street, London WC1E 6BT, UK

Abstract: Consciousness is commonly considered to be a single entity, as expressed in the term ''unity of consciousness'', and neurobiologists are fond of believing that, sooner or later, they will be able to determine its neural correlate (rather than its neural correlates). Here I propose an alternative view, derived from compelling experimental and clinical studies of the primate visual cortex, which suggest that consciousness is not a single unity but consists instead of many components (the micro-consciousnesses) which are distributed in space and time. In this article, I propose that there are multiple consciousnesses which constitute a hierarchy (Zeki and Bartels, 1998, 1999), with what Kant (1996) called the 'synthetic, transcendental' unified consciousness (that of myself as the perceiving person) sitting at the apex. Here, I restrict myself to writing about visual consciousness and, within vision, mainly about the colour and the visual motion systems, about which we know relatively more. For if it can be shown that we are conscious of these two attributes at different times, because of spatially and temporally different mechanisms, then the statement that there is a single, unified consciousness cannot be true.

Keywords: consciousness; micro-consciousness; unity of consciousness; visual brain; colour; motion; perceptual asynchrony

Functional specialization in the visual brain

The foundation stone for my argument rests on the fact of functional specialization in the visual brain (Zeki, 1978; Livingstone and Hubel, 1988; Zeki et al., 1991), from which several consequences follow. By general agreement, this functional specialization is especially true of the colour and the visual motion systems, which occupy geographically distinct locations in the visual cortex (Fig. 1). A pivotal area for the colour system is the V4 complex, and for the visual motion system the V5 complex (Zeki, et al., 1991; Wade et al., 2002). There is substantial agreement that the two systems have distinct, and characteristic, anatomical inputs, despite the many anatomical opportunities for them to interact. The geographical separation of the two systems constitutes the cornerstone of a 'theory of multiple consciousnesses'.

Further support comes from the generally accepted clinical evidence that lesions of V4 and of V5 lead to different visual disabilities, the former resulting in an achromatopsia (acquired colour blindness) (Meadows, 1974; Zeki, 1990) and the latter in an akinetopsia (Zeki, 1991; Zihl et al., 1991) (acquired visual motion blindness) (Fig. 1). Crucially, a lesion in one area does not invade and disable the perceptual territory of the other. Thus an akinetopsic patient sees colours consciously even though unable to perceive and be conscious of (fast) motion. By contrast, an

*Corresponding author. Tel.: +44(0)2076797316;
Fax: +44(0)2076797316; E-mail: zeki.pa@ucl.ac.uk

DOI: 10.1016/S0079-6123(07)68002-9

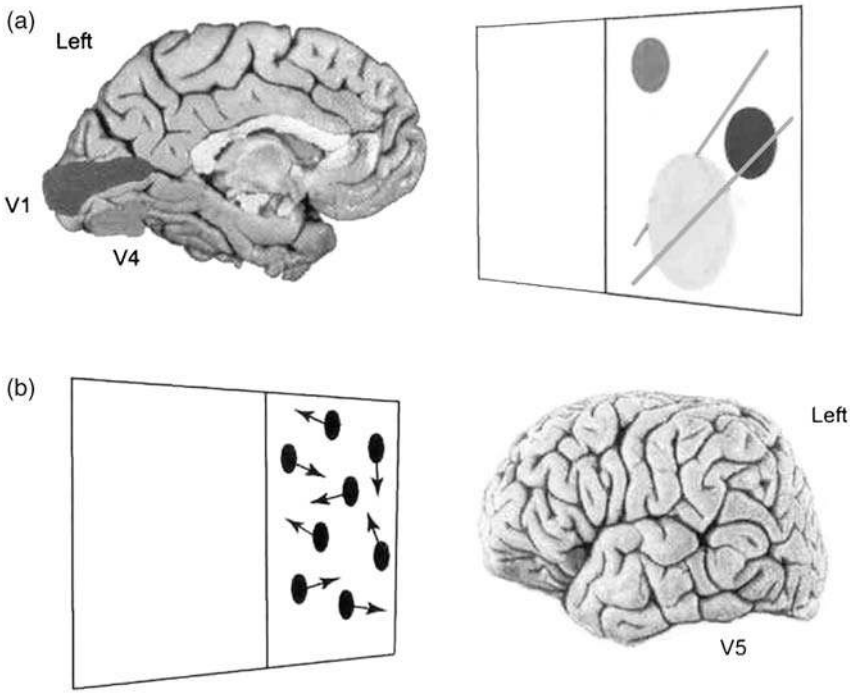


Fig. 1. Visual areas V4 (a) and V5 (b) of the human brain, specialized for colour and motion, respectively. Each receives inputs from the primary visual cortex (V1) and registers the relevant activity in the contralateral hemifield. Lesions in V4 produce achromatopsia — the inability to see colours; motion vision remains intact. Lesions in V5 produce akinetopsia, the inability to see motion; colour vision is unaffected. (See Color Plate 2.1 in color plate section.)

achromatopsic patient is unable to perceive and be conscious of colours but is able to see and be conscious of visual motion effortlessly. Hence consciousness of these elementary visual attributes are distinct from one another and I speak of them as 'micro-consciousnesses' (Zeki and Bartels, 1999). Of course, to perceive something is to be conscious of it and thus to say 'perceiving consciously' is to be tautologous, but the tautology serves to emphasize a point that is not always clearly made.

Processing sites are also perceptual sites

One conclusion from the clinical evidence is that a micro-consciousness for colour or visual motion is generated through activity at a distinct processing site, and therefore that a processing site is also a perceptual site. Such a conclusion is reinforced by studies of the visual motion centre, area V5, which receives a direct visual input that bypasses the primary visual cortex (area V1) (Fig. 2) (Cragg, 1969; Fries, 1981; Yukie and Iwai, 1981; Standage and Benevento, 1983). The perceptual consequences of this anatomical arrangement have been well studied in patient GY, blinded in one hemifield in childhood by damage to V1. Our psychophysical and imaging experiments (Barbur et al., 1993; Zeki and ffytche, 1998), since confirmed (Weiskrantz et al., 1995), have shown that, in spite of his blindness, this direct visual input to V5 (Beckers and Zeki, 1995; ffytche et al., 1995; Buchner et al., 1997) is sufficient to give GY a crude but conscious vision for fast moving, high contrast stimuli, the perception of which is mediated by V5 (Zeki and ffytche, 1998) (Fig. 2). It has also been shown that his consciousness, when visually stimulated, is visual (Stoerig and Barth, 2001). These findings

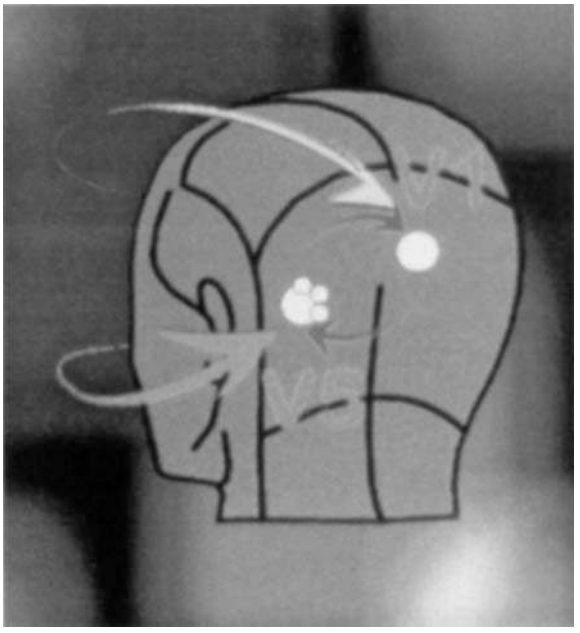


Fig. 2. The flow of visual information from the retina to V1 and V5. Notice that V5 receives a dual input from the retina, one through V1 and another that bypasses V1. (See Color Plate 2.2 in color plate section.)

suggest that, contrary to previous assumptions (Stoerig and Cowey, 1995; Lamme, 2001; Rees et al., 2002), conscious vision is possible without V1 and also that, if one can channel an appropriate visual input to a specialized visual area, then activity in it can result in a conscious correlate even if it is deprived of one of its other major sources of visual input. It is thus incorrect to think of prestriate cortex as being not 'conscious' cortex (Weiskrantz, 1990). Moreover, the switch from a state when GY is not conscious of visual stimuli and cannot therefore discriminate them correctly to one when he is conscious of them and can therefore discriminate them correctly is accompanied by a significant increase in activity in area V5, not elsewhere (Zeki and ffytche, 1998). This has led us to propose that it is heightened activity within a specialized cortical area that leads to conscious vision and that its absence (or lower activity) in the same area correlates with a lack of conscious experience, a proposal that has been confirmed in other systems, not related to visual motion (Rees et al., 2000) or even exclusively to vision (Dehaene et al., 2001).

Direct evidence that cortical processing sites are also perceptual sites comes from combined psychophysical-imaging experiments in humans (Moutoussis and Zeki, 2002). Using dichoptic stimulation, where identical visual stimuli are presented for brief periods to the two eyes separately, thus leading to binocular fusion, one finds that when the two stimuli are identical in every respect (e.g., an outline red house, or face, against a green background), subjects are able to identify (i.e., perceive) the stimulus correctly. But when the stimuli presented to the two eyes are of reverse colour contrast (e.g., outline red house against a green background to the right eye and outline green house against a red background to the left eye), subjects report seeing only yellow. Under these conditions, imaging experiments show that the same specific areas of the brain, specialized for the processing and seeing of houses (or faces), are active, regardless of whether the subjects saw the stimulus (were conscious of it) or not (see Box 1). The difference between the two states is that, in the former, the activity is higher than in the latter, although we do not know yet whether this is owing to the recruitment of previously inactive cells, to an increased discharge of already active cells, or to an increase in synaptic input without an increase in firing rate (Logothetis et al., 2001). This direct evidence obviates the need to postulate separate cortical area(s) necessary for perception, as opposed to non-conscious processing. Of course, processing–perceptual sites are not sufficient on their own in generating a conscious correlate but depend on enabling systems in the brain stem (Zeki and ffytche, 1998) and possibly additional uncharted cortical systems.

The asynchrony and temporal hierarchy of visual perception

Much has been written about the ability of an intricate system such as the visual brain, with its many parts and distributed parallel pathways, to process all attributes of the visual world simultaneously and thus provide a visual image in which

Box 1. **Combining psychophysics and imaging to demonstrate cortical processing–perceptual sites.**

In the experiment of Moutoussis and Zeki (2002) subjects viewed pictures of houses, faces, and uniformly coloured control squares dichoptically. When the stimuli presented to the two eyes were of reverse colour contrast (e.g., an outline green house against a red background to the left eye and an outline red house against a green background to the right eye, Fig. 3a), subjects reported seeing no object but only a uniform yellow field, as in the case when uniform red and green fields were presented.

The group results from the concurrent fMRI recording revealed that brain activation was correlated with both perceived and non-perceived conditions (Fig. 3c), but the level of activation was higher in the perceived condition. The contrast same houses minus same faces (SH−SF) shows bilateral stimulus-specific activation in the parahippocampal gyrus (Talaraich coordinates: −30, −44, −12 and 26, −44, −10). The contrast opposite houses minus opposite faces (OH−OF) shows unilateral stimulus-specific activation in the same region (−38, −42, −10). The contrast SF−SH reveals stimulus-specific activation in a region of the fusiform gyrus (42, −82, −12), and the contrast OF–OH reveals stimulus-specific activation in the brain region (44, −74, −14). This experiment provides direct evidence that cortical processing sites are also perceptual sites.

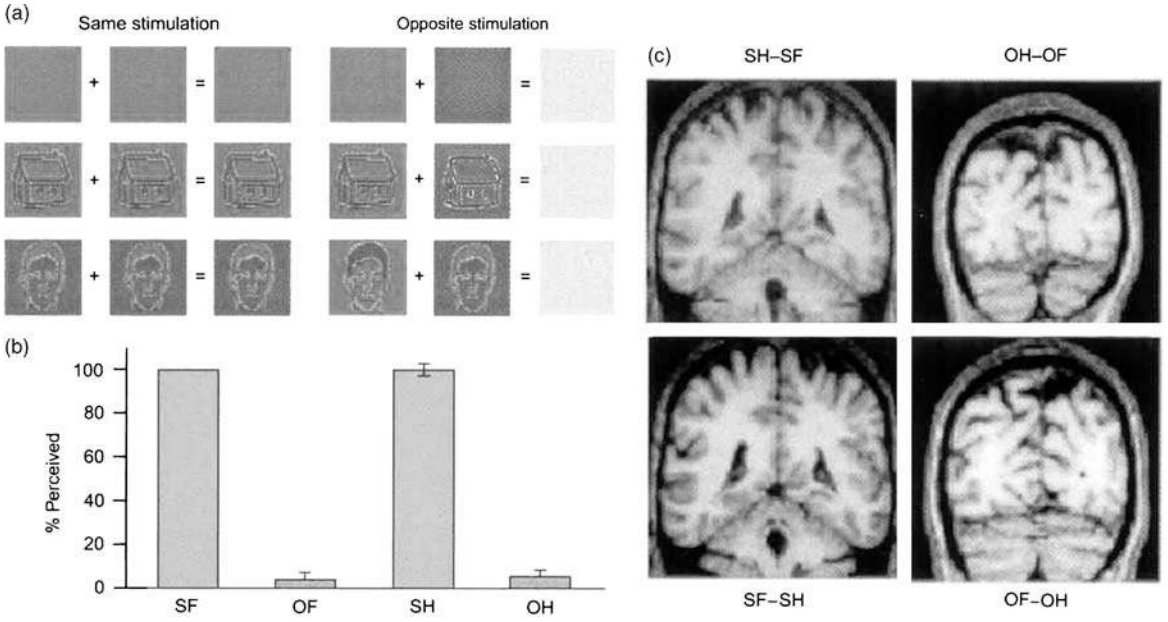


Fig. 3. The stimulation method and results from the experiment in Moutoussis and Zeki (2002). (a) The input to the two eyes and the expected perceptual output. Dichoptic stimuli of opposite colour contrast (Opposite stimulation) between the two eyes were invisible, whereas identical stimuli of the same colour contrast (Same stimulation) were easily perceived. Continuous fusion of the stimuli was achieved by using repetitive brief representations. (b) The average performance of the seven subjects in the discrimination task (face or house stimuli vs. uniform yellow). The averaged percentage of the number of stimuli perceived (of a total of 448 per subject per stimulus category), shown with the standard error between subjects. SF, same faces; OF, opposite faces; SH, same houses; OH, opposite houses. (c) Group fMRI results: brain regions showing stimulus-specific activation under conditions of same and opposite stimulation, revealing that such activation correlates with perceived and not perceived conditions (see text for details). (Modified from Moutoussis and Zeki, 2002.) (See Color Plate 2.3 in color plate section.)

all the different attributes are seen in perfect spatial and temporal registration. Our direct psychophysical results (Moutoussis and Zeki, 1997a, b; Zeki and Moutoussis, 1997), now confirmed (Barbur et al., 1998; Arnold et al., 2001), show that this is not true over brief time windows. In particular, it has been shown that colour is perceived before motion by ~80 ms. Nor is the perceptual asynchrony limited to colour and motion, because it has also been shown that locations are perceived before colours (Pisella et al., 1998), which are perceived before orientations (Moutoussis and Zeki, 1997). The perceptual delay between colour and orientation (both first-order changes) makes it difficult to accept an alternative interpretation of our results (Nishida and Johnston, 2002), which suggests that the asynchrony is the result of comparisons between a first-order (colour) and a second-order (motion direction) change. We had assumed (Zeki and Bartels, 1999) that this asynchrony is due to differences in processing time between visual attributes and this assumption has been elegantly supported by recent experiments (Arnold and Clifford, 2002).

Because we become conscious of colour before we become conscious of motion, it follows that the micro-consciousnesses generated by activity at two distinct cortical sites are distributed in time as well. From this it follows that micro-consciousnesses are distributed in time and space, and that there is a temporal hierarchy of micro-consciousnesses, that for colour preceding that for motion. Of course, it is also true that over longer periods of time, in excess of 500 ms, we do see different attributes in perfect temporal and spatial registration (the attributes are 'bound' together). This raises questions that binding studies have so far not addressed, mainly whether one area 'waits' for the other to finish its processing, and whether a time buffer is part of the physiological mechanism for this waiting period.

Binding and macro-consciousnesses

The issue of binding between different attributes has not been systematically addressed to date by physiological studies; rather, the principal concern has been how activity of cells in a single area is bound to the activity of other cells in the same area to signal, for example, a continuous straight line (Singer, 2001). This has led to the conclusion that it is the binding itself that leads to the conscious experience (Crick and Koch, 1990; Engel et al., 1999). Whatever its merits, this proposal cannot be accepted as being necessarily true of binding the activity between two specialized areas. The most compelling evidence for doubting this lies in the anatomy, which shows that there are few, if any, direct connections between V4 and V5 in the monkey. Correspondingly, chronoarchitectonic maps of human cerebral cortex (Bartels and Zeki, 2004), generated when subjects view complex scenes, show that the time courses of activity in human V4 and V5 are significantly uncorrelated, from which we can infer that there are no direct anatomical links between them. On the other hand, anatomical evidence from monkey shows that V4 and V5 project in a juxta-convergent manner to further areas, in both the parietal and temporal areas (Shipp and Zeki, 1995). This raises two inter-related questions: (1) whether binding of activity between two specialized areas such as V4 and V5 involves further area(s), and (2) whether binding occurs during the processing stage or is post-conscious, occurring only after a conscious correlate is generated in each of the two areas.

One indication that binding may be post-conscious comes from our current psychophysical experiments that demonstrate that the binding of colour to motion occurs after the binding of colour to colour or motion to motion (Bartels and Zeki, 2006). Thus subjects become conscious of the bound percept *after* they become conscious of the attributes that are bound, again suggesting a temporal hierarchy in perception. I refer to consciousness of a stimulus or of a percept that is compound, in the sense that it consists of more than one attribute, as a 'macro-consciousness', to distinguish it from consciousness of a single attribute alone (e.g., colour). Consistent with a theory of micro-consciousnesses, it is interesting to note that a macro-consciousness may be the result of false binding, as when the veridically 'wrong' colour is bound to the 'right' motion or form (Moutoussis and Zeki, 1997). We have argued that

this results from the brain's binding what it has already processed (Moutoussis and Zeki, 2002).

A macro-consciousness need not, of course, be limited to a bound visual percept. It could equally signify consciousness of a percept that includes a visual and an auditory component, or of several visual components that, together, constitute a distinct new entity, for example, a moving red bus.

Three levels of hierarchies in consciousness

It thus becomes possible to distinguish three hierarchical levels of consciousness: the levels of micro-consciousness, of macro-consciousness, and of the unified consciousness. Of necessity, one level depends on the presence of the previous one. Within each level, one can postulate a temporal hierarchy. This has been demonstrated for the level of micro-consciousness, because colour and motion are perceived at different times. It has also been demonstrated for the level of the macro-consciousnesses, because binding between attributes takes longer than binding within attributes. This in turn leads one to postulate a set of temporal hierarchies, in which the binding of one set of attributes leading to a given macro-consciousness would take longer than the binding of another set of attributes leading to another macro-consciousness, and the binding of several attributes would take still longer. The experiment has not been conducted yet, but such a result seems likely.

Micro- and macro-consciousnesses, with their individual temporal hierarchies, lead to the final, unified consciousness, that of myself as the perceiving person. This and this alone qualifies as the unified consciousness, and this alone can be described in the singular. Kant probably saw, hesitatingly, the relation between the micro-consciousness (his 'empirical consciousness') and the unified consciousness. He wrote: 'All presentations have a necessary reference to a *possible* empirical consciousness. For if they did not have this reference, and becoming conscious of them were entirely impossible, then this would be tantamount to saying that they do not exist at all. But all empirical consciousness has a necessary reference to a transcendental consciousness (a consciousness that precedes all particular experience), viz., the consciousness of myself as original apperception' (original emphasis). Here, I disagree only with the suggestion that the 'empirical' (micro) consciousness has a *necessary* reference to the unified, transcendental consciousness.

Kant also suspected that the various attributes must themselves be synthesized first before being synthesized into the 'pure consciousness', although he could not have been aware of the principles of functional specialization. He continues: 'But because every appearance contains a manifold, so that different perceptions are in themselves encountered in the mind sporadically and individually, these perceptions need to be given a combination that in sense itself they cannot have. Hence there is in us an active power to synthesize this manifold' (which he calls 'imagination') (Kant, 1996).

Kant supposed that the 'transcendental' consciousness is present a priori, before any experience is acquired. It is hard to be conclusive in this regard, but it is worth pointing out that consciousness of oneself as the perceiving person amounts to being aware of being aware, and I believe that this requires communication with others and, especially, the use of language. The cortical programs to construct visual attributes must also be present before any experience is acquired and all experience must therefore be read into them. It seems more likely that, ontogenetically, the micro-consciousnesses precede the unified consciousness and that the programs for them are also present at birth. Hence, even though in adult life the unified consciousness sits at the apex of the hierarchy of consciousnesses, ontogenetically it is the micro-consciousnesses that occupy this position.

I believe that the search for the neural correlates of consciousness will be elusive until we acknowledge the many components of consciousness and their temporally hierarchical relationship to one another. The transition from the singular neural correlate of consciousness to the plural neural correlates of the consciousnesses is a small step on paper but may yet prove to be a very important one in understanding consciousness.

Acknowledgements

This work was supported by the Wellcome Trust, London. I am grateful to my colleagues, and especially Chris Frith, for commenting critically on earlier versions of the manuscript.

References

Arnold, D.H. and Clifford, C.W. (2002) Determinants of asynchronous processing in vision. Proc. R. Soc. Lond. B Biol. Sci., 269: 579–583.

Arnold, D.H., Clifford, C.W. and Wenderoth, P. (2001) Asynchronous processing in vision: color leads motion. Curr. Biol., 11: 596–600.

Barbur, J.L., Watson, J.D., Frackowiak, R.S. and Zeki, S. (1993) Conscious visual perception without V1. Brain, 116: 1293–1302.

Barbur, J.L., Wolf, J. and Lennie, P. (1998) Visual processing levels revealed by response latencies to changes in different visual attributes. Proc. R. Soc. Lond. B Biol. Sci., 265: 2321–2325.

Bartels, A. and Zeki, S. (2004) Functional brain mapping during free viewing of natural scenes. Hum. Brain Mapp., 21: 75–83.

Bartels, A. and Zeki, S. (2006) The temporal order of binding visual attributes. Vision Res., 46(14): 2280–2286. Epub Jan 4, 2006.

Beckers, G. and Zeki, S. (1995) The consequences of inactivating areas V1 and V5 on visual-motion perception. Brain, 118: 49–60.

Buchner, H., Gobbele, R., Wagner, M., Fuchs, M., Waberski, T.D. and Beckmann, R. (1997) Fast visual evoked potential input into human area V5. Neuroreport, 8: 2419–2422.

Cragg, B.G. (1969) The topography of the afferent projections in circumstriate visual cortex studied by the Nauta method. Vis. Res., 9: 733–747.

Crick, F. and Koch, C. (1990) Towards a neurobiological theory of consciousness. Semin. Neurosci., 2: 263–275.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D.L., Mangin, J.F., Poline, J.B. and Riviere, D. (2001) Cerebral mechanisms of word masking and unconscious repetition priming. Nat. Neurosci., 4: 752–758.

Engel, A.K., Fries, P., Konig, P., Brecht, M. and Singer, W. (1999) Temporal binding, binocular rivalry, and consciousness. Conscious. Cogn., 8: 155–158.

ffytche, D.H., Guy, C.N. and Zeki, S. (1995) The parallel visual motion inputs into areas V1 and V5 of human cerebral cortex. Brain, 118: 1375–1394.

Fries, W. (1981) The projection from the lateral geniculate nucleus to the prestriate cortex of the macaque monkey. Proc. R. Soc. Lond. B Biol. Sci., 213: 73–86.

Kant, I. (1996 [1781]). Kritik der reinen Vernunft, Transl. W.S. Pluhar as Critique of Pure Reason, Hackett, IN.

Lamme, V.A. (2001) Blindsight: the role of feedforward and feedback cortical connections. Acta Psychol., 107: 209–228.

Livingstone, M. and Hubel, D. (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science, 240: 740–749.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T. and Oeltermann, A. (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature, 412: 150–157.

Meadows, J.C. (1974) Disturbed perception of colours associated with localized cerebral lesions. Brain, 97: 615–632.

Moutoussis, K. and Zeki, S. (1997a) A direct demonstration of perceptual asynchrony in vision. Proc. R. Soc. Lond. B Biol. Sci., 264: 393–399.

Moutoussis, K. and Zeki, S. (1997b) Functional segregation and temporal hierarchy of the visual perceptive systems. Proc. R. Soc. Lond. B Biol. Sci., 264: 1407–1414.

Moutoussis, K. and Zeki, S. (2002) The relationship between cortical activation and perception investigated with invisible stimuli. Proc. Natl. Acad. Sci. U.S.A., 99: 9527–9532.

Nishida, S. and Johnston, A. (2002) Marker correspondence, not processing latency, determines temporal binding of visual attributes. Curr. Biol., 12: 359–368.

Pisella, L., Arzi, M. and Rossetti, Y. (1998) The timing of color and location processing in the motor context. Exp. Brain Res., 121: 270–276.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C. and Driver, J. (2000) Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. Brain, 123: 1624–1633.

Rees, G., Kreiman, G. and Koch, C. (2002) Neural correlates of consciousness in humans. Nat. Rev. Neurosci., 3: 261–270.

Shipp, S. and Zeki, S. (1995) Segregation and convergence of specialized pathways in macaque monkey visual cortex. J. Anat., 187: 547–562.

Singer, W. (2001) Consciousness and the binding problem. Ann. N.Y. Acad. Sci., 929: 123–146.

Standage, G.P. and Benevento, L.A. (1983) The organization of connections between the pulvinar and visual area MT in the macaque monkey. Brain Res., 262: 288–294.

Stoerig, P. and Barth, E. (2001) Low-level phenomenal vision despite unilateral destruction of primary visual cortex. Conscious. Cogn., 10: 574–587.

Stoerig, P. and Cowey, A. (1995) Visual-perception and phenomenal consciousness. Behav. Brain Res., 71: 147–156.

Wade, A.R., Brewer, A.A., Rieger, J.W. and Wandell, B.A. (2002) Functional measurements of human ventral occipital cortex: retinotopy and colour. Philos. Trans. R. Soc. Lond. B Biol. Sci., 357: 963–973.

Weiskrantz, L. (1990) The Ferrier Lecture, 1989. Outlooks for blindsight: explicit methodologies for implicit processes. Proc. R. Soc. B, 239: 247–278.

Weiskrantz, L., Barbur, J.L. and Sahraie, A. (1995) Parameters affecting conscious versus unconscious visual-discrimination

with damage to the visual-cortex (V1). Proc. Natl. Acad. Sci. U.S.A., 92: 6122–6126.
Yukie, M. and Iwai, E. (1981) Direct projection from the dorsal lateral geniculate nucleus to the prestriate cortex in macaque monkeys. J. Comp. Neurol., 201: 81–97.
Zeki, S. (1978) Functional specialization in the visual cortex of the monkey. Nature, 274: 423–428.
Zeki, S. (1990) A century of cerebral achromatopsia. Brain, 113: 1721–1777.
Zeki, S. (1991) Cerebral akinetopsia (visual motion blindness): a review. Brain, 114: 811–824.
Zeki, S. and Bartels, A. (1999) Towards a theory of visual consciousness. Conscious. Cogn., 8: 225–259.
Zeki, S. and ffytche, D. (1998) The Riddoch syndrome: insights into the neurobiology of conscious vision. Brain, 121: 25–45.
Zeki, S. and Moutoussis, K. (1997) Temporal hierarchy of the visual perceptive systems in the Mondrian world. Proc. R. Soc. Lond. B Biol. Sci., 264: 1415–1419.
Zeki, S., Watson, J.D., Lueck, C.J., Friston, K.J., Kennard, C. and Frackowiak, R.S. (1991) A direct demonstration of functional specialization in human visual cortex. J. Neurosci., 11: 641–649.
Zihl, J., von Cramon, D., Mai, N. and Schmid, C. (1991) Disturbance of movement vision after bilateral posterior brain damage: further evidence and follow up observations. Brain, 114: 2235–2252.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 3

Consciousness: the radical plasticity thesis

Axel Cleeremans*

Cognitive Science Research Unit, Université Libre de Bruxelles CP 191, 50 ave. F.-D. Roosevelt, B1050 Brussels, Belgium

Abstract: In this chapter, I sketch a conceptual framework which takes it as a starting point that conscious and unconscious cognition are rooted in the same set of interacting learning mechanisms and representational systems. On this view, the extent to which a representation is conscious depends in a graded manner on properties such as its stability in time or its strength. Crucially, these properties are accrued as a result of learning, which is in turn viewed as a mandatory process that always accompanies information processing. From this perspective, consciousness is best characterized as involving (1) a graded continuum defined over "quality of representation", such that availability to consciousness and to cognitive control correlates with quality, and (2) the implication of systems of metarepresentations. A first implication of these ideas is that the main function of consciousness is to make flexible, adaptive control over behavior possible. A second, much more speculative implication, is that we learn to be conscious. This I call the "radical plasticity thesis" — the hypothesis that consciousness emerges in systems capable not only of *learning* about their environment, but also about their own internal representations of it.

Keywords: consciousness; learning; subjective experience; neural networks; emotion

Information processing can undoubtedly take place without consciousness, as abundantly demonstrated by empirical evidence, but also by the very fact that extremely powerful information-processing machines, namely, computers, have now become ubiquitous. Only but a few would be willing to grant any quantum of conscious experience to contemporary computers, yet they are undeniably capable of sophisticated information processing — from recognizing faces to analyzing speech, from winning chess tournaments to helping prove theorems. Thus, consciousness is not information processing; experience is an "extra ingredient" (Chalmers, 2007b) that comes over and beyond mere computation.

With this premise in mind — a premise that just restates Chalmers' *hard problem*, that is, the question of *why* it is the case that information processing is accompanied by experience in humans and other higher animals — there are several ways in which one can think about the problem of consciousness.

One is to simply state, as per Dennett (1991, 2001) that there is nothing more to explain. Experience is *just* (a specific kind of) information processing in the brain; the contents of experience are *just* whatever representations have come to dominate processing at some point in time ("fame in the brain"); consciousness is *just* a harmless illusion. From this perspective, it is easy to imagine that machines will be conscious when they have

*Corresponding author. Tel.: +32 2 650 32 96; Fax: +32 2 650 22 09; E-mail: axcleer@ulb.ac.be

DOI: 10.1016/S0079-6123(07)68003-0

accrued sufficient complexity; the reason they are not conscious now is simply because they are not sophisticated enough. They lack the appropriate architecture perhaps, they lack sufficiently broad and diverse information processing abilities, and so on. Regardless of what is missing, the basic point here is that there is no reason to assume, *contra* Chalmers, that conscious experience is anything special. Instead, all that is required is one or several yet-to-be-identified functional mechanisms: recurrence, perhaps (Lamme, 2003), stability of representation (O'Brien and Opie, 1999), global availability (Baars, 1988; Dehaene et al., 1998), integration and differentiation of information (Tononi, 2003), or the involvement of higher order representations (Rosenthal, 1997), to name just a few.

Another take on this most difficult question is to consider that *experience* will never be amenable to a satisfactory functional explanation. Experience, according to some (e.g., Chalmers, 1996), is precisely what is left over once all functional aspects of consciousness have been explained. Notwithstanding the fact that so defined, experience is simply not something one can approach from a scientific point of view, this position recognizes that consciousness is a unique (a *hard*) problem in the Cognitive Neurosciences. But that is a different thing from saying that a reductive account is not possible. A non-reductive account, however, is exactly what Chalmer's Naturalistic Dualism attempts to offer, by proposing that information, as a matter of ontology, has a dual aspect — a physical aspect and a phenomenal aspect. "Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing" (Chalmers, 2007a, p. 366). This position leads him to defend the possibility that experience is a fundamental aspect of reality. Thus, even thermostats, for instance, may be endowed with very simple experiences, in virtue of the fact that they can toggle in two different states.

However, what do we mean when we speak of "subjective experience" or of "quale"? The simplest definition of these concepts (Nagel, 1974) goes right to the heart of the matter: "Experience" is *what it feels like* for a conscious organism to be that organism. There is something it is like for a bat to be a bat; there is nothing it is like for a stone to be a stone. As Chalmers (2007b) puts it: "When we see, for instance, we *experience* visual sensations: The felt quality of redness, the experience of dark and light, the quality of depth in a visual field" (p. 226).

Let us try to engage in some phenomenological analysis at this point to try to capture what it means for each of us to have an experience. Imagine you see a patch of red (Humphrey, 2006). You now have a *red* experience — something that a camera recording the same patch of red will most definitely *not* have. What is the difference between you and the camera? Tononi (2007), from whom I borrow this simple thought experiment, points out that one key difference is that when you see the patch of red, the state you find yourself in is but one among billions, whereas for a simple light-sensitive device, it is perhaps one of only two possible states — thus the state conveys a lot more *differentiated information* for you than for a light-sensitive diode. A further difference is that you are able to *integrate* the information conveyed by many different inputs, whereas the chip on a camera can be thought of as a mere array of independent sensors among which there is no interaction.

Hoping not to sound presumptuous, it strikes me, however, that both Chalmers' (somewhat paradoxically) and Tononi's analyses miss fundamental facts about experience; both analyze it as a rather abstract dimension or aspect of information, whereas experience — *what it feels like* — is anything but abstract. On the contrary, what we mean when we say that seeing a patch of red elicits an "experience" is that the seeing *does something to us* — in particular, we might feel one or several emotions, and we may associate the redness with memories of red. Perhaps seeing the patch of red makes you remember the color of the dress that your prom night date wore 20 years ago. Perhaps it evokes a vague anxiety, which we now know is also shared by monkeys (Humphrey, 1971). To a synesthete, perhaps seeing the color red will evoke the number 5. The point is that if conscious experience is what it feels like to be in a certain state, then "What it feels like" can only mean the

specific set of associations that have been established by experience between the stimulus or the situation you now find yourself in, on the one hand, and your memories, on the other. This is what one means by saying that there is something it is like to be you in this state rather than nobody or somebody else. The set of memories evoked by the stimulus (or by actions you perform, etc.), and, crucially, the set of emotional states associated with each of these memories. It is interesting to note that Indian philosophical traditions have placed similar emphasis on the role that emotion plays in shaping conscious experience (Banerjee, in press).

Thus, a first point about the very notion of subjective experience I would like to make here is that it is difficult to see what experience could mean beyond (1) the emotional value associated with a state of affairs, and (2) the vast, complex, richly structured, experience-dependent network of associations that the system has learned to associate with that state of affairs. ''What it feels like'' for me to see a patch of red at some point seems to be entirely exhausted by these two points. Granted, one could still imagine an agent that accesses specific memories, possibly associated with emotional value, upon seeing a patch of red and who fails to ''experience'' anything. But I surmise that this is mere simulation. One *could* design such a zombie agent, but any real agent that is driven by self-developed motivation, and that cannot help but be influenced by his emotional states will undoubtedly have experiences much like ours.

Hence, a first point about what we mean by ''experience'' is that there is nothing it is like for the camera to see the patch of red simply because it does not care: the stimulus is meaningless; the camera lacks even the most basic machinery that would make it possible to ascribe any interpretation to the patch of red; it is instead just a mere recording device for which nothing matters. There is nothing it is like to be that camera at that point in time simply because (1) the experience of different colors does not do anything to the camera; that is, colors are not associated with different emotional valences; and (2) the camera has no brain with which to register and process its own states. It is easy to imagine how this could be different. To hint at my forthcoming argument, a camera could, for instance, keep a record of the colors it is exposed to, and come to ''like'' some colors better than others. Over time, your camera would like different colors than mine, and it would also know that in some non-trivial sense. Appropriating one's mental contents for oneself is the beginning of individuation, and hence the beginning of a *self*.

Thus a second point about experience that I perceive as crucially important is that it does not make any sense to speak of experience without an *experiencer* who experiences the experiences. Experience is, almost by definition (''what it feels like''), something that takes place not in *any* physical entity but rather only in special physical entities, namely cognitive agents. Chalmers' (1996) thermostat fails to be conscious because, despite the fact that it can find itself in different internal states, it lacks the ability to remove itself from the causal chain in which it is embedded. In other words, it lacks knowledge *that* it can find itself in different states; it is but a mere object that responds to inputs in certain ways that one can fully describe by the laws of physics. While there is indeed something to be experienced there (the different states the thermostat can find itself in), there is no one home to be the *subject* of these experiences — the thermostat simply lacks the appropriate machinery to do so.

This point can also be illustrated by means of well-known results in the connectionist, or artificial neural network modeling literature. Consider for instance Hinton's (1986) famous demonstration that a simple back-propagation network can learn about abstract dimensions of the training set. Hinton's network was a relatively simple back-propagation network trained to process linguistic expressions consisting of an agent, a relationship, and a patient, such as for instance ''Maria is the wife of Roberto''. The stimulus material consisted of a series of such expressions, which together described some of the relationships that exist in the family trees of an Italian family and of an English family. The network was required to produce the patient of each agent-relationship pair it was given as input. For instance, the network should produce

"Roberto" when presented with "Maria" and "wife". Crucially, each person and each relationship were presented to the network by activating a single input unit. Hence there was no overlap whatsoever between the input representations of, say, Maria and Victoria. Yet, despite this complete absence of surface similarity between training exemplars, Hinton showed that after training, the network could, under certain conditions, develop internal representations that capture relevant abstract dimensions of the domain, such as nationality, sex, or age!

Hinton's point was to demonstrate that such networks were capable of learning richly structured internal representations as a result of merely being required to process exemplars of the domain. Crucially, the structure of the internal representations learned by the network is determined by the manner in which different exemplars interact with each other, that is, by their *functional similarity,* rather than by their mere *physical similarity* expressed, for instance, in terms of how many features (input units) they share. Hinton thus provided a striking demonstration of this important and often misunderstood aspect of associative learning procedures by showing that under some circumstances, specific hidden units of the network had come to act as detectors for dimensions of the material that had never been presented explicitly to the network. These results truly flesh out the notion that rich, abstract knowledge can simply emerge as a by-product of processing structured domains. It is interesting to note that the existence of such single-unit "detectors" has recently been shown to exist in human neocortex (Kreiman et al., 2002). Single-neuron recording of activity in hippocampus, for instance, has shown that some individual neurons exclusively respond to highly abstract entities, such as the words "Bill Clinton" and images of the American president.

Now, the point I want to make with this example is as follows: one could certainly describe the network as being *aware* of nationality, in the sense that it is sensitive to the concept. It exhibits differential responding (hence, behavioral sensitivity) to inputs that involve Italian agents vs. English agents. But, obviously, the network does not *know* anything about nationality. It does not even know that it has such and such representations of the inputs, nor does it know anything about its own, self-acquired sensitivity or awareness of the relevant dimensions. Instead, the rich, abstract, structured representations that the network has acquired over training forever remain embedded in a causal chain that begins with the input and ends with the network's responses. As Clark and Karmiloff-Smith (1993) insightfully pointed out, such representations are "first-order" representations to the extent that they are representations *in the system* rather than representations *for the system* that is, such representations are not accessible to the network *as representations*.

What would it take for a network like Hinton's to be able to access its own representations, and what difference would that make with respect to consciousness?

To answer the first question, the required machinery is the machinery of agenthood; in a nutshell, the ability to do something not just with external states of affairs, but rather with one's own representations of such external states. This crucially requires that the agent be able to access, inspect, and otherwise manipulate its own representations, and this in turn, I surmise, requires mechanisms that make it possible for an agent to redescribe its own representations to itself. The outcome of this continuous "representational redescription" (Karmiloff-Smith, 1992) process is that the agent ends up knowing something about the geography of its own internal states. It has, in effect, *learned* about its own representations. Minimally, this could be achieved rather simply, for instance by having another network take both the input (i.e., the external stimulus as represented proximally) to the first-order network and its internal representations of that stimulus as inputs themselves and do something with them.

One elementary thing the system consisting of the two interconnected networks (the first-order, observed network and the second-order, observing network) would now be able to do is to make decisions, for instance, about the extent to which an external input to the first-order network elicits a familiar pattern of activation over its hidden units or not. This would in turn enable the system to distinguish between hallucination and blindness

(see Lau, in press), or to come up with judgments about the performance of the first-order network (Persaud et al., 2007; Dienes, in press).

To address the second question (what difference would representational redescription make in terms of consciousness), if you think this is starting to sound like a higher order thought theory of consciousness (Rosenthal, 1997), you may be right. While I do not feel perfectly happy with all aspects of Higher-Order Thought Theory, I do believe, however, that higher order representations (I will call them metarepresentations in what follows) play a crucial role in consciousness.

An immediate objection to this idea is as follows: if there is nothing intrinsic to the existence of a representation in a cognitive system that makes this representation conscious, why should things be different for metarepresentations? After all, metarepresentations are representations also. Yes indeed, but with a crucial difference. Metarepresentations inform the agent about its own internal states, making it possible for it to develop an understanding of its own workings. And this, I argue, forms the basis for the contents of conscious experience, provided of course — which cannot be the case in an contemporary artificial system — that the system has learned about its representations by itself, over its development, and provided that it cares about what happens to it, that is, provided its behavior is rooted in emotion-laden motivation (to survive, to mate, to find food, etc.).

The radical plasticity thesis

I would thus like to defend the following claim: conscious experience occurs if and only if an information processing system has *learned* about its own representations of the world. To put this claim even more provocatively: consciousness is the brain's theory about itself, gained through experience interacting with the world, and, crucially, with itself. I call this claim the "*Radical Plasticity Thesis*", for its core is the notion that learning is what makes us conscious. How so? The short answer, as hinted above, is that consciousness involves not only knowledge about the world, but crucially, knowledge about our own internal states, or mental representations. When I claim to be conscious of a stimulus, I assert my ability to discriminate cases where the stimulus is present from cases where it is not. But what is the basis of this ability, given that I have no direct access to the stimulus? The answer is obvious: some neural states correlate with the presence or absence of the stimulus, and I make judgments about these states to come to a decision.

Note that this is the way in which *all* information processing takes place, with or without consciousness. After all, we *never* have direct access to anything that is part of the world in which we are embedded; any perception necessarily involves mediation through neural states, which in this sense are appropriately characterized as internal representations of external states of affairs.

What, then, differentiates cases where one is conscious of a state of affairs from cases where one remains unaware of it? It is obvious that in the first case, the relevant representations are accompanied by subjective experience whereas in the second, they are not.

This difference is in fact what motivates Baars' "contrastive approach", through which one seeks to identify differences between information processing with and without consciousness by "treating consciousness as a variable", that is, by designing experimental paradigms in which the only difference of interest is one of conscious awareness. The same idea underpins what neuroscientists call the "search for the neural correlates of consciousness" (Frith et al., 1999). Here, the goal is to identify cerebral regions, neural processes, or processing pathways where one finds activity that correlates not with some objective state of affairs (i.e., a stimulus), but rather with people's own subjective reports that they are conscious of that state of affairs.

As Lau (this volume) points out, however, things are not so simple, for this approach rests on the premise that one can indeed design an experimental situation in which consciousness is the *only* difference. This, as it turns out, is extremely difficult to achieve, precisely because consciousness does make a difference! In other

words, performance at a given task will also be different depending on whether the subject is conscious or not of the relevant state of affairs.

In the following, I would now like to present a framework through which to characterize the relationships between learning and consciousness. If the main cognitive function of consciousness is to make adaptive control of behavior possible, as is commonly accepted, then consciousness is necessarily closely related to processes of learning, because one of the central consequences of successful adaptation is that conscious control is no longer required over the corresponding behavior. Indeed, it might seem particularly adaptive for complex organisms to be capable of behavior that does not require conscious control, for instance because behavior that does not require monitoring of any kind can be executed faster or more efficiently than behavior that does require such control. What about conscious experience? Congruently with our intuitions about the role of consciousness in learning, we often say of somebody who failed miserably at some challenging endeavor, such as completing a paper by the deadline, that the failure constitutes ''a learning experience''. What precisely do we mean by this? We mean that the person can now learn from her mistakes, that the experience of failure was sufficiently imbued with emotional value that it has registered in that person's brain. The experience *hurt*, it made one realize what was at stake, it made us think about it, in other words, it made us consciously aware of what failed and why.

But this minimally requires what Kirsh (1991) has called ''explicit representation'', namely the presence of representations that directly represent the relevant information. By ''direct'' here, I mean that the information is represented in such a manner that no further computation is required to gain access to it. For instance, a representation that is explicit in this sense might simply consist of a population of neurons that fire whenever a specific condition holds: a particular stimulus is present on the screen, my body is in a particular state (i.e., pain or hunger).

By assumption, such ''explicit'' representations are not necessarily conscious. Instead, they are merely good candidates to enter conscious awareness in virtue of features such as their stability, strength, or distinctiveness (Cleeremans, 2005, 2006). What is missing, then? What is missing is that such representations be themselves the target of other representations. And how would this make any difference? It makes a crucial difference, for the relevant first-order *representations* are now part of the agent's repertoire of mental states; such representations are then, and only then, recognized as playing the function of representing some other (external) state of affairs.

A learning-based account of consciousness

I would now like to introduce the set of assumptions that together form the core of the framework (see Cleeremans and Jiménez, 2002; Cleeremans, in preparation, for more detailed accounts). It is important to keep it in mind that the framework is based on the connectionist framework (e.g., Rumelhart and McClelland, 1986). It is therefore based on many central ideas that characterize the connectionist approach, such as the fact that information processing is graded and continuous, and that it takes place over many interconnected modules consisting of processing units. In such systems, long-term knowledge is embodied in the pattern of connectivity between the processing units of each module and between the modules themselves, while the transient patterns of activation over the units of each module capture the temporary results of information processing.

This being said, a first important assumption is that *representations are graded, dynamic, active, and constantly causally efficacious* (Cleeremans, 1994). Patterns of activation in neural networks and in the brain are typically distributed and can therefore vary on a number of dimensions, such as their stability in time, their strength, or their distinctiveness. *Stability* in time refers to how long a representation can be maintained active during processing. There are many indications that different neural systems involve representations that differ along this dimension. For instance, prefrontal cortex, which plays a central role in working memory, is widely assumed to involve

circuits specialized in the formation of the enduring representations needed for the active maintenance of task-relevant information. *Strength* of representation simply refers to how many processing units are involved in the representation, and to how strongly activated these units are. As a rule, strong activation patterns will exert more influence on ongoing processing than weak patterns. Finally, *distinctiveness* of representation is inversely related to the extent of overlap that exists between representations of similar instances. Distinctiveness has been hypothesized as the main dimension through which cortical and hippocampal representations differ (McClelland et al., 1995; O'Reilly and Munakata, 2000), with the latter becoming active only when the specific conjunctions of features that they code for are active themselves.

In the following, I will collectively refer to these different dimensions as "quality of representation" (see also Farah, 1994). The most important notion that underpins these different dimensions is that representations, in contrast to the all-or-none propositional representations typically used in classical theories, instead have a *graded* character that enables any particular representation to convey the extent to which what it refers to is indeed present.

Another important aspect of this characterization of representational systems in the brain is that, far from being static propositions waiting to be accessed by some process, representations instead continuously influence processing regardless of their quality. This assumption takes its roots in McClelland's (1979) analysis of cascaded processing, which by showing how modules interacting with each other need not "wait" for other modules to have completed their processing before starting their own, demonstrated how stage-like performance could emerge out of such continuous, non-linear systems. Thus, even weak, poor-quality traces are capable of influencing processing, for instance through associative priming mechanisms, that is, in *conjunction* with other sources of stimulation. Strong, high-quality traces, in contrast have *generative capacity*, in the sense that they can influence performance independently of the influence of other constraints, that is, whenever their preferred stimulus is present.

A second important assumption is that *learning is a mandatory consequence of information processing*. Indeed, every form of neural information processing produces adaptive changes in the connectivity of the system, through mechanisms such as long-term potentiation (LTP) or long-term depression (LTD) in neural systems, or hebbian learning in connectionist systems. An important aspect of these mechanisms is that they are mandatory in the sense that they take place whenever the sending and receiving units or processing modules are co-active. O'Reilly and Munakata (2000) have described hebbian learning as instantiating what they call *model learning*. The fundamental computational objective of such unsupervised learning mechanisms is to enable the cognitive system to develop useful, informative models of the world by capturing its correlational structure. As such, they stand in contrast with *task learning* mechanisms, which instantiate the different computational objective of mastering specific input–output mappings (i.e., achieving specific goals) in the context of specific tasks through error-correcting learning procedures.

Having put in place assumptions about representations and learning, the central ideas that I would now like to explore are (1) that the extent to which a particular representation is available to consciousness depends on its quality, (2) that learning produces, over time, higher quality (and therefore adapted) representations, and (3) that the function of consciousness is to offer necessary control over those representations that are strong enough to influence behavior, yet not sufficiently adapted that their influence does not require control anymore.

Figure 1 aims to capture these ideas by representing the relationships between quality of representation (X-axis) on the one hand and (1) potency, or the extent to which a representation can influence behavior, (2) availability to control, (3) availability to subjective experience. I discuss the figure at length in the following section. Let us simply note here that the X-axis represents a continuum between weak, poor-quality representations on the left

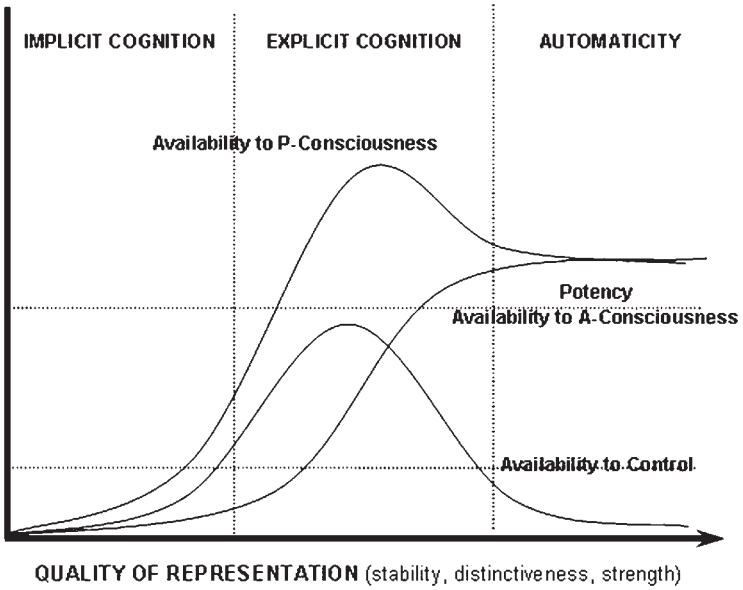


Fig. 1. Graphical representation of the relationships between quality of representation (X-axis) and (1) potency, (2) availability to control, (3) availability to subjective experience. See text for further details.

and very strong, high-quality representations on the right.

Two further points are important to be kept in mind with respect to Fig. 1. First, the relationships depicted in the figure are intended to represent *availability* to some dimension of behavior or consciousness independently of other considerations. Many potentially important modulatory influences on the state of any particular module are thus simply not meant to be captured neither by Fig. 1, nor by the framework presented here. Second, the figure is intended to represent what happens in *each* of the many processing modules involved in any particular cognitive task. Thus, at any point in time, there will be many such modules active, each contributing to some extent to behavior and to conscious experience; each modulating the activity of other modules. With these caveats in mind, let me now turn to four assumptions about consciousness and its relationship with learning:

Assumption C1: *Consciousness involves two dimensions: subjective experience and control*

As argued by Block (1995, 2005) and even though there is continuing debate about this issue, consciousness involves at least two separable aspects, namely access consciousness (A-consciousness) and phenomenal consciousness (P-consciousness). According to Block (1995), "A perceptual state is access-conscious roughly speaking if its content — what is represented by the perceptual state — is processed via that information processing function, that is, if its content gets to the Executive system, whereby it can be used to control reasoning and behavior" (p. 234). In other words, whether a state is A-conscious is defined essentially by the causal efficacy of that state; the extent to which it is available for global control of action. Control refers to the ability of an agent to control, to modulate, and to inhibit the influence of particular representations on processing. In this framework, control is simply a function of potency, as described in assumption C3. In contrast, P-consciousness refers to the phenomenal aspects of subjective experience discussed in the introduction: a state is P-conscious to the extent that there is something it is like to be in that state: I am currently experiencing a pain, hearing a beautiful piece of music, entertaining the memory of a joyful event. While the extent to which potency (i.e., availability to access consciousness) and

control on the one hand, and subjective experience (i.e., availability to phenomenal consciousness) on the other, are dissociable is debatable, the framework suggests that potency, control, and phenomenal experience are closely related to each other.

Assumption C2: *Availability to consciousness correlates with quality of representation*

This assumption is also a central one in this framework. It states that explicit, conscious knowledge involves higher quality memory traces than implicit knowledge. ''Quality of representation'' designates several properties of memory traces, such as their relative strength in the relevant information-processing pathways, their distinctiveness, or their stability in time. The assumption is consistent with the theoretical positions expressed by several different authors over the last few years. O'Brien and Opie (1999) have perhaps been the most direct in endorsing a characterization of phenomenal consciousness in terms of the properties of mental representations in defending the idea that ''consciousness equals stability of representation'', that is, that the particular mental contents that one is aware of at some point in time correspond to those representations that are sufficiently stable in time. Mathis and Mozer (1996) have also suggested that consciousness involves stable representations, specifically by offering a computational model of priming phenomena in which stability literally corresponds to the state that a dynamic ''attractor'' network reaches when the activations of a subset of its units stops changing and settle into a stable, unchanging state.

A slightly different perspective on the notion of ''quality of representation'' is offered by authors who emphasize not stability, but strength of representation as the important feature by which to characterize availability to consciousness. One finds echoes of this position in the writings of Kinsbourne (1997), for whom availability to consciousness depends on properties of representations such as duration, activation, or congruence.

In Fig. 1, I have represented the extent to which a given representation is available to the different components of consciousness (phenomenal consciousness, access-consciousness/potency, and control) as functions of a single underlying dimension expressed in terms of the quality of this representation. Availability to access-consciousness is represented by the curve labeled ''potency'', which expresses the extent to which representations can influence behavior as a function of their quality: high-quality, strong, distinctive representations, by definition, are more potent than weaker representations and hence more likely to influence behavior. ''Availability to control processes'' is represented by a second curve, so labeled. We simply assume that both weak and very strong representations are difficult to control, and that maximal control can be achieved on representations that are strong enough that they can begin to influence behavior in significant ways, yet not so strong that have become utterly dominant in processing. Finally, availability to phenomenal experience is represented by the third curve, obtained simply by adding the other two. The underlying intuition, discussed in the context of assumption C4, is that which contents enter subjective experience is a function of both availability to control and of potency.

Assumption C3: *Developing high-quality representations takes time*

This assumption states that the emergence of high quality representations in a given processing module takes time, both over training or development, as well as during processing of a single event. Figure 1 can thus be viewed as representing not only the relationships between quality of representation and their availability to the different components of consciousness, but also as a depiction of the dynamics of how a particular representation will change over the different time scales corresponding to development, learning, or within-trial processing (see Destrebecqz and Cleeremans, 2001, 2003; Destrebecqz et al., 2005, for further developments of this specific idea; Cleeremans and Sarrazin, 2007).

Both skill acquisition and development, for instance, involve the long-term progressive emergence of high-quality, strong memory traces based on early availability of weaker traces. Likewise, the extent to which memory traces can influence

performance at any moment (e.g., during a single trial) depends both on available processing time, as well as on overall trace strength. These processes of change operate on the connection weights between units, and can involve either task-dependent, error-correcting procedures, or unsupervised procedures such as hebbian learning. In either case, continued exposure to exemplars of the domain will result in the development of increasingly congruent and strong internal representations that capture more and more of the relevant variance. Although I think of this process as essentially continuous, three stages in the formation of such internal representations (each depicted as separate regions in Fig. 1) can be distinguished: implicit representations, explicit representations, and automatic representations.

The first region, labeled "Implicit Cognition" in Fig. 1, is meant to correspond to the point at which processing starts in the context of a single trial, or to some early stage of development or skill acquisition. In either case, this stage is characterized by weak, poor-quality representations. A first important point is that representations at this stage are already capable of influencing performance, as long as they can be brought to bear on processing together with other sources of constraints, that is, essentially through mechanisms of associative priming and constraint satisfaction. A second important point is that this influence is best described as "implicit", because the relevant representations are too weak (i.e., not distinctive enough) for the system as a whole to be capable of exerting control over them: you cannot control what you cannot identify as distinct from something else.

The second region of Fig. 1 corresponds to the emergence of explicit representations, defined as representations over which one can exert control. In the terminology of attractor networks, this would correspond to a stage during learning at which attractors become better defined — deeper, wider, and more distinctive, so corresponding to the best "constraint-satisfaction" interpretation of a state of affairs (Maia and Cleeremans, 2005). It is also at this point that the relevant representations acquire generative capacity, in the sense that they now have accrued sufficient strength to have the potential to determine appropriate responses when their preferred stimulus is presented to the system alone. Such representations are also good candidates for redescription and can thus be recoded in various different ways, for instance, as linguistic propositions.

The third region involves what I call automatic representations, that is, representations that have become so strong that their influence on behavior can no longer be controlled (i.e., inhibited). Such representations exert a mandatory influence on processing. Importantly, however, one is aware both of possessing them (i.e., one has relevant metaknowledge) and of their influence on processing (see also Tzelgov, 1997), because availability to conscious awareness depends on the quality of internal representations, and that strong representations are of high quality. In this perspective then, one can always be conscious of automatic behavior, but not necessarily with the possibility of control over these behaviors.

In this framework, skill acquisition and development therefore involve a continuum at both ends of which control over representations is impossible or difficult, but for very different reasons: implicit representations influence performance but cannot be controlled because they are not yet sufficiently distinctive and strong for the system to even know it possesses them. This might in turn be related to the fact that, precisely because of their weakness, implicit representations cannot influence behavior on their own, but only in conjunction with other sources of constraints. Automatic representations, on the other hand, cannot be controlled because they are too strong, but the system is aware both of their presence and of their influence on performance.

Assumption C4: *The function of consciousness is to offer flexible, adaptive control over behavior*

The framework gives consciousness a central place in information processing, in the sense that its function is to enable flexible control over behavior. Crucially, however, consciousness is not necessary for information processing, or for adaptation in general, thus giving a place for implicit learning in cognition. I believe this

perspective to be congruent with theories of adaptation and optimality in general.

Indeed, another way to think about the role of learning in consciousness is to ask: "When does one need control over behavior?" Control is perhaps not necessary for implicit representations, for their influence on behavior is necessarily weak (in virtue of the fact that precisely because they are weak, such representations are unlikely to be detrimental to the organism even if they are not particularly well-adapted). Likewise, control is not necessary for automatic representations, because presumably, those representations that have become automatic after extensive training should be adapted (optimal) as long as the processes of learning that have produced them can themselves be assumed to be adaptive. Automatic behavior is thus necessarily optimal behavior in this framework, except, precisely, in cases such as addiction, obsessive–compulsive behavior, or laboratory situations where the automatic response is manipulated to be non-optimal, such as in the Stroop situation. Referring again to Fig. 1, my analysis therefore suggests that the representations that require the most control are the explicit representations that correspond to the central region of Fig. 1: representations that are strong enough that they have the potential to influence behavior in and of themselves (and hence that one should really care about, in contrast to implicit representations), but not sufficiently strong that they can be assumed to be already adapted, as is the case for automatic representations. It is for those representations that control is needed, and, for this reason, it is these representations that one is most aware of.

Likewise, this analysis also predicts that the dominant contents of subjective experience at any point in time consist precisely of those representations that are both strong enough that they can influence behavior, yet weak enough that they still require control. Figure 1 reflects these ideas by suggesting that the contents of phenomenal experience depend both on the potency of currently active representations as well as on their availability to control. Since availability to control is inversely related to potency for representations associated with automatic behavior, this indeed predicts weaker availability to phenomenal experience of "very strong" representations as compared to "merely strong" representations. In other words, such representations can become conscious if appropriate attention is directed towards their contents — as in cases where normally automatic behavior (such as walking) suddenly becomes conscious because the normal unfolding of the behavior has been interrupted (e.g., because I've stumbled upon something) — but they are not normally part of the central focus of awareness nor do they require cognitive control. It is interesting to note that these ideas are roughly consistent with Jackendoff's (1987) and Prinz's (2007) "Intermediate Level Theory of Consciousness".

The framework thus leaves open four distinct possibilities for knowledge to be implicit. First, knowledge that is embedded in the connection weights within and between processing modules can never be directly available to conscious awareness and control. This is simply a consequence of the fact that I assume that consciousness necessarily involves representations (patterns of activation over processing units). The knowledge embedded in connection weights will, however, shape the representations that depend on it, and its effects will therefore be detectable — but only indirectly, and only to the extent that these effects are sufficiently marked in the corresponding representations.

Second, to enter conscious awareness, a representation needs to be of sufficiently high quality in terms of strength, stability in time, or distinctiveness. Weak representations are therefore poor candidates to enter conscious awareness. This, however, does not necessarily imply that they remain causally inert, for they can influence further processing in other modules, even if only weakly so.

Third, a representation can be strong enough to enter conscious awareness, but fail to be associated with relevant metarepresentations. There are thus many opportunities for a particular conscious content to remain, in a way, implicit, not because its representational vehicle does not have the appropriate properties, but because it fails to be integrated with other conscious contents. Dienes and Perner (2003) offer an insightful analysis of

the different ways in which what I have called high-quality representations can remain implicit. Likewise, phenomena such as inattentional blindness (Mack and Rock, 1998) or blindsight (Weiskrantz, 1986) also suggest that high-quality representations can nevertheless fail to reach consciousness, not because of their inherent properties, but because they fail to be attended to or because of functional disconnection with other modules (see Dehaene et al., 2006).

Finally, a representation can be so strong that its influence can be no longer be controlled. In such cases, it is debatable whether the knowledge should be taken as genuinely unconscious, because it can certainly become fully conscious as long as appropriate attention is directed to it, but the point is that such very strong representations can trigger and support behavior without conscious intention and without the need for conscious monitoring of the unfolding behavior.

Metarepresentation

Strong, stable, and distinctive representations are thus *explicit* representations, at least in the sense put forward by Koch (2004): they indicate what they stand for in such a manner that their reference can be retrieved directly through processes involving low computational complexity (see also Kirsh, 1991, 2003). Conscious representations, in this sense, are explicit representations that have come to play, through processes of learning, adaptation, and evolution, the functional role of denoting a particular content for a cognitive system. Importantly, quality of representation should be viewed as a *graded* dimension.

Once a representation has accrued sufficient strength, stability, and distinctiveness, it may be the target of metarepresentations: the system may then ''realize'', if it is so capable, that is, if it is equipped with the mechanisms that are necessary to support self-inspection, that it has learned a novel partition of the input; that it now possesses a new ''detector'' that only fires when a particular kind of stimulus, or a particular condition, is present. Humphrey (2006) emphasizes the same point when he states that ''This self-monitoring by the subject of his own response is the prototype of the 'feeling sensation' as we humans know it'' (p. 90). Importantly, my claim here is that such metarepresentations are learned in just the same way as first-order representations, that is, by virtue of continuously operating learning mechanisms. Because metarepresentations are also representations, the same principles of stability, strength, and distinctiveness therefore apply. An important implication of this observation is that activation of metarepresentations can become automatic, just as it is the case for first-order representations.

What might be the function of such metarepresentations? One intriguing possibility is that their function is to indicate the mental attitude through which a first-order representation is held: is this something I know, hope, fear, or regret? Possessing such metaknowledge about one's knowledge has obvious adaptive advantages, not only with respect to the agent himself, but also because of the important role that communicating such mental attitudes to others plays in both competitive and cooperative social environments.

However, there is another important function that metarepresentations may play: they can also be used to anticipate the future occurrences of first-order representations. Thus for instance, if my brain learns that SMA is systematically active before M1, then it can use SMA representations to explicitly represent their consequences downstream, that is, M1 activation, and ultimately, action. If neurons in SMA systematically become active before an action is carried out, a metarepresentation can link the two and represent this fact explicitly in a manner that will be experienced as intention. That is, when neurons in the SMA become active, I experience the feeling of intention *because* my brain has learned, unconsciously, that such activity in SMA precedes action. It is this knowledge that gives qualitative character to experience, for, as a result of learning, each stimulus that I see, hear, feel, or smell is now not only represented, but also re-represented through metarepresentations that enrich and augment the original representation(s) with knowledge about (1) how similar the manner in which the stimulus'

representation is with respect to that associated with other stimuli, (2) how similar the stimulus' representation is now with respect to what it was before, (3) how consistent is a stimulus' representation with what it typically is, (4) what other regions of my brain are active at the same time that the stimulus' representation is, etc. This perspective is akin to the sensorimotor perspective (O'Regan and Noë, 2001) in the sense that awareness is linked with knowledge of the consequences of our actions, but, crucially, the argument is extended to the entire domain of neural representations.

Conclusion

Thus we end with the following idea, which is the heart of the ''radical plasticity thesis'': the brain continuously and unconsciously learns not only about the external world, but about its own representations of it. The result of this unconscious learning is conscious experience, in virtue of the fact that each representational state is now accompanied by (unconscious learnt) metarepresentations that convey the mental attitude with which these first-order representations are held. From this perspective thus, there is nothing intrinsic to neural activity, or to information per se, that makes it conscious. Conscious experience involves specific mechanisms through which particular (i.e., stable, strong, and distinctive) unconscious neural states become the target of further processing, which I surmise involves some form of representational redescription in the sense described by Karmiloff-Smith (1992). These ideas are congruent both with higher order theories in general (Rosenthal, 1997; Dienes and Perner, 1999; Dienes, in press), but also with those of Lau (in press) who characterizes consciousness as ''signal detection on the mind''. Finally, one dimension that I feel is sorely missing from contemporary discussion of consciousness is emotion (see Damasio, 1999; LeDoux, 2002; Tsuchiya and Adolphs, 2007). Conscious experience would not exist without experiencers who *care* about their experiences!

Acknowledgments

A.C. is a Research Director with the National Fund for Scientific Research (FNRS, Belgium). This work was supported by an institutional grant from the Université Libre de Bruxelles to A.C. and by Concerted Research Action 06/11-342 titled ''Culturally modified organisms: What it means to be human in the age of culture'', financed by the Ministère de la Communauté Française — Direction Générale l'Enseignement non obligatoire et de la Recherche scientifique (Belgium). Substantial portions of this article were adapted from the following publication: Cleeremans (2006). I would like to thank the organizers of the *Models of Brain and Mind: Physical, Computational and Psychological Approaches* workshop, and Rahul Banerjee in particular, for inviting me to contribute this piece.

References

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

Banerjee, R. (in press). Buddha and the bridging relations. Prog. Brain Res.

Block, N. (1995) On a confusion about a function of consciousness. Behav. Brain Sci., 18: 227–287.

Block, N. (2005) Two neural correlates of consciousness. Trends Cogn. Sci., 9(2): 46–52.

O'Brien, G. and Opie, J. (1999) A connectionist theory of phenomenal experience. Behav. Brain Sci., 22: 175–196.

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

Chalmers, D.J. (2007a) Naturalistic dualism. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 359–368.

Chalmers, D.J. (2007b) The hard problem of consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 225–235.

Clark, A. and Karmiloff-Smith, A. (1993) The cognizer's innards: a psychological and philosophical perspective on the development of thought. Mind Lang., 8: 487–519.

Cleeremans, A. (1994) Awareness and abstraction are graded dimensions. Behav. Brain Sci., 17: 402–403.

Cleeremans, A. (2005) Computational correlates of consciousness. In: Laureys S. (Ed.), Progress in Brain Research, Vol. 150. Elsevier, Amsterdam, pp. 81–98.

Cleeremans, A. (2006) Conscious and unconscious cognition: a graded, dynamic perspective. In: Jing Q., Rosenzweig M.R., d'Ydewalle G., Zhang H., Chen H.-C. and Zhang C. (Eds.), Progress in Psychological Science around the World. Vol. 1:

Neural, Cognitive, and Developmental IssuesPsychology Press, Hove, UK, pp. 401–418.
Cleeremans, A. (in preparation) Being Virtual. Oxford University Press, Oxford, UK.
Cleeremans, A. and Jiménez, L. (2002) Implicit learning and consciousness: a graded, dynamic perspective. In: French R.M. and Cleeremans A. (Eds.), Implicit Learning and Consciousness: An Empirical, Computational and Philosophical Consensus in the Making? Psychology Press, Hove, UK, pp. 1–40.
Cleeremans, A. and Sarrazin, J.-C. (2007) Time, action, and consciousness. Hum. Mov. Sci., 26(2): 180–202.
Damasio, A. (1999) The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace & Company, New York, NY.
Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. and Sergent, C. (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci., 10(5): 204–211.
Dehaene, S., Kerszberg, M. and Changeux, J.-P. (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. U.S.A., 95(24): 14529–14534.
Dennett, D.C. (1991) Consciousness Explained. Little, Brown & Co., Boston, MA.
Dennett, D.C. (2001) Are we explaining consciousness yet? Cognition, 79: 221–237.
Destrebecqz, A. and Cleeremans, A. (2001) Can sequence learning be implicit? New evidence with the Process Dissociation Procedure. Psychon. Bull. Rev., 8(2): 343–350.
Destrebecqz, A. and Cleeremans, A. (2003) Temporal effects in sequence learning. In: Jiménez L. (Ed.), Attention and Implicit Learning. John Benjamins, Amsterdam, pp. 181–213.
Destrebecqz, A., Peigneux, P., Laureys, S., Degueldre, C., Del Fiore, G. Aerts, J. et al. (2005) The neural correlates of implicit and explicit sequence learning: interacting networks revealed by the process dissociation procedure. Learn. Mem., 12(5): 480–490.
Dienes, Z. (in press). Subjective measures of unconscious knowledge. Prog. Brain Res.
Dienes, Z. and Perner, J. (1999) A theory of implicit and explicit knowledge. Behav. Brain Sci., 22: 735–808.
Dienes, Z. and Perner, J. (2003) Unifying consciousness with explicit knowledge. In: Cleeremans A. (Ed.), The Unity of Consciousness: Binding, Integration, and Dissociation. Oxford University Press, Oxford, UK, pp. 214–232.
Farah, M.J. (1994) Visual perception and visual awareness after brain damage: a tutorial overview. In: Umiltà C. and Moscovitch M. (Eds.), Attention and Performance XV: Conscious and Nonconscious Information Processing. MIT Press, Cambridge, MA, pp. 37–76.
Frith, C.D., Perry, R. and Lumer, E. (1999) The neural correlates of conscious experience: an experimental framework. Trends Cogn. Sci., 3: 105–114.
Hinton, G.E. (1986) Learning distributed representations of concepts. In: Proceedings of the Eighth Annual Conference of the Cognitive Science Society (Amherst, MA), Hillsdale, Erlbaum, pp. 1–12.
Humphrey, N. (1971) Colour and brightness preferences in monkeys. Nature, 229: 615–617.
Humphrey, N. (2006) Seeing Red. Harvard University Press, Cambridge, MA.
Jackendoff, R. (1987) Consciousness and the Computational Mind. MIT Press, Cambridge, MA.
Karmiloff-Smith, A. (1992) Beyond Modularity: A Developmental Perspective on Cognitive Science. MIT Press, Cambridge.
Kinsbourne, M. (1997) What qualifies a representation for a role in consciousness? In: Cohen J.D. and Schooler J.W. (Eds.), Scientific Approaches to Consciousness. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 335–355.
Kirsh, D. (1991) When is information explicitly represented? In: Hanson P.P. (Ed.), Information, Language, and Cognition. Oxford University Press, New York, NY.
Kirsh, D. (2003) Implicit and explicit representation. In: Nadel L. (Ed.), Encyclopedia of Cognitive Science, Vol. 2. Macmillan, London, UK, pp. 478–481.
Koch, C. (2004) The Quest for Consciousness. A Neurobiological Approach. Roberts & Company Publishers, Englewood, CO.
Kreiman, G., Fried, I. and Koch, C. (2002) Single-neuron correlates of subjective vision in the human medial temporal lobe. Proc. Natl. Acad. Sci. U.S.A., 99: 8378–8383.
Lamme, V.A.F. (2003) Why visual attention and awareness are different? Trends Cogn. Sci., 7(1): 12–18.
Lau, H. (in press). A higher-order Bayesian Decision Theory of consciousness. Prog. Brain Res.
LeDoux, J. (2002) Synaptic Self. Viking Penguin, Harmondsworth, UK.
Mack, A. and Rock, I. (1998) Inattentional Blindness. MIT Press, Cambridge, MA.
Maia, T.V. and Cleeremans, A. (2005) Consciousness: converging insights from connectionist modeling and neuroscience. Trends Cogn. Sci., 9(8): 397–404.
Mathis, W.D. and Mozer, M.C. (1996) Conscious and unconscious perception: a computational theory. In: Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 324–328.
McClelland, J.L. (1979) On the time-relations of mental processes: an examination of systems in cascade. Psychol. Rev., 86: 287–330.
McClelland, J.L., McNaughton, B.L. and O'Reilly, R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev., 102: 419–457.
Nagel, T. (1974) What is like to be a bat? Philos. Rev., 83: 434–450.
Persaud, N., McLeod, P. and Cowey, A. (2007) Post-decision wagering objectively measures awareness. Nat. Neurosci., 10: 257–261.
Prinz, J.J. (2007) The intermediate level theory of consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Oxford University Press, Oxford, UK, pp. 248–260.

O'Regan, J.K. and Noë, A. (2001) What it is like to see: a sensorimotor theory of visual experience. Synthèse, 129(1): 79–103.

O'Reilly, R.C. and Munakata, Y. (2000) Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain. MIT Press, Cambridge, MA.

Rosenthal, D. (1997) A theory of consciousness. In: Block N., Flanagan O. and Güzeldere G. (Eds.), The Nature of Consciousness: Philosophical Debates. MIT Press, Cambridge, MA.

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

Tononi, G. (2003) Consciousness differentiated and integrated. In: Cleeremans A. (Ed.), The Unity of Consciousness: Binding, Integration, and Dissociation. Oxford University Press, Oxford, UK, pp. 253–265.

Tononi, G. (2007) The information integration theory. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 287–299.

Tsuchiya, N. and Adolphs, R. (2007) Emotion and consciousness. Trends Cogn. Sci., 11(4): 158–167.

Tzelgov, J. (1997) Automatic but conscious: that is how we act most of the time. In: Wyer R.S. (Ed.), The Automaticity of Everyday Life, Vol. X. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 217–230.

Weiskrantz, L. (1986) Blindsight: A Case Study and Implications. Oxford University Press, Oxford, England.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 4

A higher order Bayesian decision theory of consciousness

Hakwan C. Lau[1,2,*]

[1]*Wellcome Trust Functional Imaging Laboratory, University College London, 12 Queen Square, London WC1N 3BG, UK*
[2]*Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, UK*

Abstract: It is usually taken as given that consciousness involves superior or more elaborate forms of information processing. Contemporary models equate consciousness with global processing, system complexity, or depth or stability of computation. This is in stark contrast with the powerful philosophical intuition that being conscious is more than just having the ability to compute. I argue that it is also incompatible with current empirical findings. I present a model that is free from the strong assumption that consciousness predicts superior performance. The model is based on Bayesian decision theory, of which signal detection theory is a special case. It reflects the fact that the capacity for perceptual decisions is fundamentally limited by the presence and amount of noise in the system. To optimize performance, one therefore needs to set decision criteria that are based on the behaviour, i.e. the probability distributions, of the internal signals. One important realization is that the knowledge of how our internal signals behave statistically has to be learned over time. Essentially, we are doing statistics on our own brain. This 'higher-order' learning, however, may err, and this impairs our ability to set and maintain optimal criteria for perceptual decisions, which I argue is central to perception consciousness. I outline three possibilities of how conscious perception might be affected by failures of 'higher-order' representation. These all imply that one can have a dissociation between consciousness and performance. This model readily explains blindsight and hallucinations in formal terms, and is beginning to receive direct empirical support. I end by discussing some philosophical implications of the model.

Keywords: consciousness; Bayesian; signal detection; fMRI

Introduction

This article describes a theoretical framework for characterizing perceptual consciousness. People receive information from the outside world through their sense organs, and produce actions in reaction to the external stimuli. However, the brain seems to perform more than just the mechanical transformation of sensory inputs into motor outputs. Often, the person is also said to be subjectively and consciously aware of the objects of perception. I call this phenomenon 'perceptual consciousness', or sometimes 'consciousness' for short. Here I describe a model that formally characterizes the conditions under which this occurs.

Many theories have already been proposed on this topic, but the framework described here differs

*Corresponding author. Tel.: +44 (0)1865 271 444;
Fax: +44 (0)1865 310447; E-mail: hakwan@gmail.com

DOI: 10.1016/S0079-6123(07)68004-2

from them in an important aspect, which is that I do not treat consciousness as the same thing as superior information processing. In other words, I consider that, when the subject is consciously aware of the stimuli, the basic effectiveness of information processing is not necessarily higher. This may differ from common interpretations of most contemporary models of consciousness, which equate consciousness with global processing (Baars, 1988; Dehaene et al., 2003), system complexity (Tononi, 2004), or depth or stability of computation, etc. (Cleeremans, 2005). Commitment to these models often leads to the prediction that consciousness, as compared to the lack of it, will lead to some absolute advantage in terms of information processing. Specifically, within the context of perception, the prediction would be that when one consciously perceives something, rather than sensing it unconsciously, one is always better at identifying it or discriminating it from something else. Intuitively, this may seem plausible enough. In fact, this is often a tacit assumption in experimental studies of perceptual consciousness, even in studies that are conducted by scientists who are not committed to any specific model. This is reflected by the fact that many researchers (Rees et al., 2002) take forced-choice identification or discrimination performance as an index of consciousness: if the performance is high (hits, or high average accuracy), we consider the stimuli consciously perceived, and if the performance is low (misses, or near-chance average accuracy), we consider the stimuli not consciously perceived.

However, despite its ubiquity, the assumption that conscious perception is associated with high performance is unsupported by current empirical data (Lau, in press). In the author's opinion, the general question of whether consciousness plays any special function is still open to empirical investigation. Many have assumed that consciousness might be necessary in executive control or in the generation of spontaneous voluntary action, but recent studies revealed several surprising contradictions to these assumptions (Wegner and Wheatley, 1999; Wegner, 2003; Wegner et al., 2003; Dijksterhuis et al., 2006; Lau et al., 2006, 2007). Whereas none of these show that consciousness has no special function at all, at least they remind us that we should be cautious in accepting theoretical speculations. Consciousness may be functionally less powerful than being assumed previously.

Specifically, within the context of visual perception, we have good reason to think that consciousness is not necessary for good performance in forced-choice detection or discrimination tasks (Lau, in press). This is due to the well-documented phenomenon of blindsight (Weiskrantz, 1986, 1999). After lesions to the primary visual area, blindsight patients report a lack of visual consciousness in the affected region of their visual field. However, when forced to make a decision as to whether something was presented in the region, or to discriminate between two stimuli such as gratings with different orientations, the patients performed well above chance, even though they said they were merely guessing. In some circumstances, they could even guess correctly $\sim$80–90% of the time. This challenges the view that consciousness is the same as high basic effectiveness of information processing.

The foregoing considerations motivate the formulation of a new model that can allow for the dissociation between perceptual consciousness and the basic effectiveness of information processing. We will first discuss a standard method to characterize the basic effectiveness of information processing, which is signal detection theory, and then we consider how we can further extend it so that it could also characterize consciousness.

Signal detection theory

In studies of perception, the subject's performance is often characterized by using signal detection theory (Green and Swet, 1966; Macmillan and Creelman, 1991). Let us take the example in which the subject is presented with a visual stimulus in half of the trials, and a blank screen in the other half. The subject is required to say whether the stimulus is present or absent in each trial. According to the theory of signal detection, the subject's behaviour could be characterized by the detection sensitivity (d') of the subject and the criterion for detection (c). The former is a measure

of perceptual capacity and the latter reflects the decision strategy used. If we assume that there is an internal decision signal (e.g. firing rate in the visual cortex) with which the subject determines whether a stimulus is presented or not, one could construct the probability distribution function for the decision signal given that the stimulus is present, and for the decision signal given that the stimulus is absent (Fig. 1). The fact that the signal strengths for both conditions are reflected by probability distributions means that there is variability or fluctuation in the signal. One usually assumes that these distributions are Gaussians and of equal variance, and the mean for the "stimulus present" distribution is higher than the mean for the "stimulus absent" distribution.

On each trial/presentation, the subject has an internal signal of a particular strength, and has to decide based on this whether the stimulus is present. According to the theory, the subject sets a criterion (c) and responds "yes" ("there is a stimulus") if the internal signal exceeds c, or responds "no" if the internal signal is lower than c. When c is low, we can say that the observer is adopting a liberal strategy (saying "yes" frequently). When c is high, we can say that the observer is adopting a conservative strategy (saying "no" frequently). So long as the two distributions overlap, the subject makes errors, because at the signal strength where the distributions overlap, sometimes the signal is present and sometimes it is not. The subject can only make an informed guess, but in the long run there will still be errors. These errors could take the form of false positives or false negatives, which means the subjects either say "yes" when the signal is actually absent, or say "no" when the signal is actually present. The degree of overlap of the distributions characterizes the perceptual sensitivity, and is measured by d', which is the distance between the two distributions in terms of their variance. The smaller the degree of overlap, the higher the sensitivity.

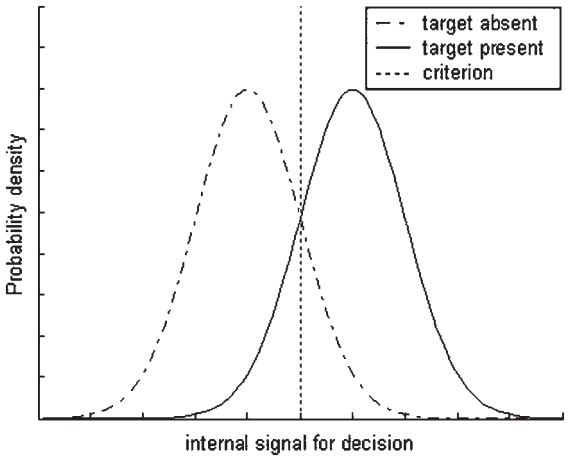


Fig. 1. Standard model of signal detection. Assuming that the target is present in 50% of trials, to perform optimal detection (i.e. to minimize errors), one would set the criterion at a point that best separates the two Gaussian distributions, that is right between their means.

Note that this model also applies to the discrimination between two stimuli, because we can think of the blank as a stimulus that differs from the target visual stimulus. So instead of distinguishing between the stimulus and blank, the subject distinguishes between stimulus A and stimulus B. Formally it is the same.

Criterion setting and maintenance reflect consciousness

Many studies take d' as a measure of perceptual consciousness. If $d' = 0$ the experimenter claims that the subject does not consciously perceive the stimulus. Other studies compare 'hits' and 'misses'. 'Hits' are just trials where the target is present, and the internal signal strength is higher than the criteria and therefore the subject responds "yes". Misses are trials where the target is present, but the signal strength is lower than the criteria, and thus the subject responds "no". Therefore, comparing 'hits' against 'misses' is essentially comparing trials with high internal signal strength and trials with low internal signal strength. In terms of task performance or accuracy, 'hits' are by definition 100% correct, 'misses' are 0% correct.

I argue that neither d' nor internal signal strength is a good measure of consciousness.

It is useful to remind ourselves that signal detection theory was developed partly in order to characterize the behaviour of simple electronics.

A functional photodiode has a $d' > 0$, but it is hard to argue that it is conscious of light. Similarly, blindsight patients show high d' in detecting stimulus in regions of the impaired visual field, and yet, they do not report perceptual consciousness. Sensitivity measure d' characterizes the basic effectiveness of information processing, which we have argued is not necessarily the same as consciousness.

Similarly, the variability of internal signal strength is a basic feature of any noisy detection system. The fact that SDT is useful in characterizing performance means that the internal signal fluctuates, which is why we need to represent the signal strength in terms of probability distributions. The fact that the internal signal fluctuates means that there will naturally be 'hits' and 'misses', if the appropriate criterion is being set. Comparing trials with high internal signal strength against trials with low internal signal strength is just comparing different degrees of effectiveness of information processing (100% performance vs. 0% performance). Similar to d', this captures the objective aspect of perceptual processes, but do not reflect the subjective nature of consciousness (Lau, in press).

I argue that the criterion for perceptual decisions is more relevant to study of consciousness, because of studies of both blindsight and normal observers.

One account of why blindsight patients deny conscious perception of the stimuli is that they adopt an extreme criterion (c) for detection. In other words, despite their high d', they use a very conservative strategy, and respond "no" all the time in a detection situation (Campion and Latto, 1985). This explains their apparent lack of awareness and is also compatible with SDT, because c and d' are independent, in the sense that subjects can set whatever criterion they see fit, regardless of their d'. Alternatively, Azzopardi and Cowey (1997, 1998) have reported that blindsight patients fail to maintain a stable criterion in a detection situation (i.e. a Yes–No task), but has no such problem when they perform '2-alternative forced-choice' tasks, which involve distinguishing the spatial or temporal arrangement of two stimuli. This means that when they perform in a detection task, the criterion they use for decision changes across trials. This leads to an inflation of the measured sensitivity for the detection, but not for the '2-alternative forced-choice' task. The authors argue that this is why the behaviour of blindsight patients is so unusual. Taken together these suggest that failing to set and maintain the criterion properly might be an important factor that contributes to blindsight.

Another reason why we consider decision criteria to be important for consciousness is to due to the results of a recent study (Lau and Passingham, 2006). In that study, I have shown that given the same d', discrimination accuracy, reaction time, and similar stimuli, the same normal subjects can report different levels of perceptual consciousness in two different conditions. After a forced-choice response to discriminate between a square or a diamond, the subjects were asked to also say whether they actually saw the target or they had just guessed. This procedure is based on the "commentary key" paradigm used to test blindsight patients (Weiskrantz, 1999). One can consider the additional Seen/Guessed judgements within the framework of SDT, in that they require the adoption of additional criteria. In other words, instead of setting one criterion to classify the signal strength as high or low, and thus respond "yes it is a square" and "no it is not a square" respectively, one could set three criteria to classify the signal into four ranges, and respond "yes I see it is a square", "yes I guess it is a square", "no I guess it is not a square", and "no I see that it is not a square". Considered this way, the change of proportions of trials reported as 'Seen' rather than 'Guessed' reflects a change in the criteria between seeing and guessing, which could occur even if d' is kept constant across the different conditions.

These considerations suggest that setting and maintaining the criteria appropriately might be an important aspect to perceptual consciousness. Or at the very least, to the extent to which consciousness could be characterized by SDT, the setting and maintenance of criteria seem more relevant than d' and signal strength. However, an account is still needed as to why in some cases subjects set the criterion in an unusual way, and how that affects consciousness.

Optimal criterion setting and its failure

Bayesian decision theory (Kersten et al., 2004) offers a general perspective to the optimal setting of a criterion setting, and may therefore help us to understand its failure. The probability distributions in Fig. 1 represent the probability that the internal signal would be of a certain strength, given that the target stimulus is present or not. When making an optimal decision, one would like to know the opposite, i.e. the probability that the target stimulus is present, given a certain internal signal strength. This could be easily worked out by using Bayes theorem, which mathematically relates any pair of reverse conditional probabilities $P(A|B)$ (probability that A given B) and $P(B|A)$ (probability that B given A) by taking into account the prior probabilities $P(A)$ and $P(B)$. The important prior information here is how frequently the stimulus in presented in general. Quite often, in psychophysical experiments we present the stimulus on 50% of the trials, and tell the subjects so. In this case, the prior information does not bias the optimal criterion; one could determine the criterion by looking at Fig. 1 alone, if the objective is just to maximize accuracy.

After working out the probability that the stimulus is present, given different internal signal strengths, we just set a criterion such that a signal beyond that strength implies that it is likely that the stimulus is present ($P > 0.5$), and a signal below that strength implies it is likely to be absent. We should note that Bayesian decision theory allows a more generalized view to optimality, in that it could also take into account the payoffs and punishments for different types of correct responses and errors, so that one maximizes the expected payoffs instead of accuracy. Here, for simplicity we assume that the objective is just to maximize accuracy, which is equivalent to stipulating that any correct response is worth the same as the avoidance of any incorrect response.

Under these assumptions, i.e. unbiased prior and maximizing accuracy, one would set the criterion for decision at a level that best divides the two probability distributions as shown in Fig. 1, right between the two distributions. This is the optimal criterion, in the sense that it produces a minimum amount of errors. Why do blindsight patients not fix the criterion stably at this optimal point? In the case of the study where d' was matched in two conditions, why would the same subjects set the additional Seen/Guessed criteria differently in two conditions, given that the same d' in the conditions implies similar probability distributions for the stimuli (assuming that they are both Gaussians and of equal variance)?

The solution I offer is as follows: given that the optimal criteria are determined by the probability distributions for the stimuli, one's knowledge of these distributions is important. However, the distributions describe the probabilistic behaviour of the internal signal, which has to be learned over time. The learning of one's own internal signal produces representations concerning the internal signal, which itself is a representation of the external stimuli. In this sense, we are creating representations of representations, and thus they are described as "higher-order" representations (Rosenthal, 2000, 2002). I propose that perceptual consciousness depends on our Bayesians decisions, i.e. criterion setting, based on these higher order representations.

Let us take a simple imaginary example by assuming that we try to detect a dim light by using the firing rate in the primary visual cortex as the only source of evidence. If there is no light, the neurons may fire on average at 10 Hz, with a standard deviation of 5 Hz. If there is light, the neurons may fire on average at 15 Hz, again with a standard deviation of 5 Hz. So given this information, a reasonable subject who has the goal to maximize accuracy would set a criterion at 12.5 Hz, and then say there is light when the firing rating exceeds that criterion and say that there is no light if it does not. It follows, then, that what criterion is set depends on what the subjects *think* their own average firing rates are when there is a light, and when there is no light. To make an estimate of the average level of a fluctuating signal, one has to statistically sample the data over time. So essentially, the subject has to learn the probabilistic behaviour for their own neurons. If this learning fails, or is incomplete, such that the subject makes a biased or incorrect estimate of the firing rates, the criterion they set would be suboptimal.

How exactly this learning occurs is beyond the scope of this paper. Previous work on criterion setting in psychology (Treisman, 1984) has suggested methods of learning that do not involve explicitly modelling the probability distributions, but that could be treated as a special heuristic that satisfies the same goal of learning the distributions and basing the criterion on the result of that learning. Without making any strong assumptions about the form of the distributions, a general solution to this learning problem could take the form of a standard Bayesian learning procedure. In the simple case where there are only two stimulus conditions (e.g. dim light or no light) and the subjects are told whether they make correct decisions after responding, the learning should be a fairly straightforward problem. However, in real life when stimuli are multidimensional and feedback is not always available, this could take more complicated forms of unsupervised learning. The critical point here is that the result of this learning affects the setting and maintaining of criteria in the detection or discrimination of stimuli, which might be important for the normal functioning of perceptual consciousness. We now turn to how different cases of failure of perfect learning can produce behaviour that characterizes disturbance of perceptual consciousness.

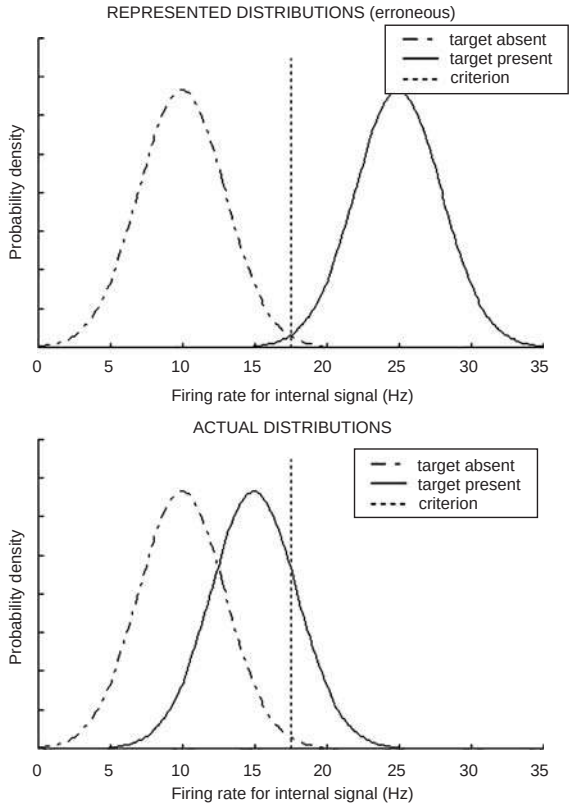


Fig. 2. Misrepresentation. Because one could only set the criterion based on the learned distributions (upper graph), instead of the actual distributions (lower graph), if the two are different, i.e. the learned distributions are incorrect, one sets the criterion suboptimally. Here only one form of misrepresentation is depicted. However, one could imagine the other case, such as representing both the target absent and target present distributions as higher than reality, or both lower than reality, etc.

Misrepresentations

One obvious way of failing to learn the probability distributions is to grossly misrepresent the mean or the variance of the signal (Fig. 2). So in the example given above, the subject could overestimate the firing rate given that the dim light is present, so that it is thought that the average rate is 25 Hz instead of 15 Hz. The subject might then set the criterion, reasonably, at the mid-point (i.e. 17.5 Hz) between the expected averages for the signal present and signal absent conditions, which are 10 and 25 Hz. This high criterion of 17.5 Hz would lead to a high portion of false-negatives, because actually the average firing rate is only 15 Hz when the dim light is on; the subject will be missing most of it. The subject would behave as if the dim light is not perceived most of the time, because the subject responds ''no'' even when the stimulus is present and the firing rate is at its most likely frequency. The actual d, however, remains the same, because it measures the distance between the actual distributions, but not the learned distributions.

This would be similar to the behaviour exhibited by blindsight patients, where a negative response is usually given in detection, although a fairly high d' is implied by other indirect forced-choice measures. The primary visual cortex is likely to be one important major source of the internal signal. After a lesion to the cortex, the actual internal

signal decreases dramatically, with the remaining weak signal possibly routing through subcortical pathways. If the subject fails to learn that the signal has dropped, and uses the old criterion for detection, this is equivalent to overestimating the signal as described above. In fact, it is likely that the primary visual cortex is also a source of the baseline noise, i.e. the signal when no stimulus is present. A lesion to the primary visual cortex is thus likely to shift both probability distributions (signal present and absent) to the negative direction. Failing to learn this shift can result in a dramatic positive deviation from the optimal criteria. Understood this way, blindsight could be partly due to a failure to learn the reduction in signal strength after lesion to the primary visual cortex; after the lesion, because of the inappropriate criterion, blindsight patient mis-classifies stimuli as noise.

Hallucinations could be treated as the opposite within this framework. Hallucination could be treated as the production of false-positives, in the sense that noise is being mis-classified as stimuli. By this definition, one hallucinates while dreaming; in dreams we consciously perceive stimuli that are not really there. According to the present framework, this could be due to the underestimation of the signal strength. When brain activity is monitored by electroencephalogram (EEG), sleep can be divided into different stages by the EEG pattern. Dreams are more likely to be reported during a stage of sleep that is characterized by rapid eye movement (REM), and brain activity of relatively high frequency and intensity. Let us assume that the overall signal during REM-sleep is higher. If the brain maintains the same criterion for detection over alternations of REM and non-REM sleep, it would be predicted that false-positives are a lot more likely during REM-sleep, because of the higher signal intensity. Perhaps during sleep, when the brain is not actively learning, it only makes a general estimation of the probability distributions of signal for both REM and non-REM sleep combined, and that is why we fail to set a appropriately high criterion during REM sleep. With this inappropriately low criterion for detection, one mis-classifies noise as stimuli.

Ambiguity

Another way to fail to fully learn the probability distributions is to learn it with high ambiguity (Fig. 3). Ambiguity is formally defined as uncertainty about probabilities (Ellsberg, 1961). The graphs in Figs. 1 and 2 represent the probability that the signal intensity is of a certain strength, given that the signal is absent or present. The fact that we use a line to represent the bell shape curves means that at each signal intensity level, there is a definite number that represents the probability. In reality, this is possible: the probability that when there is a stimulus, the probability that the signal strength is between x and $x+1$ could be exactly 0.01245, for example. However, for any subject who is learning this distribution, one could only learn this with a certain degree of uncertainty. The probability might be expected to be at 0.01245

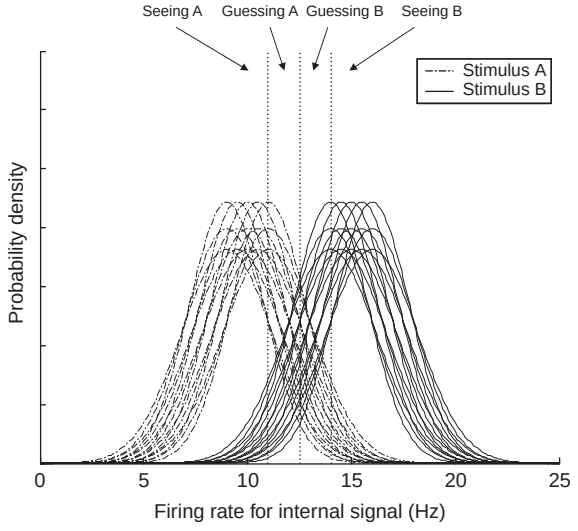


Fig. 3. Ambiguity. Rather than representing the distributions with absolute certainty, one might represent them as a possible range of values (ambiguity). Here one represents the probability distributions with standard deviations 1.8–2.2 Hz, and means of 9–11 Hz and 14–16 Hz respectively for Stimulus A and B. These ranges could be considered as similar to confidence intervals or error bars. When subjects were asked whether they were "guessing", they could be setting three criteria as depicted in the diagram. It is possible that when the level of ambiguity increases, the criteria for guessing might change accordingly so that more trials are classified as "guessing".

(i.e. it is the most likely), but the 95% confidence interval may cover between 0.01241 and 0.001248. In other words, when we try to learn the information on Fig. 1, which depicts reality, in our brain we try to create the same graphs that best represent the same information, but possibly with error bars on it, to reflect how certain we are about our estimation of reality (Fig. 3). In this sense, representing ambiguity is not a failure, but rather a useful way to capture the degree to which we are certain. However, having very high ambiguity means that we are not really certain about our estimation, which is not favourable for us.

Recall that under the assumptions we have taken, the optimal way to set the criterion is to set it such that it best classifies the two distributions for stimulus present and stimulus absent, which is the mid-point between the two means. For signal strength above the criterion, we expect that it is more likely that the stimulus is present than not. However, even if we estimate that the probability that the stimulus is present is bigger than 0.5, it may not be significantly bigger than 0.5. Maybe the error bar or confidence interval is so large that it covers the point of 0.5. In other words, we estimate the stimulus to be there, rather than not there, but we are not so sure about this estimation.

This formally characterizes guessing. Obviously, having guessed A instead of B we think that A might be marginally more likely. But it is counted as guessing because we are not 'significantly' certain about our decision. Significance here could be defined as in statistics, i.e. when we are significantly certain about a decision, the 95% confidence interval (or whatever other confidence level) for the likelihood that we are correct does not cover 0.5 (chance). The amount of guessing depends on the level of ambiguity.

This offers a possible explanation as to why sometimes given the same d', discrimination accuracy, reaction time, and similar stimuli, the same subjects can report different levels of perceptual consciousness in two different conditions (Lau et al., 2006). This is because although the underlying distributions are the same, the subjects may learn the distributions for the two conditions with different level of ambiguity (uncertain regarding the distributions). Subjects set the criteria for Seen/Guessed at the signal strength at which they become significantly certain about the discrimination, and this depends on the level of ambiguity. In other words, if the error bars are large and they overlap greatly, for the distributions, they set the criteria accordingly so that many trials are classified as ''Guesses''. The underlying actual distributions, however, could be exactly the same, and thus the actual performance levels are matched.

Dynamic fluctuation

Finally, we briefly consider another type of failure to learn the distributions properly, which is to fail to make stable estimates of the distributions. In a way, this is a form of misrepresentation, but here the emphasis is on how this misrepresentation fluctuates dynamically. Unless the subjects have learned the distributions perfectly and precisely, one would expect that in every trial, the subjects would acquire new information and thereby change the estimation accordingly. In this sense, dynamic fluctuation is a sign that we are learning the distributions. If the learning is effective, one would expect that the estimation of the distributions to fluctuate less and less, and eventually converge to the true forms. However, if the learning itself is not optimal, this fluctuation may continue.

As mentioned above, it has been reported that blindsight patients fail to maintain a stable criterion in a detection situation (i.e. a Yes–No task), but have no such problem when they perform '2-alternative forced-choice' tasks (Azzopardi and Cowey, 1997, 1998). This might seem hard to explain, because if they fail to maintain a stable criterion in the Yes–No task because they fail to remember where to put the criterion, they should also fail to maintain a stable criterion in the other task. However, if the failure is due to the fluctuation in the learned representation for the stimulus-absent distribution only, but not for the other distributions, this could explain the jittering of criterion for the task of detection but not discrimination tasks.

Fluctuation is interesting to consider because unlike the other two forms of failure, it actually affects d' as measured by conventional methods. This is because if one's learned distributions fluctuate, one's criterion also fluctuates; presumably one sets the criterion according to the most up-to-date estimation. When the criterion fluctuates, effectively we have a reduced d'. This has been suggested to be the explanation for why in detection tasks, blindsight patients have a d' that is lower than expected, when one estimates it from the d' for the 2-alternative forced-choice' tasks (Azzopardi and Cowey, 1997, 1998).

Empirical support

It is important to note that the aforementioned three types of failure are not incompatible with each other. One could exhibit all three problems, or just one, or two of them. Therefore it is not really an issue at the moment to determine which is *the* failure that causes disturbance of perceptual consciousness. In reality all three of them may play some role. It is a matter for future research to empirically investigate the relative importance of each type of failure of representation in different contexts. At this stage, we consider empirical evidence that support the general notion that perceptual consciousness depends on the representation of the probability distributions that describe the behaviour of the internal signal.

One basic prediction of this framework is the dissociation between consciousness and detection/discrimination performance. This is because except in the case where dynamic fluctuation is significant, the basic detection and discrimination performance is determined by the actual probability distributions. However, consciousness depends on *the representation* of the probability distributions, by which we set and maintain the criteria. This representation, as we discussed above, may err. Therefore the framework predicts that given the same d', the same subjects performing the same detection/discrimination task may report different levels of consciousness under different conditions, if the distributions are represented differently.

In a recent study that has been mentioned above (Lau et al., 2006), we showed exactly that. In that task, subjects were required to discriminate between a square and a diamond figure (Fig. 4). We also asked subjects, after the discrimination in each trial, to state whether they consciously saw the identity of the target or that they just guessed what it was. Therefore, we have both an objective forced-choice measure of performance, as well as a subjective measure of perceptual consciousness. The target was metacontrast masked at different stimulus onset asynchrony (SOA, i.e. the temporal distance between the target and the mask), so as to produce different conditions with various levels of difficulty. We capitalized on the fact that the masking function (performance against SOA) is U-shaped for metacontrast masking, which means that there will always be two SOA points at which the performance levels will be matched (Fig. 5). We found that at the two SOA points, the subjective levels of perceptual consciousness differed, in that in the shorter SOA condition subjects claimed to be guessing more frequently. Within the present theoretical framework, this could be understood in terms of different levels of ambiguity for the two SOA conditions.

If this difference in levels of consciousness is due to a difference in the higher order representations, that is the representation of how the internal signal behaves statistically, it should be associated with a difference in neural activity in the relevant brain area. The prefrontal cortex is likely to play an important role in forming and maintaining these higher order representations, because it has also been implicated in studies of uncertainty and learning (Daw et al., 2005; Huettel et al., 2006; Yoshida and Ishii, 2006). Also, it receives anatomical projections from areas of all sensory modalities, and has been considered as situated at the top of the information processing hierarchy of the brain (Goldman-Rakic, 1995; Fuster, 1997; Passingham et al., 2005). The internal signal, on the other hand, is likely to be represented in the occipital and temporal cortices, where neurons code specific visual information. Therefore, our models predicts that if we compare the trials for the two SOA points, where the subjective level of perceptual consciousness differs but forced-choice

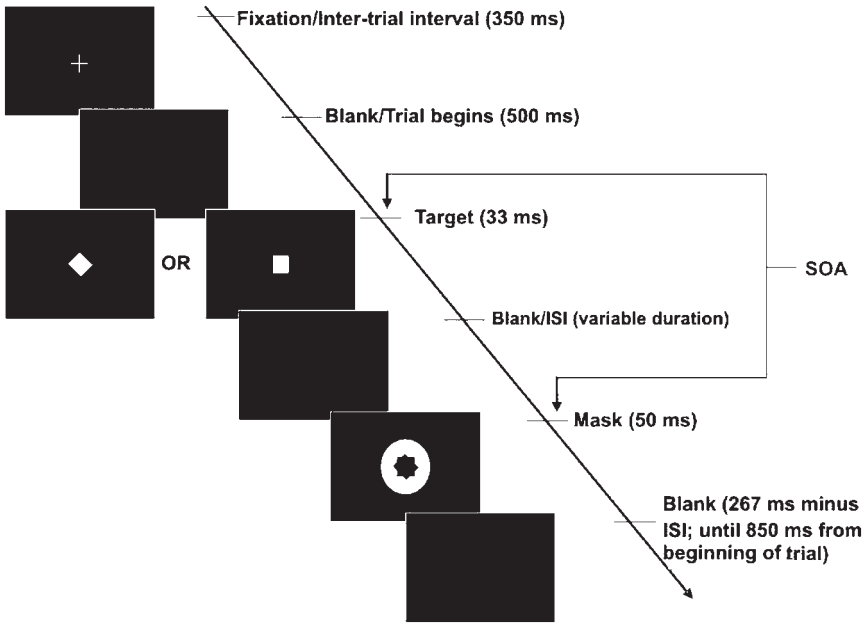


Fig. 4. Visual discrimination task with metacontrast masking. After the presentation of the target and the mask, the participants were first asked to decide whether a diamond or a square was presented. Then, they had to indicate whether they actually saw the target, or that they simply guessed the answer. Shown in the brackets are the durations of each stimulus.

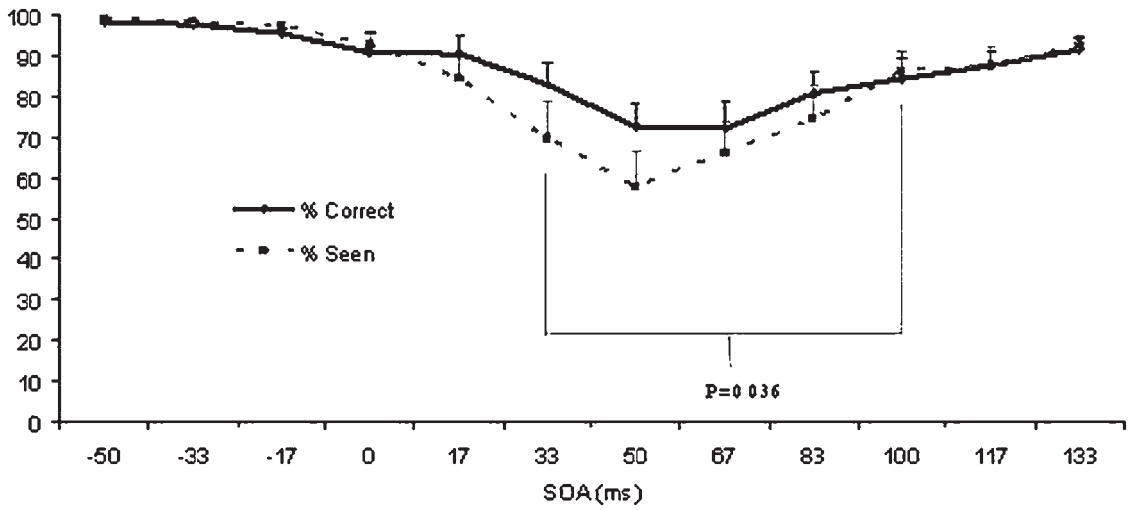


Fig. 5. Consciousness and performance. These results were obtained using the procedure described in Fig. 1, except that it also included trials where the mask was presented before the target (paracontrast masking). Note that at the SOAs where the performance levels (% correct) were the same (e.g. 33 and 100 ms), the awareness levels (% seen) differed significantly.

performance does not, there should be a difference in activity in the prefrontal cortex but not the in the early visual areas. In fact, this is what we observed (Lau et al., 2006). This is a counter-intuitive finding because most theories of visual consciousness suggest that the critical neural correlate should be in visual areas (Zeki and Bartels, 1999; Lamme and Roelfsema, 2000; Lamme, 2003). Even though some researchers have proposed that a 'frontal-parietal network' that might be important for consciousness, they typically suggest that this is only important in addition to the visual areas (Rees et al., 2002). However, in our study, the dorsolateral prefrontal (DLPFC) cortex is the only area where we could find a significant difference in activity. This fits with the central idea of the model that perceptual consciousness depends on higher order representations (in the prefrontal cortex), and it can change in the absence of a difference in the internal visual signal (in occipital and temporal visual areas).

Interestingly, the DLPFC has also been implicated in a study of blindsight. Sahraie et al. (1997) have reported results from a study on a blindsight subject (known as GY). The subject has a lesion to the primary visual cortex that affects roughly half of his visual field, stimuli presented to which yield no phenomenal visual awareness. The authors presented to this ''blind field'' slowly moving (3°/s) stimuli of which the subject was unaware, and found that the subject could nonetheless discriminate the direction of the horizontal movement at slightly above 80% correct. This visual stimulation was associated with a lack of significant activation in the DLPFC. However, when the speed of the movement of the stimuli was increased to 20°/s, the subject reported a sense of awareness even though the visual presentation was to the ''blind field'' (a phenomenon known as type II blindsight), and the performance of discrimination was above 90% correct. This visual stimulation was also associated with a significant activation of DLPFC. The performance levels for discrimination task in these conditions were different, but could be considered roughly matched, because they were both well above chance. Incidentally, when visual stimuli were presented to the unimpaired field of the blindsight subject, there was also significant activation in the DLPFC. These results, as in the study discussed above (Lau et al., 2006), support the idea that activity in the DLPFC vary in relation to the level of awareness, even when performance level is not an important contributing factor.

Finally, the DLPFC has also been implicated in studies concerning sleep and dreams (Maquet et al., 1996; Muzur et al., 2002). As explained earlier, within the present framework dreams and hallucinations could be considered as the opposite of blindsight. If the internal signal intensity increases, but the higher order representations and thus also the criterion remain the same, we are likely to produce false-positives for conscious perception, and this formally characterizes dreams and hallucinations. We have argued that in REM sleep the internal signal is likely to be higher than in non-REM sleep. And we know that dreams are likely to be reported during REM sleep, and unlikely to be reported during non-REM sleep. Therefore, one way that the aspects of dreams regarding perceptual consciousness could be explained is that activity in the DLPFC should be similar for both REM and non-REM sleep. This possibility reflects that the higher order representations and criterion remain the same between REM and non-REM sleep. If this is true, the higher internal signal intensity during REM sleep would produce false-positives. This pattern of activity is in fact found in the DLPFC when neural activity during REM and non-REM sleep was compared using positron emission topography (PET) (Maquet et al., 1996; Muzur et al., 2002). Compared to wakefulness, during non-REM sleep many areas are deactivated. During REM sleep, most areas are reactivated to normal wakefulness level. However, the DLPFC remains deactivated in REM sleep, in a level similar to that during non-REM sleep. In fact, the DLPFC is the only area in the prefrontal cortex that remains deactivated (Muzur et al., 2002).

Taken together, these results are compatible with the notion that the subjective aspects of perceptual consciousness depend on the higher order representations of how the internal signal behave, which are likely to be associated with the prefrontal cortex. Future studies should further

examine their interactions with areas that are likely to represent the internal signal, such as occipital and temporal areas as in the case of vision.

Finally, the model predicts that manipulating the activity in the prefrontal cortex, and thus the higher order representations, should change the level of perceptual consciousness but not forced-choice performance in detection/discrimination tasks. This offers a solution to the controversy as to whether lesion to the prefrontal cortex impairs consciousness. Critics (Pollen, 1995, 1999) have argued that there has been a lack of reported cases of such prefrontal damaged patients. However, most previous studies of visual consciousness have focused on forced-choice performances, which we predict should not show any difference if only the higher order representations are manipulated. Instead, subjectively reported perceptual consciousness should change. We are currently using transcranial magnetic stimulation (TMS) to test these hypotheses.

Philosophical issues and final remarks

We started off by arguing that signal detection theory is too simple to characterize perceptual consciousness. One could argue that the model we presented here, that perceptual consciousness depends on the setting and maintaining of criterion based on representations of the statistical behaviour of internal signals, is not substantially more complicated. Does this really solve the ''hard problem'' of explaining the subjective nature of consciousness in terms of physical facts (Chalmers, 1996)? If d' is a bad measure of consciousness because a photodiode could have a d' value as high as a conscious human being, does the same criticism not apply to our model? One could imagine building a device that learns dynamically its only internal signal for detection/discrimination, and maintains optimal criterion according to Bayesian decision theory. Does this make the device conscious?

I do not intend to claim that the present framework readily solves all these problems, or at least I am not going to argue so within the space of this paper. The foregoing thought experiment is an interesting one, but one should not overlook the complexity of a moving robot that could dynamically learn the probability distributions for its internal signal for performing detection/discrimination in a changing environment. However, here I only recommend a minimal interpretation of the model: it formally characterizes perceptual consciousness, in the sense that it describes the conditions under which consciousness is intact or disturbed in a human subject. It is not supposed to explain all features of consciousness. The main point I wish to make is that perceptual consciousness depends on some form of criterion setting, though it does not mean that all forms of criterion setting is directly relevant to consciousness. For instance, there might be a specific criterion for conscious perception, and a different one for responding or communicating the information to others. The important hypothesis is that the principles for the maintaining and the setting of the criterion for conscious perception should be compatible with the framework described here.

I also argue that because the present model takes the form of a higher order representational theory as discussed frequently in philosophy (Rosenthal, 2000, 2002), it shares similar philosophical explanatory power. To the degree to which higher order representational theories in general can solve the philosophical problems associated with 'explaining' consciousness, I speculate that the present model probably does at least equally well. In fact, I am going to argue that it is likely to be more attractive than many other versions of higher order representational theories.

One problem of higher order representational theories is the problem of mismatch between higher and lower levels of representations. Normally, in the current philosophical literature, both the higher and lower level of representation is taken to represent information regarding the external world. If the first level representation represents 'red' but the higher order representation represents 'green', it is unclear what the conscious experience should be (Neander, 1998). However, in the present model, the higher order representation represents a scale by which the first-order representation (the internal signal) could be interpreted. The internal signal carries no fixed meaning unless

one is to have some access to the higher order representations; a firing rate of 5 Hz in the early visual cortex could mean that a signal is very likely to be present, or very unlikely to be so, depending on the higher order representations. Similarly, the higher order representations do not make sense outside of the context of the internal signal. This way, a mismatch between the levels in the above sense is simply not possible: their content cannot directly contradict, because they are never meant to duplicate each other.

Finally, the present model should also shed light on the function of perceptual consciousness. I have argued earlier than one motivation of the model is that the function of consciousness is unclear. However, if perceptual consciousness depends on the correct knowledge of the variability of the internal signal, when one is conscious one should also be able to perform functions that depend also on this knowledge. An obvious example might be optimal betting, based on one's own performance (Persaud et al., 2007).

As a closing remark, I further suggest that some form of social interaction may also depend on this knowledge which underlies perceptual consciousness. Several observers could make optimal joint decisions by combining their information regarding the same external stimulus. This optimal joint decision could be predicted by a Bayesian an approach, in which the team of observers could be considered as a ''Bayesian committee''. These approaches typically assume that each of the observers know their own variance of their internal signal. In other words, if the present model is correct, that perceptual consciousness depends on the correct knowledge of the variability of one's own internal signal, subjects who are perceptually conscious should be suitable candidates for joining a Bayesian committee which gives optimal team responses. One could imagine being required to team up with another observer who has a certain detection sensitivity. In order to maximize joint performance, when opinions differ one would discuss with the partner and compromise based on the relative levels of confidence in the decision in a particular trial. It is not difficult to see that one would rather team up with normal observers rather than with blindsight patients who have the same sensitivity. If the team partners claim that their confidence is low and they are guessing all the time even when their responses are correct, negotiating and compromising for an optimal joint response becomes impossibly difficult.

Acknowledgments

I thank Chris Frith, Peter Dayan, Peter Latham, Allan Hobson, Karl Friston, and Tim Behrens for discussions that contributed to the development of the presented ideas. I also thank Chris Frith and Dick Passingham for comments on an earlier version of this manuscript. This work is supported by the Wellcome Trust.

References

Azzopardi, P. and Cowey, A. (1997) Is blindsight like normal, near-threshold vision? Proc. Natl. Acad. Sci. U.S.A., 94: 14190–14194.

Azzopardi, P. and Cowey, A. (1998) Blindsight and visual awareness. Conscious. Cogn., 7: 292–311.

Baars, B.J. (1988) A Cognitive Theory of Consciousness. Cambridge [Cambridgeshire]. Cambridge University Press, New York.

Campion, J. and Latto, R. (1985) Apperceptive agnosia due to carbon monoxide poisoning. An interpretation based on critical band masking from disseminated lesions. Behav. Brain Res., 15: 227–240.

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

Cleeremans, A. (2005) Computational correlates of consciousness. Prog. Brain Res., 150: 81–98.

Daw, N.D., Niv, Y. and Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci., 8: 1704–1711.

Dehaene, S., Sergent, C. and Changeux, J.P. (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc. Natl. Acad. Sci. U.S.A., 100: 8520–8525.

Dijksterhuis, A., Bos, M.W., Nordgren, L.F. and van Baaren, R.B. (2006) On making the right choice: the deliberation-without-attention effect. Science, 311: 1005–1007.

Ellsberg, D. (1961) Risk, ambiguity, and the savage axioms. Q. J. Econ., 75: 643–669.

Fuster, J.M. (1997) The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe (3rd ed.). Lippincott-Raven, Philadelphia, PA.

Goldman-Rakic, P.S. (1995) Architecture of the prefrontal cortex and the central executive. Ann. N.Y. Acad. Sci., 769: 71–83.

Green, D. and Swet, S. (1966) Signal Detection Theory and Psychophysics. Wiley, New York.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T. and Platt, M.L. (2006) Neural signatures of economic preferences for risk and ambiguity. Neuron, 49: 765–775.

Kersten, D., Mamassian, P. and Yuille, A. (2004) Object perception as Bayesian inference. Annu. Rev. Psychol., 55: 271–304.

Lamme, V.A. (2003) Why visual attention and awareness are different. Trends Cogn. Sci., 7: 12–18.

Lamme, V.A. and Roelfsema, P.R. (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci., 23: 571–579.

Lau, H.C. (in press). Are we studying consciousness yet? In: Weiskrantz L., Davies M. and Parker A. (Eds.), Chichele Lectures 2006: Frontiers of Consciousness. Oxford University Press, Oxford.

Lau, H.C. and Passingham, R.E. (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. Proc. Natl. Acad. Sci. U.S.A., 103: 18763–18768.

Lau, H.C., Rogers, R.D. and Passingham, R.E. (2006) On measuring the perceived onsets of spontaneous actions. J. Neurosci., 26: 7265–7271.

Lau, H.C., Rogers, R.D. and Passingham, R.E. (2007) Manipulating the experienced onset of intention after action execution. J. Cogn. Neurosci., 19(1): 81–90.

Macmillan, N. and Creelman, C. (1991) Detection Theory: A User's Guide. Cambridge University Press, Cambridge, England.

Maquet, P., Peters, J., Aerts, J., Delfiore, G., Degueldre, C. Luxen, A. et al. (1996) Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. Nature, 383: 163–166.

Muzur, A., Pace-Schott, E.F. and Hobson, J.A. (2002) The prefrontal cortex in sleep. Trends Cogn. Sci., 6: 475–481.

Neander, K. (1998) The division of phenomenal labor: a problem for representational theories of consciousness. Philos. Perspect., 12: 411–434.

Passingham, R.E., Rowe, J.B. and Sakai, K. (2005) Prefrontal cortex and attention to action. In: Humphreys G.W. (Ed.), Attention in Action. Taylor & Francis Group, Inc., London, UK.

Persaud, N., McLeod, P. and Cowey, A. (2007) Post-decision wagering objectively measures awareness. Nat. Neurosci., 10(2): 257–261.

Pollen, D.A. (1995) Cortical areas in visual awareness. Nature, 377: 293–295.

Pollen, D.A. (1999) On the neural correlates of visual perception. Cereb. Cortex, 9: 4–19.

Rees, G., Kreiman, G. and Koch, C. (2002) Neural correlates of consciousness in humans. Nat. Rev. Neurosci., 3: 261–270.

Rosenthal, D.M. (2000) Consciousness, content, and metacognitive judgments. Conscious. Cogn., 9: 203–214.

Rosenthal, D.M. (2002) How many kinds of consciousness? Conscious. Cogn., 11: 653–665.

Sahraie, A., Weiskrantz, L., Barbur, J.L., Simmons, A., Williams, S.C. and Brammer, M.J. (1997) Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. Proc. Natl. Acad. Sci. U.S.A., 94: 9406–9411.

Tononi, G. (2004) An information integration theory of consciousness. BMC Neurosci., 5: p. 42.

Treisman, M. (1984) A theory of criterion setting: an alternative to the attention band and response ratio hypotheses in magnitude estimation and cross-modality matching. J. Exp. Psychol. Gen., 113: 443–463.

Wegner, D.M. (2003) The mind's best trick: how we experience conscious will. Trends Cogn. Sci., 7: 65–69.

Wegner, D.M., Fuller, V.A. and Sparrow, B. (2003) Clever hands: uncontrolled intelligence in facilitated communication. J. Pers. Soc. Psychol., 85: 5–19.

Wegner, D.M. and Wheatley, T. (1999) Apparent mental causation. Sources of the experience of will. Am. Psychol., 54: 480–492.

Weiskrantz, L. (1986) Blindsight. A case study and implications. Oxford University Press, Oxford, UK.

Weiskrantz, L. (1999) Consciousness Lost and Found. Oxford University Press, Oxford.

Yoshida, W. and Ishii, S. (2006) Resolution of uncertainty in prefrontal cortex. Neuron, 50: 781–789.

Zeki, S. and Bartels, A. (1999) Toward a theory of visual consciousness. Conscious. Cogn., 8: 225–259.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 5

Subjective measures of unconscious knowledge

Zoltán Dienes*

Department of Psychology, University of Sussex, Brighton BN1 9QH, UK

Abstract: The chapter gives an overview of the use of subjective measures of unconscious knowledge. Unconscious knowledge is knowledge we have, and could very well be using, but we are not aware of. Hence appropriate methods for indicating unconscious knowledge must show that the person (a) has knowledge but (b) does not know that she has it. One way of determining awareness of knowing is by taking confidence ratings after making judgments. If the judgments are above baseline but the person believes they are guessing (guessing criterion) or confidence does not relate to accuracy (zero-correlation criterion) there is evidence of unconscious knowledge. The way these methods can deal with the problem of bias is discussed, as is the use of different types of confidence scales. The guessing and zero-correlation criteria show whether or not the person is aware of knowing the content of the judgment, but not whether the person is aware of what any knowledge was that enabled the judgment. Thus, a distinction is made between judgment and structural knowledge, and it is shown how the conscious status of the latter can also be assessed. Finally, the use of control over the use of knowledge as a subjective measure of judgment knowledge is illustrated. Experiments using artificial grammar learning and a serial reaction time task explore these issues.

Keywords: conscious; unconscious; mental state; subjective measures; higher order thoughts; implicit learning; confidence; artificial grammar learning; serial reaction time; exclusion task

Ever since the birth of experimental psychology at the end of the 19th century, psychologists have been interested in the distinction between conscious and unconscious mental states (Sidis, 1898). Recently, there has been a resurgence of interest in the distinction, as demonstrated by both purely behavioural and also brain imaging research (e.g. trying to find the neural correlates of consciousness). All such research requires a methodology for determining the conscious status of a mental state. This chapter will argue for the use of 'subjective measures' for assessing the conscious status of knowledge states (see also Gaillard et al., 2006). Subjective measures measure the extent to which people think they know, as opposed to measuring how much people simply know. The assumption is that knowledge is conscious if it *subjectively* seems to people that they know when they do know, i.e. if people are aware of knowing. First, we discuss the philosophical basis of subjective measures and illustrate the application of two subjective measures — the guessing criterion and the zero-correlation criterion — to learning artificial grammars. Then we will consider the problem of bias and some practical details concerning the best ways of implementing the criteria. Next we show how the criteria indicate the conscious status of only some knowledge contents

*Corresponding author. Tel.: +44 1273 678550;
Fax: +44 1273 678058; E-mail: dienes@sussex.ac.uk

DOI: 10.1016/S0079-6123(07)68005-4

(judgment knowledge) but not others (structural knowledge). We show how the conscious status of structural knowledge can also be assessed with subjective measures. Finally, we show how the control a person has over the use of knowledge can be used as an alternative subjective measure of the conscious status of judgment knowledge. In sum, the chapter should give a researcher the tools to use subjective measures in a wide variety of research settings.

Philosophical basis of subjective measures

A first-order state is a mental state that is about the world. Forming a representation in the visual system that an object is moving up is an example of a first-order representation. First-order representations allow facilitated interaction with the world, for example discriminations about the world. That is their function, that is what makes them knowledge at all. Blindsight patients, who have damage to an area of the cortex called V1, can say whether an object is moving up or down at above 80% accuracy. Yet they often claim not to be seeing, often just to be purely guessing[1] (Weiskrantz, 1997). Our strong intuition is to say the seeing is unconscious precisely because the blindsight patient is not aware of seeing; they do not have an accurate mental state about the mental state of seeing. That is, it is *because* they lack a second-order state (a mental state about a mental state) that it seems right to say their seeing is unconscious. In general, subjective measures ask people to report the mental state they are in, not just to make discriminations about the world. Subjective measures test for the presence of suitable second-order states.

One reason for urging subjective measures as the appropriate method for measuring the conscious status of mental states would be the theory that a mental state being conscious is constituted by there being (or potentially being) suitable second-order states. The other reason would be the theory that a state's being conscious happens to enable higher order states, at least for us humans. Either sort of theory justifies the use of subjective measures. We discuss both in turn.

In the 1980s, David Rosenthal from New York and Peter Carruthers, then at Sheffield, independently took up an idea that can be traced back to Aristotle (in the Western tradition), namely that a mental state's being conscious arises because of actual (Rosenthal, 2005) or potential (Carruthers, 2000) higher order states. Rosenthal's argument consists of two very plausible premises.

The first premise: a conscious mental state is a mental state of which we are conscious. Although this premise might seem circular, it is not. Rosenthal distinguishes two separate senses of 'conscious' used in the premise. There is *transitive consciousness*, namely consciousness *of* something. Transitive consciousness always takes an object. For example, looking at a tree and thereby being conscious of the tree. Another sense is *state consciousness* — a mental state can be conscious, as when one consciously sees, consciously thinks or consciously knows. The two senses are different. A blindsight patient is, in one way of talking, conscious of the object moving up — not consciously aware of it, but just conscious of it — because he has a visual representation that makes him sensitive to the object's motion. There is transitive consciousness (of the motion). Yet that mental state of seeing is not conscious, not state conscious. The first premise relates the two very different senses of conscious by proposing that state consciousness consists of transitive consciousness of the state. But how does transitive consciousness arise?

The second premise: the way we become conscious of mental states is by thinking about them. There are two ways we can become conscious of anything: by perceiving it or by thinking about it being there. I can see you there or I can close my eyes and think of you being there; either way I am conscious of you being there. Philosophers debate about whether we perceive mental states. But certainly we can think of them. (Note the theory, thus far, does not state what having a mental state — like thinking — consists in. But we do not have to solve all problems at once to be making progress.)

Putting the two premises together, Rosenthal concludes that a mental state being conscious consists of there being a higher order thought

[1] Depending on speed, at fast speeds they are aware of seeing, see also Zeki (this volume).

(HOT) asserting we are in that state. (For example, we see when we are aware of seeing by thinking that we see.) If you want to deny the conclusion you need to consider which of the premises you wish to deny.

There are broadly two major philosophical intuitions concerning what consciousness consists in. One is the higher order state theory just mentioned. The other is the idea that conscious states are those that are 'inferentially promiscuous' or in other words 'globally available' (Baars, 1988). If you consciously know something you can in principle use that knowledge in conjunction with anything else you consciously know or want in order to draw inferences, make plans or form intentions. In contrast, unconscious knowledge may be available for only a limited set of uses. In adult humans it follows — given they have concepts of mental states — when you see something consciously that knowledge is quite capable of being used to make other inferences, like to think that you are seeing. The property of inferential promiscuity ensures that in adult humans conscious knowledge will enable the inference that one has that knowledge; in other words, it will enable HOTs. If when you probe for HOTs you cannot find any suitable ones, the knowledge was not inferentially promiscuous, and hence not conscious.

In summary, both major (Western) philosophical intuitions concerning the nature of consciousness — higher order state and inferential promiscuity — justify the use of subjective measures as tests of the conscious status of mental states. In higher order theories, a state being conscious is constituted by a higher order state, and in inferential promiscuity theories, a state's being conscious will allow a higher order thought if you ask for one.

Particular theories of consciousness may elaborate these themes in different ways. For example, Cleeremans (this volume) proposes a graded representation theory of consciousness. Low-quality representations remain unconscious, only high-quality representations may become consciousness. But importantly, Cleeremans regards representational quality as necessary but not sufficient. Conscious states also require meta-representation, i.e. one must represent oneself as having the first-order representation. Meta-representation is a higher order state; thus, subjective measures are also directly motivated by Cleereman's theory. Lau (this volume) develops a higher order state theory of perception and similarly urges the use of subjective rather than objective measures.

In Indian philosophy there has been a debate whether mental states are 'self-luminous' (see Gupta, 2003, pp. 49–55). Self-luminosity implies 'I know' is part of each first-order cognition (a notion in Western philosophy that goes back to Descartes and that has been incorporated in some current higher order theories, e.g. Van Gulick, 2004). On the other hand the Nyāya schools deny self-luminosity; cognitions are followed by higher order cognitions making us aware of the first-order ones.[2] Whichever line of argument one takes, it follows that a conscious state allows the person to know what state they are in.

The alternative to subjective measures is objective measures. Objective measures were promoted in psychology from the 1960s onwards by those people who were skeptical about the existence of unconscious mental states. An objective measure uses the ability of a person to discriminate states of the world (e.g. object moving up or down) to measure whether the mental state is conscious. When people are found able to make such worldly discriminations, the conclusion drawn is that there was conscious knowledge. But worldly discrimination only tests for the existence of first-order states. It is true that a failure to make a worldly discrimination indicates the absence of conscious knowledge, but it also likely indicates absent or at least degraded unconscious knowledge (see Lau, this volume). Relying on objective measures gives a distorted picture of the nature of unconscious mental states.

[2]However, I do not think the Nyāya school provides a higher order state theory of conscious mental states. It is not in virtue of the higher order cognitions that the lower order ones are conscious. In fact, the higher order cognitions simply reveal the conscious nature of the first-order states (see Bhattacharya, 1989, p. 144). Nonetheless, on this approach conscious states enable higher order states, and that is all we need to justify the use of subjective measures. I am assuming that, for the sake of argument, both schools would allow non-conscious states.

The guessing and zero-correlation criteria

Knowledge is typically shown when a person makes a worldly discrimination (like 'the object is going up', 'this sequence is grammatical'). To test for relevant higher order thoughts we can ask the person for their confidence in each such judgment. The simplest confidence scale is just 'guess' if the person believes the judgment had no firm basis whatsoever, and 'know' if the person believes the judgment constituted knowledge to some extent. If on all the trials when the person says 'guess' nonetheless the discrimination performance is above baseline, then there is evidence that the person does have knowledge (performance above baseline) that she does not know she has (she says she is guessing). This is unconscious knowledge by the *guessing criterion*. If a person's knowledge states are conscious, she will know when she knows and when she is just guessing. In this case, there should be a relation between confidence and accuracy. Thus, a relation between confidence and accuracy indicates conscious knowledge and zero relation indicates unconscious knowledge by the *zero-correlation criterion*. Both criteria are illustrated in Fig. 1 (see Dienes, 2004; Dienes and Perner, 2001, 2004, for further discussion).[3]

The criteria can be illustrated with the phenomenon of *implicit learning*, a coin termed by Reber (1967) to indicate the process by which we acquire unconscious knowledge of the structure of an environment. An everyday example is how we learn the grammar of our native language. By age 5 we have learnt the main regularities, but we did not know we were learning them and could not have said what they were. (Indeed, the mechanism is so powerful, it still beats our best attempts at conscious learning: no linguist has produced a complete grammar of any natural language.) Reber investigated implicit learning by

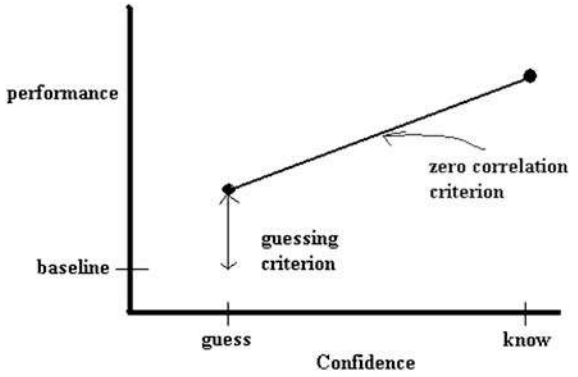


Fig. 1. The guessing and zero-correlation criteria. In the example, the guessing criterion indicates the presence of some unconscious knowledge and the zero-correlation criterion indicates the presence of some conscious knowledge.

constructing artificial grammars, i.e. arbitrary rules for determining the sequences of elements. The elements he used were letters. The strings of letters produced looked more or less random but were in fact structured. Initially, people were asked to just look at, copy down or memorize such strings for a few minutes. Then they were informed that the order of letters within each string was determined by a complex set of rules, and people classified new strings as grammatical or not. Reber found people could classify at above chance levels (typically 65% after a few minutes exposure) while being unable to freely report what the rules of the system were. Reber did not use the guessing or zero-correlation criteria as measures of the conscious status of knowledge; he used free report.

Free report is a type of subjective measure because the person normally has to believe they know something in order to report it. However, critics have been unhappy with free report as an indicator of unconscious knowledge (Berry and Dienes, 1993). Free report gives the subject the option of not stating some knowledge if they choose not to (because they are not certain enough of it); and if the free report is requested some time after the decision, the subject might momentarily forget some of the knowledge. Similarly, what the subject freely reports depends on what sort of response the subject thinks the experimenter wants. For example, if the subject classified on the basis of similarity to memorized exemplars, but

[3]Note that the zero-correlation criterion should be applied by finding the relationship between confidence and accuracy within each subject (enabling one to then test the significance of the relationship over subjects). If one subject contributes just one confidence and accuracy point, the relation between confidence and accuracy may be confounded with personality variables.

thinks the experimenter wants to hear about rules, then free report may not be very informative about the subject's conscious knowledge. That is, a test must tap the knowledge that was in fact responsible for any changes in performance (the *information* criterion of Shanks and St. John, 1994, and the problem of correlated hypotheses highlighted by Dulany, 1968). One way around the information criterion is to use confidence ratings, because then the experimenter does not need to know exactly what the knowledge is that participants use. Any knowledge the participant is conscious of using as knowledge, no matter what its content, should be reflected in the participant's confidence. Further, using confidence ratings has an advantage over free report in that low confidence is no longer a means by which relevant conscious knowledge is excluded from measurement; rather the confidence itself becomes the object of study and can be directly assessed on every trial. Indeed, Ziori and Dienes (2006) provided empirical evidence for the greater sensitivity of confidence-based methods over free report in detecting conscious knowledge. These are major benefits of the use of confidence measures of conscious knowledge.

A further strength of the zero-correlation and guessing criteria is that they do not assume that people have only conscious or only unconscious knowledge in any one condition (a desideratum of tests of conscious and unconscious knowledge repeatedly advocated by Jacoby, e.g. 1991). A guessing criterion analysis indicating the presence of some unconscious knowledge does not rule out the existence of conscious knowledge on other trials. Conversely, a zero-correlation criterion analysis indicating the presence of some conscious knowledge does not rule out the existence of unconscious knowledge in the same trials. If both criteria are statistically significant then there is evidence for both conscious and unconscious knowledge, a typical state of affairs.

In terms of the guessing criterion, Dienes et al. (1995) showed that when people believed they were guessing in the test phase of artificial grammar learning paradigm, they nonetheless classified above baseline levels. These results were replicated by Dienes and Altmann (1997), Tunney and Shanks (2003), Dienes and Perner (2003) and Dienes and Scott (2005); by Dienes and Longuet-Higgins (2004) with musical stimuli; and by Ziori and Dienes (in press) in another concept formation paradigm. In terms of the zero-correlation criterion, Chan (1992) showed subjects were no more confident in correct than incorrect decisions in artificial grammar learning. Typically, though not always, the zero-correlation criterion does indicate the presence of some conscious knowledge in artificial grammar learning. Nonetheless, Dienes et al. (1995), Dienes and Altmann (1997), Allwood et al. (2000), Channon et al. (2002), Tunney and Altmann (2001) and Dienes and Perner (2003) replicated Chan in finding some conditions under which there was no within-subject relationship between confidence and accuracy; as did Dienes and Longuet-Higgins (2004) and Kuhn and Dienes (2006) with musical stimuli. Subjects could not discriminate between mental states providing knowledge and those just corresponding to guessing; hence, there must have existed unconscious mental states. Kelly et al. (2001) and Newell and Bright (2002) used the same lack of relationship between confidence and accuracy to argue for the use of unconscious knowledge in other learning paradigms.

The problem of bias

In order to determine the conscious status of mental states we need to make a distinction between first-order and second-order states. The English language does not respect that distinction very well. If I say 'Bill saw the tree' I usually mean there not only was a first-order seeing of the tree but Bill was also aware of seeing the tree. Similarly, normally 'knowing' means awareness of knowing as well. Now let us try to keep first-order and second-order states conceptually separate, as illustrated in Fig. 2. When a person makes a judgment with a certain content, e.g. 'this string is grammatical', the first-order state itself may be one of guessing, one in which the system itself has no commitment to that particular content. On the other hand, the system may have a lot of commitment to the judgment (because the system

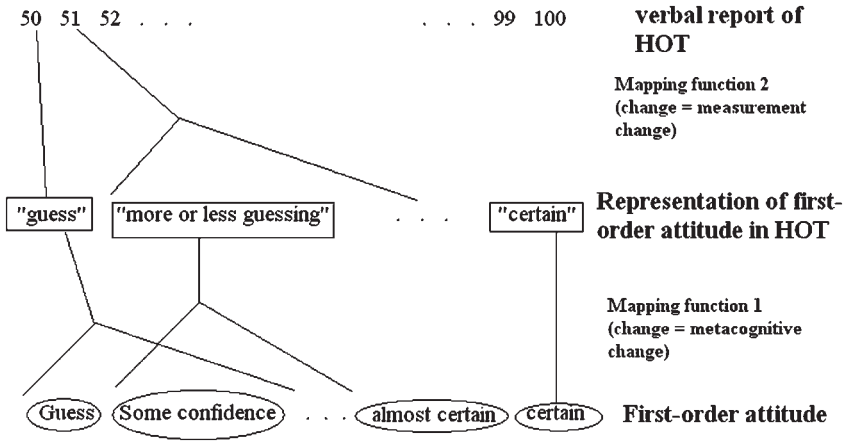


Fig. 2. Relations between first-order states, HOTs and verbal report.

has used a generally reliable method in arriving at it); the commitment may show itself in the consistency with which the same judgment is made on repeated trials, or the amount of counter-evidence or punishment needed to reverse it. In a sense, the system is 'sure', though it does not need to represent itself as being sure in order to be sure. That is, one can have purely first-order states of guessing, being fairly sure or certain. In English, when we say the person is sure we mean also they think they are sure. But in Fig. 2 the terms at the bottom are just meant to refer to first-order states and not imply the person necessarily thinks they are in those states (see Twyman and Dienes, in press).

The first-order states at the bottom of Fig. 2 are represented in higher order thoughts via the first mapping function illustrated. For example, perhaps when the first-order state is *guessing* and also *some confidence*, the person represents herself as just guessing. That is, there is bias in the mapping from first-order states to HOTs. This is the bias that allows unconscious mental knowledge to exist at all (Dienes, 2004); it is bias that researchers interested in unconscious states want to happen. It is precisely when the person is in a first-order state of some confidence (i.e. based on a generally reliable learning mechanism) but represents herself as guessing that we have unconscious knowledge. The manipulations that affect this mapping (motivations, rewards, types of structures to be learnt, conscious distractions, feedback on accuracy, etc.) are the manipulations that affect amount of unconscious knowledge. Establishing the effect of such manipulations is one of the major tasks of research into unconscious knowledge.

The experimenter asks the subject to express their HOTs on a confidence scale, for example a percentage scale. For a binary first-order judgment (e.g. ''this sequence is grammatical''/''this sequence is non-grammatical''), one could use a 50–100% confidence scale. The person gives a number between 50 and 100%. If they say 50%, it means they expect to get 50% of such answers correct, they could have well just flipped a coin. If they felt they knew to some extent, they could give a number to reflect that fact; for example 54%, meaning the person expects to get 54% of such judgments correct. And if the person was completely certain they could give 100%. Figure 2 illustrates a possible mapping from the thoughts the person actually has to the form of verbal expression allowed by the experimenter, in this case a 50–100% confidence scale (mapping function 2). Unlike the bias in the first mapping function, bias in the second mapping function is undesirable. As illustrated in Fig. 3, it is possible that when people *say* they are guessing they actually *think* they have some knowledge. The knowledge demonstrated when people say they are guessing may all be due to those cases where the person thinks they do have

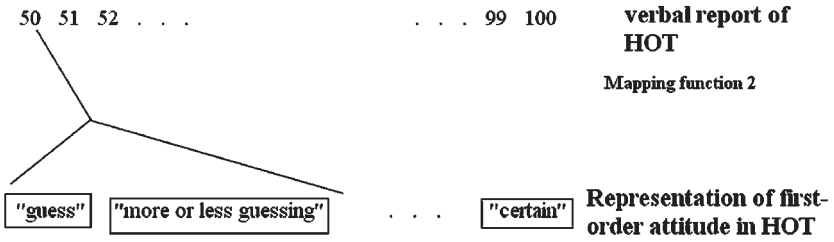


Fig. 3. The problem of bias.

some knowledge. Then, despite what the guessing criterion seems to indicate, there would not be any unconscious knowledge.

The problem of validity is faced by all tests in psychology: does the test measure what it says it does? Even if we solved this problem with certainty for the guessing criterion, there is also a more general point: scientists seek not to *classify* the world according to a priori criteria but to *identify* interesting kinds in nature. In the same way, the ultimate aim of the guessing criterion is not to classify knowledge according to an a priori notion of what is unconscious, but to identify an interesting kind in nature, namely, we speculate, unconscious knowledge. The evidence that it does so is provided by evidence that the criterion is useful in separating qualitatively different types of knowledge (the conscious and the unconscious) that differ specifically in ways predicted by interesting theories of the difference between conscious and unconscious. The guessing criterion's long-term ability to do this is the evidence that it measures what it says it does and, more importantly, that it picks out a kind in nature that is worth studying. Notice its validation depends on its being used with substantial theories. One cannot establish whether or not there is unconscious knowledge as an isolated question in itself separate from a theoretical research program.

One common theory of conscious knowledge is that it relies on a frontal working memory system for formulating and testing hypothesis and drawing inferences, but that unconscious (implicit) learning does not. Unconscious learning, it is proposed, involves changes in the weights of neural networks that happen by automatic learning rules, requiring only minimal levels of attention to the stimuli (Berry and Dienes, 1993; Cleeremans, this volume). If so, then loading working memory with a difficult task should interfere with the application of conscious knowledge but not unconscious knowledge. One task for loading working memory is random number generation: producing a digit (0..9), one every second or two, such that the sequence is random is very consciously demanding. Dienes et al. (1995) found that random number generation during the test phase of an artificial grammar learning task interfered with the accuracy of classification when people had some confidence but not with accuracy when people believed they were guessing. There is no reason why a failure to verbalize a HOT should render knowledge resilient to the effects of a demanding secondary task; but the results are consistent with the claim that people had unconscious knowledge, and its application does not rely on working memory. On another concept formation task, Ziori and Dienes (in press) also found performance associated with guess responses was resilient to a demanding secondary task. The effects of secondary tasks on performance on implicit learning tasks more generally is variable (see Jiménez, 2003); perhaps, the effects would be clarified if subjective measures were used to separate conscious and unconscious knowledge, which has rarely been done to date (for a refinement of this claim, see the section Judgment versus Structural Knowledge).

Merikle (1992) reviewed evidence that the guessing criterion picks out a qualitatively different type of knowledge in perception as well. For quickly flashed stimuli, people based plans for action on stimuli they say they saw, but not on the presence of stimuli they say they just guessed at.

This provides some evidence that when people said they were guessing they also *thought* they were guessing and were not just saying it. Knowledge associated with 'guess' responses was not inferentially promiscuous, people were unwilling to base actions on it.

Skeptics of the existence of unconscious knowledge point out that unconscious rather than conscious knowledge is often associated with lower performance and so seeming qualitative differences between 'conscious' and 'unconscious' knowledge may arise from the scale effects of having different amounts of conscious knowledge (Holender, 1986). Lau (this volume) describes a perception experiment in which overall detection performance was equalized between two conditions that differed in terms of the proportion of times it seemed to the subject they saw anything. In this situation the quality of knowledge is the same for conscious and unconscious cases, an ideal method for future work exploring their qualitative differences without the confound of differing performance levels. Interestingly, in fMRI the two conditions differed in the activation of only the dorsolateral prefrontal cortex. In contrast, Spence et al. (2001) found that subjects asked to lie showed increased activation in the ventrolateral prefrontal cortex, as well as many other areas (but only minimal activation of the dorsolateral prefrontal cortex). Thus the subjects in Lau's experiment were unlikely to have differed in the extent to which their words were true to their thoughts across the two conditions. That is, there does not seem to have been a bias problem in Lau's experiment.

The above arguments suggest that the guessing criterion often does track what it says it tracks: unconscious knowledge. The bias problem, while a possible problem, is not necessarily an actual problem for the guessing criterion. That is not to say the guessing criterion will always track unconscious knowledge reliably; the conditions under which it does so is a substantial problem for future research.

The zero-correlation criterion can escape the bias problem, as illustrated in Fig. 4. In this example the confidence rating is just a 'guess' or

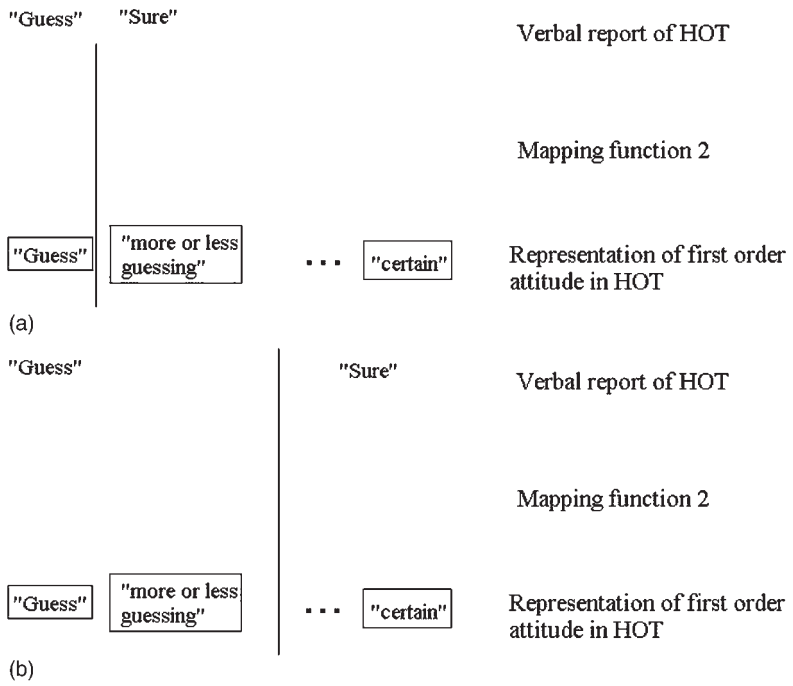


Fig. 4. Zero-correlation criterion insensitive to criterion placement. (a) Good placement of criterion; (b) sloppy placement of criterion.

'sure' response. In Fig. 4a, the criterion separating the *verbal* response 'guess' from 'sure' occurs at the boundary between the *thought* that one is guessing and the thought that one is only more or less guessing. If the HOTs are accurate there will be a relation between confidence and accuracy. Figure 4b shows the subject has a bias that would be problematic for the guessing criterion — the criterion placement means the 'guess' verbal response includes some HOTs where the person thinks they know to some extent. BUT note if the HOTs are accurate, there will still be a relation between confidence and accuracy. The change in bias does not affect the use of the zero-correlation criterion for indicating conscious or unconscious knowledge.

One way of measuring the zero-correlation criterion is the Chan difference score (Chan, 1992; Dienes et al., 1995; Ziori and Dienes, in press). For a binary confidence rating, this is the difference in the proportion of 'sure' responses when correct and when incorrect. The proportion of 'sure' responses when correct is called a *hit* in type 2 signal detection theory (STD) and the proportion of 'sure' responses when incorrect is called a *false alarm*. The Chan difference score is hence equal to hits minus false alarms, which is the commonest way of dealing with the possibility of bias in memory research.

The slope in Fig. 1 is similar to the Chan difference score but it conditionalizes the other way round: *P*(correct/'know') minus *P*(correct/'guess'). This may be a better way of dealing with bias in artificial grammar learning because the slope is undefined if the subject uses only one confidence category (possibly indicating an extreme bias).

If we define *misses* as (1−hits) and *correct rejections* as (1−false alarms), then the following quantity

$$\mathrm{Ln}\left(\frac{\text{hits} \times \text{correct rejections}}{\text{false alarms} \times \text{misses}}\right)$$

also gives a measure of the relation between confidence and accuracy controlling for bias. If it is scaled by the factor $\sqrt{3}/\pi$, it is called (logisitic) d'("d prime"). Tunney and Shanks (2003) implemented the zero-correlation criterion with d'.

In sum, the various ways of implementing the zero-correlation criterion allow the bias problem to be addressed.

The relative insensitivity of the zero-correlation criterion to bias does not imply it is better than the guessing criterion or that it should replace it. Typically, subjects presumably develop both conscious and unconscious knowledge and the use of both criteria is useful for picking these out. As mentioned earlier, the proof of the usefulness of the criteria is in their heuristic value and this has scarcely been tested yet. Also, the interpretation of the zero-correlation criterion depends on one's model of underlying processes. A lack of relation between confidence and accuracy does not automatically mean all the knowledge is unconscious. All knowledge may be unconscious if people, not being aware of any knowledge, but not wanting to give just one confidence response all the time, chose 'guess' and 'know' responses randomly. Or it may be that unconscious and conscious knowledge have the same accuracy, and the unconscious knowledge expresses itself in the 'guess' responses and the conscious knowledge expresses itself in the 'sure' responses. Indeed, if the unconscious knowledge were superior to the conscious knowledge, there may be a negative relation between confidence and accuracy. (We are currently working on a paradigm for finding just this outcome.) The criteria are not operational definitions in the literal sense of defining; they are just tools and like any tool must be used with intelligence and sensitivity in each application.

Types of confidence scales

Now we come to a practical matter. Does it matter what sort of confidence scale one uses? Tunney and Shanks (2003) compared a binary scale (high vs. low) with the 50–100% scale with a particularly difficult type of artificial grammar learning task (classification performance $\sim$55%). They found the binary scale indicated a relation between confidence and accuracy where the 50–100% scale did not, so the binary scale was more sensitive

(even after a median split on the continuous scale was used to make it binary for the purposes of analysis). Tunney (2005) obtained the same result with another artificial grammar learning task (again a difficult one with performance around 55%). The result is surprising: one would think giving people more categories than two would focus their mind on finer distinctions. On the other hand, presumably HOTs are not typically expressed as numbers. The person may think something like 'I am more or less guessing' and the process by which this is converted into a number for the experimenter may be more variable and noisy than the process of converting it into everyday words. Still, it remains an open question whether the type of scale used does make any consistent difference to the sensitivity of the zero-correlation criterion.

In an ongoing study, I have asked different groups of subjects to express their confidence in one of six different scales: binary (high vs. low); binary (guess vs. sure: more useful than high vs. low because it asks the subject to put the divide where we want it); numerical 50–100%, any number in the range allowed; the same again, 50–100% but with detailed explanation of what the numbers mean (i.e. as was explained above, they are expected performances); numerical categories (50, 51–59, 60–69, …., 90–99, 100); and verbal categories (complete guess, more or less guessing, somewhat sure, fairly sure, quite sure, almost certain, certain).

The first study used the same difficult materials as Tunney and Shanks (2003). The relation between confidence and accuracy was expressed in different ways: Chan difference score, d' and a correlation measure much favoured by psychologists interested in metacognition, gamma (used in Kuhn and Dienes, 2006, in an implicit learning context). The precise measure used did not affect the results at all. The results for gamma are shown in Fig. 5.

There is no indication that any scale was more sensitive than any other (perhaps surprisingly; I was betting on the verbal categories being most

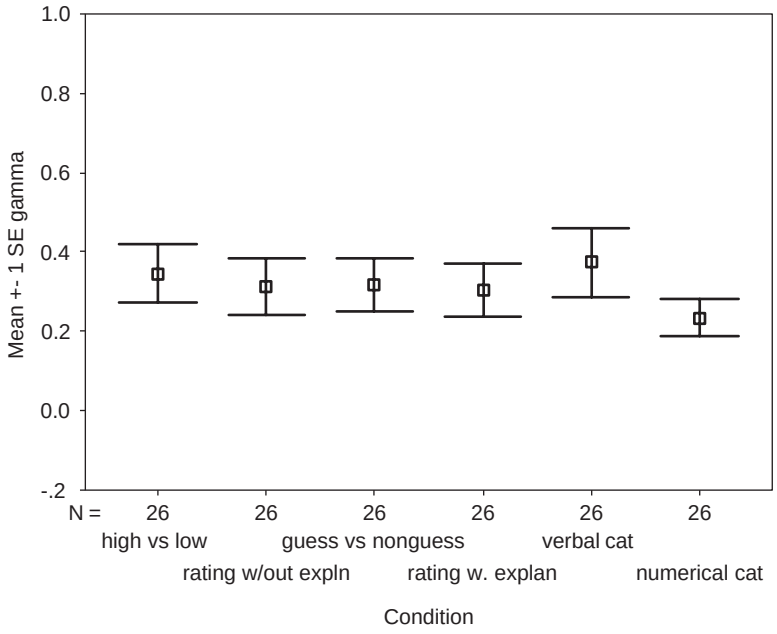


Fig. 5. Zero-correlation criterion with different confidence scales: difficult artificial grammar learning task (overall classification performance 54%). Small squares indicate the mean for each condition, and the lines go out one standard error (SE) either side. The conditions refer to different confidence scales. For example "high versus low" is a binary scale with values 'high confidence' and 'low confidence'. See text for full explanation.

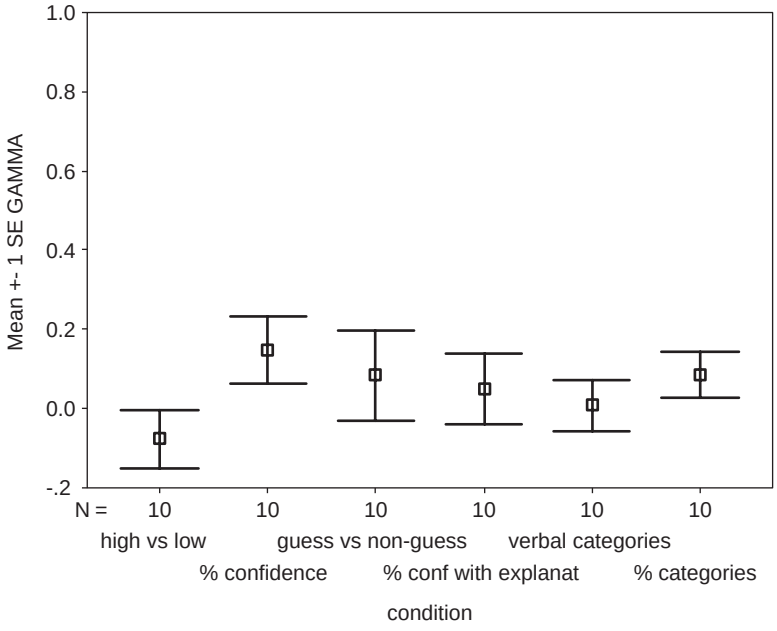


Fig. 6. Zero-correlation criterion with different confidence scales: easier artificial grammar learning task (overall classification performance 61%). Small squares indicate the mean for each condition, and the lines go out one standard error (SE) either side. The conditions refer to different confidence scales. For example "high versus low" is a binary scale with values 'high confidence' and 'low confidence'. See text for full explanation.

sensitive). Maybe the type of scale does not consistently make a major difference in this situation, but when people have more knowledge overall then surely more fine-grained scales would show their greater sensitivity. The next study, still in progress, used materials typically leading to classification performance around 65%. The results for gamma are shown in Fig. 6.

A one-way omnibus ANOVA comparing gamma across conditions was non-significant.[4] In sum, the evidence to date does not definitively indicate one type of confidence scale is consistently more sensitive than any other overall.

A rather different type of scale was introduced by Persaud et al. (in press). On each trial subjects chose to wager either a small amount (one pound) or a large amount of money (two pounds). If they got the trial right they received the sum wagered and if they got it wrong they lost the sum. Optimally, if one had any confidence at all one should go with the large wager. In an artificial grammar learning task, the percentage of correct decisions when a small wager was chosen was 77%, significantly above chance, indicating unconscious knowledge by the guessing criterion. The probability of a high wager after a correct decision was higher than after an incorrect decision (a Chan difference score), indicating the presence of some conscious knowledge by the zero-correlation criterion. The relation between wagering (putting your money where your mouth is) and verbal confidence ratings is an issue Persaud, Lau and myself have just started to explore.

[4]However, an uncorrected *t*-test comparing just the two conditions Tunney compared (high vs. low and percentage confidence) is significant. The percentage confidence scale was more sensitive than the high/low scale.

Judgment versus structural knowledge

When a person is exposed to strings from an artificial grammar, she learns about the structure

of the strings. Call this knowledge *structural knowledge*. It might consist of the knowledge that an M can start a string, about whole strings that were presented, about what letters can repeat, and so on. In the test phase the structural knowledge is brought to bear on a test item to form a new piece of knowledge: the judgment, for example that this string is grammatical. Call this knowledge *judgment knowledge*. When confidence ratings are taken, the confidence is confidence in the judgment; hence confidence ratings test for HOTs about judgment knowledge. The guessing and zero-correlation criteria test the conscious status of judgment knowledge only; that is their job. They do this job very well, but sometimes people criticize them for not testing the conscious status of structural knowledge. They say, 'But surely unconscious knowledge might be influencing confidence ratings, so they are not a good measure of conscious knowledge' (Allwood et al., 2000). To reword the criticism with our new concepts, it states that structural knowledge may be unconscious when judgment knowledge is conscious. This is true, but not a criticism of the guessing and zero-correlation criteria.

Consider natural language. You can tell of a sentence in your native tongue whether it is grammatical or not, be reliably right in your judgment, and be confident that you are right. You have conscious judgment knowledge. But your structural knowledge is almost entirely unconscious (try explaining to a second language learner why your version of what they were trying to say is better). When structural knowledge is unconscious and judgment knowledge is conscious the phenomenology is that of intuition. When both structural knowledge and judgment knowledge is unconscious it just feels like guessing. When people are using intuition, there may be no unconscious knowledge by the guessing criterion and a strong relation between confidence and accuracy indicating judgment knowledge is all conscious.

Dienes and Scott (2005) employed a simple way of testing for the conscious status of structural knowledge as well as judgment knowledge. In the test phase of an artificial grammar learning task, after each classification decision, as well as giving a confidence rating, subjects ticked one of four boxes to indicate the basis of their judgment: pure guess, intuition, a rule or rules they could state or memory for part or all of a training string. The guess and intuition attributions are prima facie cases of unconscious structural knowledge and rules and memory attributions cases of conscious structural knowledge. To check subjects were identifying useful internal *kinds* by ticking boxes, two manipulations were included. First, half the subjects were informed of the rules and asked to search for them in the training phase and the other half were, as normal, not informed about rules (the first group should acquire more conscious structural knowledge than the second group); and half the subjects generated random numbers in the test phase and the other half classified with full attention (the secondary task should interfere with the application of conscious structural knowledge). Dienes and Scott found that people used the four attributions about equally often, but used the conscious structural knowledge attributions more when they had been asked to search for rules rather than just memorize; and less when they generated random numbers at test (just as one would expect). Further, the level of classification performance was above baseline for each of the attributions. There was prima facie evidence of simultaneous unconscious structural and judgment knowledge (guess attributions); unconscious structural knowledge with conscious judgment knowledge (intuition attributions; the zero correlation criterion also indicated conscious judgment knowledge in this case); and finally both conscious structural and judgment knowledge (rules and memory attributions).

Possibly measurements of the conscious status of judgment knowledge have indicated dissociations in the past because the conscious status of judgment knowledge is often confounded with the conscious status of structural knowledge. Indeed, comparing intuition with guess attributions (conscious status of structural knowledge constant, conscious status of judgment knowledge differs) revealed no differential effect of the manipulations. However, the division of knowledge into conscious and unconscious structural knowledge was relevant. Unconscious structural knowledge (classification performance based on guess and

intuition attributions) was unaffected by the manipulations; but conscious structural knowledge (rules and memory attributions) was harmed by a secondary task after searching for rules. The relevant distinction for capturing a kind in nature seemed to be the difference between conscious and unconscious structural knowledge, not conscious and unconscious judgment knowledge.

Scott and Dienes (submitted) drew a similar conclusion. They found that continuous ratings of the familiarity of an item (hypothesized to reflect the continuous output of the neural network responsible for learning the grammar) predicted grammaticality classification for all structural knowledge attributions. However, when subjects searched for rules, the actual grammaticality of an item had additional predictive power above that of familiarity for only the conscious structural knowledge attributions, indicating they involved an additional source of knowledge. Again the joint in nature appeared to be between conscious and unconscious structural knowledge, not conscious and unconscious judgment knowledge.

Other recent work in my lab indicates the importance of the conscious status of structural knowledge. Riccardo Pedersini, following a study by Bierman et al. (2005), rewarded or punished subjects in an artificial grammar learning task after they made correct or incorrect choices. On each trial a test string was shown and skin conductance was recorded for 3 s before subjects made a response. As shown in Fig. 7, and replicating Bierman et al., skin conductance was higher for incorrect than correct choices. Somehow the subjects knew when they were getting it wrong, and this created arousal, increased sweating and

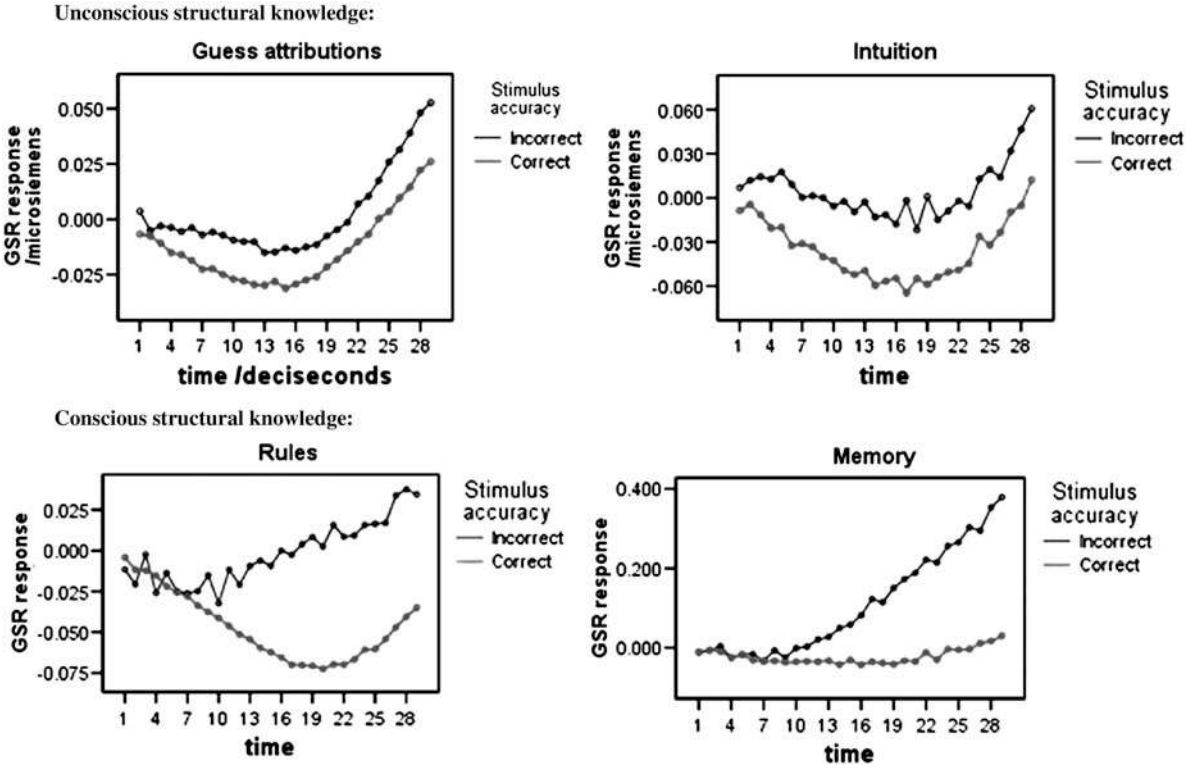


Fig. 7. Galvanic skin response (GSR) for 3 s after a test stimulus is presented. The subject responded just as the graph finishes. The GSR measures how much the subject is sweating, i.e. arousal. The graphs show a greater GSR to incorrect than correct responses; subjects had some knowledge of when they were right, even when they thought they were guessing. (See Color Plate 5.7 in color plate section.)

hence a higher skin conductance. Pedersini also asked on each trial for a structural knowledge attribution. Interestingly, even when subjects thought they were guessing, their skin conductance revealed they knew when they correct or incorrect. A striking finding shown in Fig. 7 is that for unconscious structural knowledge attributions (guess and intuition), the skin conductance separated correct from incorrect responses within the first second after a test string was shown (10 ds shown on the axes); by contrast, when structural knowledge was conscious (rules or memory), a full second was needed before correct and incorrect responses separated. In terms of this time course, the relevant distinction is between conscious and unconscious structural knowledge not conscious and unconscious judgment knowledge (guess and intuition behaved similarly). The finding is consistent with unconscious structural knowledge being embedded in the weights of a neural network so the knowledge is applied in 'one time step' as activation flows through the network. By contrast, application of rules and recollection often require multiple processing steps.

In sum, the interesting distinction in implicit learning paradigms may be between conscious and unconscious structural knowledge. This needs further testing and currently stands as only a hypothesis. Implicit learning research should habitually take both confidence ratings and structural knowledge attributions when first-order judgments are made. In perception, it goes without saying that structural knowledge is unconscious. The useful dividing line there — between conscious perception and subliminal perception — is between conscious and unconscious judgment knowledge.

Controlling the use of knowledge

Destrebecqz and Cleeremans (2001, 2003) used people's ability to control the use of their knowledge as a measure of its conscious status, a method developed for general use in perception and memory research by Jacoby (1991). Destrebecqz and Cleeremans used an implicit learning paradigm called the serial reaction time (SRT) task. One of four locations on a computer is indicated on each trial; the subject responds by pressing a corresponding key. From the subject's point of view it is a straightforward reaction time task. But unbeknownst to the subject the order of locations is structured; we know people learn this structure because they come to respond faster when the sequence follows the structure than when it violates it. But is the knowledge conscious or unconscious? Destrebecqz and Cleeremans asked subjects to generate a sequence. Following Jacoby's methodology, there were two conditions: subjects either tried to generate the same sequence they had been trained on as best they could (the so-called *inclusion* condition, subjects aim to include the sequence) or to make sure they did not generate that sequence (the *exclusion* condition). When people were trying to generate the sequence (inclusion), they could do so to some extent. But both conscious and unconscious knowledge would enable this. The key finding was that when people were trying not to generate the sequence (i.e. in the exclusion condition) they nonetheless generated the sequence at above baseline levels. As consciously knowing the sequence would lead one to perform below baseline, Destrebecqz and Cleeremans concluded subjects had acquired unconscious knowledge. Further, they showed that above baseline exclusion was associated with rapid trials; when subjects could take their time, subjects excluded more effectively. With slow trials, there was a clear difference between the extent to which the sequence was generated in inclusion and exclusion. The latter results are consistent with the claim that conscious knowledge takes time to apply.

In the exclusion task, so long as subjects have keyed in to the structural knowledge, it will make them tend to generate grammatical continuations. Now they need to make a judgment before they press the key: do they know it is grammatical or is it just a random guess? If it seems like a random guess they can go ahead and press the key. If they believe it is the product of knowledge they should withhold the response and choose another. In other words, the exclusion task is an intuitively good measure of the conscious status of knowledge because it relies on a covert assessment by the subject of whether they know. If exclusion were

not controlled by HOTs, it would loose its face validity. Logically, subjects could exclude simply on the basis of pure guesses; but it would be strange to conclude from such successful exclusion that it indicated conscious knowledge when the subject denies having any knowledge whatsoever. (Indeed, Dienes et al. (1995) found that subjects by purely guessing could choose to exclude the use of one grammar and apply another in classifying strings in an artificial grammar learning experiment. Here exclusion was based on unconscious judgment knowledge.)

In excluding based on conscious knowledge, all that is required is consciously knowing whether this continuation is grammatical or not. That is, when subjects are instructed to base exclusion on conscious knowledge, exclusion only requires conscious judgment knowledge. The subject does not need to know why the continuation is grammatical. Below baseline exclusion performance is prima facie evidence of conscious judgment knowledge, but is mute about whether structural knowledge is also conscious.

Fu et al. (in press) replicated the Destrebecqz and Cleeremans finding of above baseline exclusion knowledge, showing it was particularly likely early in training and with statistically noisy training sequences. Conscious judgment knowledge was shown by the difference in performance between inclusion and exclusion conditions. Fu et al. showed that when people made guess attributions there was no difference between inclusion and exclusion. Both measures (control and verbal attribution) agreed in showing no conscious judgment knowledge. With intuition attributions, however, there was a difference between inclusion and exclusion, indicating unconscious structural knowledge with conscious judgment knowledge. We are currently using this paradigm to explore further the qualitative differences between conscious and unconscious knowledge.

Conclusion

For many decades research into the distinction between conscious and unconscious knowledge was regarded with suspicion. William James regarded the field of the unconscious as a 'tumbling ground for whimsies'. As late as 1994 when I gave a talk in my department on the distinction between conscious and unconscious knowledge, afterwards a colleague told me he really liked the talk, but he wondered if I could give it without referring to consciousness. Finally things have changed and it is OK to address what must be one of the most important problems in understanding minds. Please, come and have a tumble. And I urge you to seriously consider using subjective measures – despite more than 100 years of research it seems we have barely started in seeing how useful they might be.

References

Allwood, C.M., Granhag, P.A. and Johansson, H. (2000) Realism in confidence judgements of performance based on implicit learning. Eur. J. Cogn. Psychol., 12: 165–188.

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

Berry, D.C. and Dienes, Z. (1993) Implicit Learning: Theoretical and Empirical Issues. Lawrence Erlbaum, Hove.

Bhattacharya, G. (1989) Tarkasamgraha-Dipka on Tarkasamgraha. Progressive Publishers, Kolkata.

Bierman, D., Destrebecqz, A. and Cleeremans, A. (2005) Intuitive decision making in complex situations: somatic markers in an implicit artificial grammar learning task. Cogn. Affect. Behav. Neurosci., 5: 297–305.

Carruthers, P. (2000) Phenomenal Consciousness: A Naturalistic Theory. Cambridge University Press, Cambridge.

Chan, C. (1992) Implicit cognitive processes: theoretical issues and applications in computer systems design. Unpublished DPhil thesis, University of Oxford.

Channon, S., Shanks, D., Johnstone, T., Vakili, K., Chin, J. and Sinclair, E. (2002) Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. Neuropsychologia, 40: 2185–2197.

Destrebecqz, A. and Cleeremans, A. (2001) Can sequence learning be implicit? New evidence with the Process Dissociation Procedure. Psychon. Bull. Rev., 8: 343–350.

Destrebecqz, A. and Cleeremans, A. (2003) Temporal effects in sequence learning. In: Jiménez L. (Ed.), Attention and Implicit Learning. John Benjamins Publishing Company, Amsterdam, pp. 181–213.

Dienes, Z. (2004) Assumptions of subjective measures of unconscious mental states: higher order thoughts and bias. J. Conscious. Stud., 11: 25–45.

Dienes, Z. and Altmann, G.T.M. (1997) Transfer of implicit knowledge across domains? How implicit and how abstract? In: Berry D. (Ed.), How Implicit is Implicit Learning? Oxford University Press, Oxford, pp. 107–123.

Dienes, Z., Altmann, G.T.M., Kwan, L. and Goode, A. (1995) Unconscious knowledge of artificial grammars is applied strategically. J. Exp. Psychol. Learn. Mem. Cogn., 21: 1322–1338.

Dienes, Z. and Longuet-Higgins, H.C. (2004) Can musical transformations be implicitly learned? Cogn. Sci., 28: 531–558.

Dienes, Z. and Perner, J. (2001) When knowledge is unconscious because of conscious knowledge and vice versa. Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society, 1–4 August, Edinburgh, Scotland. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 255–260.

Dienes, Z. and Perner, J. (2003) Unifying consciousness with explicit knowledge. In: Cleeremans A. (Ed.), The Unity of Consciousness: Binding, Integration, and Dissociation. Oxford University Press, Oxford, pp. 214–232.

Dienes, Z. and Perner, J. (2004) Assumptions of a subjective measure of consciousness: three mappings. In: Gennaro R.J. (Ed.), Higher Order Theories of Consciousness. John Benjamins Publishers, Amsterdam, pp. 173–199.

Dienes, Z. and Scott, R. (2005) Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. Psychol. Res., 69: 338–351.

Dulany, D. (1968) Awareness, rules, and propositional control: a confrontation with S-R behavior theory. In: Dixon T. and Horton D. (Eds.), Verbal Behavior and General Behavior Theory. Prentice-Hall, Englewood Cliffs, NJ, pp. 340–387.

Fu, Q., Fu, X. and Dienes, Z. (in press) Implicit sequence learning and conscious awareness. Conscious. Cogn.

Gaillard, V., Vandenberghe, M., Destrebecqz, A. and Cleeremans, A. (2006) First- and third-person approaches in implicit learning research. Conscious. Cogn., 15: 709–722.

Gupta, B. (2003) Cit consciousness. Oxford University Press, Oxford.

Holender, D. (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: a survey and appraisal. Behav. Brain Sci., 9: 1–23.

Jacoby, L.L. (1991) A process dissociation framework: separating automatic from intentional uses of memory. J. Mem. Lang., 30: 513–541.

Jiménez, L. (Ed.) (2003) Attention and Implicit Learning. John Benjamins, Amsterdam.

Kelly, S.W., Burton, A.M., Kato, T. and Akamatsu, S. (2001) Incidental learning of real world regularities in Britain and Japan. Psychol. Sci., 12: 86–89.

Kuhn, G. and Dienes, Z. (2006) Differences in the types of musical regularities learnt in incidental and intentional learning conditions. Q. J. Exp. Psychol., 59: 1725–1744.

Merikle, P.M. (1992) Perception without awareness: critical issues. Am. Psychol., 47: 792–795.

Newell, B.R. and Bright, J.E.H. (2002) Well past midnight: calling time on implicit invariant learning? Eur. J. Cogn. Psychol., 14: 185–205.

Persaud, N., McLeod, P. and Cowey, A. (2007) Post-decision wagering objectively measures awareness. Nat. Neurosci., 10: 257–261.

Reber, A.S. (1967) Implicit learning of artificial grammars. J. Verbal Learn. Verbal Behav., 6: 317–327.

Rosenthal, D.M. (2005) Consciousness and Mind. Oxford University Press, Oxford.

Scott, R. and Dienes, Z. (submitted) The conscious, the unconscious, and familiarity.

Shanks, D.R. and St. John, M.F. (1994) Characteristics of dissociable human learning systems. Behav. Brain Sci., 17: 367–447.

Sidis, B. (1898) The Psychology of Suggestion. Appleton and Company, New York.

Spence, S.A., Farrow, T.F.D., Herford, A.E., Wilkinson, I.D., Zheng, Y. and Woodruff, P.W. (2001) Behavioural and functional anatomical correlates of deception in humans. NeuroReport, 12: 2849–2853.

Tunney, R.J. (2005) Sources of confidence judgments in implicit cognition. Psychon. Bull. Rev., 12: 367–373.

Tunney, R.J. and Altmann, G.T.M. (2001) Two modes of transfer in artificial grammar learning. J. Exp. Psychol. Learn. Mem. Cogn., 27: 1322–1333.

Tunney, R.J. and Shanks, D.R. (2003) Does opposition logic provide evidence for conscious and unconscious processes in artificial grammar learning? Conscious. Cogn., 12: 201–218.

Twyman, M. and Dienes, Z. (in press) Are subjective measures biased? Conscious. Cogn.

Van Gulick, R. (2004) Higher-order global states (HOGS): an alternative higher-order model of consciousness. In: Gennaro R.J. (Ed.), Higher-Order Theories of Consciousness: An Anthology. John Benjamins, Amsterdam.

Weiskrantz, L. (1997) Consciousness Lost and Found. Oxford University Press, Oxford.

Ziori, E. and Dienes, Z. (2006) Subjective measures of unconscious knowledge of concepts. Mind Soc., 5: 105–122.

Ziori, E. and Dienes, Z. (in press) How does prior knowledge affect implicit and explicit concept learning? Q. J. Exp. Psychol.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 6

Interdependence of attention and consciousness

Narayanan Srinivasan*

Centre for Behavioural and Cognitive Sciences, University of Allahabad, Allahabad 211002, India

Abstract: Research on attention has been closely linked with possible advances in the study of consciousness. Various theories and models have been proposed for attention in the past 50 years. Behavioural, computational, and neuroscientfic approaches have been successful in improving our understanding of attentional processes. Given the current status of attention research, what can we say about the relationship between attention and consciousness? This paper discusses the possible relationships between attention and consciousness. Findings from cognitive science and neuroscience relevant to the elucidation of this relationship are discussed. Recent findings from phenomena that have a bearing on this relationship such as inattentional amnesia, change blindness, attentional blink, perceptual stabilization, and afterimages are described. The implications of the results of these phenomena for attention and awareness are also discussed. It is proposed that top-down attention is not a unitary phenomena and such a characterization may provide a way to interpret some of the results from these findings.

Keywords: attention; consciousness; perception; selective attention; change blindness; attentional blink; inattentional blindness

Introduction

Attention is not a unitary process. Different types of attentional processes include processes like selective attention and vigilance. Selective attention itself is thought of in terms of the basis on which selection is made (location, feature, or object). The most typical way attention is visualized is as a process of selecting information from the visual field for further processing. Various metaphors have been proposed to describe this selective nature of processing including spotlight (Posner, 1980) and zoom lens (Eriksen and Yeh, 1985). Directing attention to a particular location or object typically enhances information processing at that location or for that object. The changes in processing due to attentional processes may be accompanied by eye movements which is called overt attention or may not involve any eye movements which is called covert attention. Cueing paradigms are commonly used in studies based on orienting of attention or spatial attention (Posner, 1980; Posner and Cohen, 1984). A typical cuing paradigm involves participants fixating on a central point and then directing attention to either the left visual field (LVF) or right visual field (RVF) when the central fixation is replaced by left or right directing cues. The participants respond to the targets presented at the cued or uncued locations. There are two types of cueing: exogenous and endogenous. Exogenous cueing or reflexive orienting is involuntary in nature and depends on the properties of objects in space.

*Corresponding author. Tel.: 0532-2460738;
Fax: 0532-2460738; E-mail: ammuns@yahoo.com

DOI: 10.1016/S0079-6123(07)68006-6

Exogenous cueing effects reveal themselves even when reflexive eye-movements are suppressed. Attention moves reflexively to the place of onset or offset of the stimulus in space over time. Many studies have shown that a valid peripheral cue facilitates target detection. The cueing effect is found to be large when the stimulus onset asynchrony is around 100–200 ms (Muller and Rabbitt, 1989). At large SOAs, performance at invalidly cued locations is better than validly cued locations, which has been termed the inhibition of return (Posner and Cohen, 1984). Unlike exogenous or peripheral cuing, typical endogenous cuing utilizes a symbolic cue and attention is shifted voluntarily to the cued location. While voluntary attention results in better performance, it also differs from the processes involved in stimulus-driven attention.

A way in which attentional processes have been characterized is the stage at which selection occurs. Early selection theories proposed by Broadbent (1958) argue that selection occurs at an early stage in perceptual processing. Late selection theories argue that selection occurs after identification of stimuli and is usually thought of in terms of response selection (Deutsch and Deutsch, 1963). Intermediate views on the stage at which selection occurs have also been proposed (Treisman, 1960). Both behavioral and electrophysiological evidence have been used to argue for or against these different views on attention. The majority of attention theorists seem to tend towards early selection theories while accepting that selection occurs at late stages as well including at the level of a response. This chapter mainly focuses on selective aspects of attention.

Studies based on visual search by Triesman and others (Treisman and Gelade, 1980) has led to a two-stage model consisting of a preattentive stage and an attentive stage. Preattentive processing can be defined as quick and basic feature analysis of the visual field, on which attention can subsequently operate. These basic featural computations are combined through the process of spatial attention which is called the binding problem (Treisman and Gelade, 1980). Binding is also necessary for having a unitary conscious experience pointing to the strong relationship between attention and consciousness.

Similar to research on the mechanisms involved in attention, consciousness research has been flourishing in the past 20 years. Reflecting the different views on consciousness, a number of definitions of consciousness have been proposed. Typically awareness is considered a main characteristic of consciousness and it is awareness that will be mainly discussed in this chapter. This includes subjective awareness of the environment (external) and of one's own mental processes (internal). Different characterizations or distinctions have been proposed for studying consciousness. One way in which consciousness has been characterized is in terms of primary consciousness and access consciousness (Block, 2005). Primary consciousness refers to the phenomenal aspects of experience, i.e. qualia. Access consciousness refers to the functional aspects of consciousness which is related to cognitive processes like executive attention, planning, and voluntary control. Essential aspects of consciousness as awareness include its subjective nature and reportability. In most of the studies, the report of participants as "aware" of a particular stimulus is taken as evidence for participant's awareness of the stimulus.

Different views have been proposed on consciousness in terms of its psychological as well as neural underpinnings. For example, consciousness has been visualized as a global work space (Baars, 1997, 2005). According to global workspace theory, unconscious systems process information in parallel and these are made available for access to all the processing systems through the global workspace or a theater. Some have argued against the notion of a Cartesian theater (Dennett, 1991). Dennett (1991) has proposed a multiple drafts model of consciousness in which there is no specific place where everything comes together. On a related line, Zeki (2003) has argued that consciousness of a particular feature is dependent on activation of the extra striate area that specializes in the processing of that feature. Activation of an area will result in a "microconsciousness" of that feature.

One approach to study consciousness has been to look for the neural correlates of consciousness (NCC) by Koch and his colleagues (Crick and Koch, 2003; Koch, 2004). NCC has been defined as the "minimal set of neuronal mechanisms or

events jointly sufficient for any one specific conscious percept or experience'' (Koch, 2004). NCC has been studied by keeping the visual input constant and studying dynamic changes in awareness or visibility of a stimulus. The study of NCC tries to identify neural activity accompanied by awareness and neural activity that is not accompanied by awareness. Within 120 ms of the stimulus presentation, most areas are activated in essentially a feed forward manner and later activations are dependent on recurrent activations. It has been argued that such recurrent activity is necessary for consciousness (Lamme, 2003). On the lines of the distinction proposed by Block (2005), primary awareness is argued to be due to recurrent activations in early perceptual areas and access consciousness is argued to be due to activations in higher centers like prefrontal cortex. Given this brief overview of attention and consciousness, we move on to a discussion on linkages between these two important concepts.

Relationship between attention and consciousness

The relationship between attention and consciousness has been debated for a long time. One of the most quoted definitions of attention, ''Every one knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrain state'' (James, 1890/1950) makes a strong connection between the two concepts. Views that emphasize the close relationship between attention and consciousness are common (Posner, 1994). According to this view, what is attended and what is conscious are one and the same or at least closely coupled.

One of the most compelling set of studies that provided evidence for this close link came from the studies on inattentional blindness (Mack and Rock, 1998). They proposed the concept of ''Inattentional blindness'' to explain failure in detection of unattended stimuli. Their hypothesis is that we do not consciously perceive objects which we have not attended. In their experiment, observers were briefly presented a cross, whose vertical and horizontal component differed slightly in length. Observers were asked to judge whether the vertical or horizontal component of it is longer. In one of the trials an irrelevant stimulus was flashed in one of the quadrants formed by the cross. After the trial, observers were asked if they saw anything out of the ordinary. With their attention focused on the discrimination task, a large number of observers failed to notice the target stimulus. Around 25% of participants said that they did not notice the unexpected stimulus that appeared parafoveally and the cross was presented at fixation. Interestingly around 75% of the participants reported not perceiving the target stimulus that appeared at fixation with the cross presented parafoveally. Observers failed to report the irrelevant stimulus when they were not aware that such a stimulus might appear, although the normally irrelevant stimulus was easily visible. Mack and Rock (1998) argued that in the absence of attention, the irrelevant stimuli never rose to the level of conscious perception.

The inattentional blindness argument is used to explain the failure of change detection in several change blindness (CB) experiments (Grimes, 1996). Grimes (1996), in his experiment, tracked observers' eye movements while they viewed scenes for 10 s, in a change detection experiment. Scenes were altered during eye movements, and a single object was changed either in size, color, or location or they could disappear. Observers failed to detect these changes and inattentional blindness argues that the changes are not seen because the changed object was not attended and thus not consciously perceived. However, inattentional blindness fails to explain convincingly results of Simons and Levin (1997) or Rensink et al. (1997) experiments in which stimuli is presented for a very long time. In their CB experiments observers may have attended to the object and yet not detected changes to them. CB is the phenomena where we fail to perceive large changes, in our surroundings as well as in experimental conditions. Change could be in existence, properties, semantic identity, and spatial layout. Attention is required to perceive change,

and in the absence of localized transient motion signals (that may attract or grab attention), attention is directed by high level of interest (Rensink et al., 1997). Only when attention is focused on an object, can change in the object be perceived. The contents of visual short-term memory are simply over written with succeeding stimuli without focused attention (Rensink, 2002). Studies do show that more information is available than what is consciously available for report. For example, it has been shown that localization was above chance level even in undetected trials (Fernandez-Duque and Thornton, 2000). In addition, response times are longer in failed change detection trials in which change actually occurred (Williams and Simons, 2000). Change detection has been shown for changes in the background (Driver et al., 2001). More interesting are claims of blindsight in which observers claimed to sense the change before they were aware of the change suggesting that sensing occurs due to conscious non-visual awareness (Rensink, 2004).

A slightly different perspective on the close relationship between attention and consciousness is provided by studies in which load was manipulated and awareness of stimuli were evaluated (Cartwright-Finch and Lavie, 2006; Lavie, 2006). One was an inattentional blindness task in which the primary task was easier (low load) or difficult (high load). They found that inattentional blindness was more in the high-load condition compared to the low-load condition (Lavie, 2006). They also performed a change detection study in which the primary task (low load or high load) was presented at fixation and change between two scenes had to be detected at peripheral locations. Once again, change detection was better in the low-load condition compared to the high-load condition indicating that focused attention is necessary and plays a critical role in awareness (Lavie, 2006). In addition to better performance, a recent study has shown that attention can alter phenomenal appearance (Carrasco et al., 2004). They showed that the contrast of an attended (using an exogenous cue) grating was higher than the contrast of the unattended grating.

While acknowledging the close relationship between attention and consciousness, a large number of recent studies have convincingly argued that attention is different from consciousness (LaBerge, 1995; Baars, 1997; Hardcastle, 1997; Beilock et al., 2002; Naccache et al., 2002; Crick and Koch, 2003; Lamme, 2003; Woodman and Luck, 2003; Kentridge et al., 2004). According to Lamme (2003), consciousness operates prior to attention. Attentional selection operates on conscious stimuli leading to verbal report or store for later conscious, typically verbal access. Unconscious stimuli are outside the control of attention.

According to Dehaene et al. (2006), consciousness and top-down attention can be thought of in terms of a 2×2 matrix in which one of the dimensions is bottom-up stimulus strength (weak or sufficiently strong) and the other is top-down attention (absent or present). They ended with four classes of processing: subliminal-unattended, subliminal-attended, preconscious, and conscious. These different types of processes are sub served by different neural networks. Conscious processing refers to the case in which stimulus strength is high and top-down attention is present. This class is characterized by reportability, intense activation, and long-range interaction across cortical areas. The subliminal (unattended) is characterized by absence of priming and is not affected by top-down attention. These processes are also characterized as essentially feed-forward processes in the brain. Unlike the subliminal (unattended) processes, the processes in the other subliminal class are supposed to show stronger activation and short-term priming. Both the subliminal classes of processes are not associated with reportability. The preconscious, mainly sensorimotor in nature, display priming effects and are also not reportable in the absence of top-down attention. They also argue that global synchronization is characteristic of conscious processes and local synchronization is characteristic of preconscious processes.

Koch and Tsuchiya (2007) have proposed a fourfold classification scheme in which attention and consciousness are different. Certain processes are analyzed in terms of whether is attention necessary or not and whether they may give rise to consciousness resulting in a 2×2 matrix of possibilities. Some processes like early rapid vision does not need attention and may not give rise to

consciousness. This will also cover a significant amount unconscious information processing. Some processes may need attention and will give rise to consciousness. Some processes like priming and thoughts may require attention and may not give rise to consciousness. It is quite possible that some processes benefit from attentional processing without the involvement of consciousness. The most interesting possibility is the case of processes for which attention is not required but gives rise to consciousness. In the context of the differing views on the relationship on attention and consciousness, questions can be raised on whether attention is necessary for consciousness and whether consciousness is necessary for attention.

Is consciousness necessary for attention?

If consciousness is necessary for attention, then it implies a view of attention as a process that operates on stimuli that are consciously perceived. This view of attention belongs to the late selection theories of attention (Deutsch and Deutsch, 1963; Neuman, 1987; Lamme, 2003) which typically argue that early perceptual processing until conscious identification does not have capacity limitations and the purpose of attention is selection of appropriate actions. Using findings from a change detection study (Landman et al., 2003), it has been argued that change detection performance is very high. In the study, an initial stimulus display consisting of eight rectangular bars (vertically or horizontally oriented) was followed by a blank screen and a subsequent stimulus (with or without change). They showed that change detection performance with a spatial cue during the blank interval was as good as the performance with a spatial cue presented with the initial stimulus display indicating that attention operates on items that are in working memory and are conscious for a short amount of time. However, using a similar paradigm manipulating both the number of items as well as the duration of blank interval, Naaz and Srinivasan (2005) have shown that the better performance shown in the Landman et al. (2003) study saturate and is also dependent on the duration of blank interval with better performance with a longer interval of 1500 ms. These results question the interpretation of results from change detection for arguing that attention operates on items that were conscious.

Results from other studies also question the hypothesis that consciousness is necessary for attention (He et al., 1996; Naccache et al., 2002; Kentridge et al., 2004; Jiang et al., 2006). Top-down influences of attention on subliminal processing question the position that attention operates on information available through consciousness. Subliminal priming is present under attention but not when attention is diverted (Naccache et al., 2002). In a study with the blindsight patient GY, Kentridge et al. (2004) showed that a conscious cue helps in the detection of unconscious stimuli in his blind visual field. A peripheral grating that is masked by other stimuli and is not consciously perceived never the less produce orientation specific after effects (He et al., 1996). Unconscious emotional information has also been shown to attract attention (Jiang et al., 2006). Pictures of nudes which were not seen however had an effect on ratings of attraction indicating that unconscious information can attract attention. These results suggest that attention can operate on essentially unconscious information.

Is attention necessary or sufficient for consciousness?

The strongest arguments for the necessity of attention for consciousness have been the findings from inattentional blindness (Mack and Rock, 1998) and CB studies (Rensink et al., 1997; Simons and Levin, 1997). As discussed earlier, stimuli that were not attended were not seen in studies on inattentional blindness (Mack and Rock, 1998). However, some stimuli were seen compared to others indicating that possibly salient stimuli can capture attention. While these findings strongly support the claim that attention is necessary for explicit reporting of stimuli or change in stimuli, attention might not be the only critical factor in conscious detection of stimuli. Simons and Levin (1997) in their experiment showed that change detection was poor even when the changed object

was attended. Forty observers were shown a movie clip in which a single actor was performing some actions. When the actor was switched across the camera positions, observers failed to notice the change. When observers were warned about the change, performance was better. It was concluded that object features do not integrate automatically to form different views of the scene.

Results of experiment by Mack and Rock (1998) were interpreted by Braun Julesz (1998) as an argument to support the view that attention is not sufficient to detect change. The failure of change detection in their experiment was due to lack of expectation rather than attention. Rensink et al. (1997) found in their study that changes to objects of central interest are easier to detect than changes to objects of marginal interest. This faster detection of change to objects of interest could be due to the fact that we are expecting change to significant objects of the stimulus. Simons and Levin (1998) found that people did not notice even when the person to whom they were talking was changed. This failure to detect change could be because of our expectation of a stable world. We do not expect people to suddenly change into someone else. These studies on inattentional blindness and CB suggest that blindness is more from the observer's inability to anticipate the stimulus than from lack of attention. Thus, it might be concluded that attending to an object is necessary but not sufficient for change detection.

Studies with different paradigms have been used to argue that attention may be neither necessary nor sufficient for consciousness. One paradigm that is commonly used is the rapid serial visual presentation (RSVP) paradigm. In this paradigm, targets are presented one at a time very briefly. Typically the presentation of the target as well as the blank interval has a duration of around 100 ms. Participants have to detect two targets (T1 and T2) and the rest of the stimuli are distractors. Target T1 appears first followed by target T2 and the temporal gap (lag) between T1 and T2 is varied. The basic finding is that accurate identification of target T2 is poor for lag 2, i.e. when there is one distractor presented between T1 and T2. This phenomenon is called attentional blink (Raymond et al., 1992; Chun and Potter, 1995). The performance improves with higher lag and reaches asymptote around lag 6 or 7. In attentional blink, participants attend to the stimuli but are still not able to consciously detect T2 indicating that attention may not be sufficient for consciousness. When participants were asked to rate T2 visibility on a continuous scale, conscious report of T2 was dependent on a threshold for visibility of T2, i.e. attentional blink results from an all or none process of conscious perception rather than a gradual one (Sergent et al., 2005).

It has been shown that the T2 performance improves in the AB task when a task-irrelevant mental activity is concurrently performed (Olivers and Nieuwenhuis, 2005). Some hypotheses have been investigated regarding the distracting tasks affecting performance on T2 (Olivers and Nieuwenhuis, 2005). The critical claim of the overinvestment hypothesis is that the processing interference in the second stage is a direct consequence of allocating too many attentional resources to the RSVP stream. Conversely, a reduction of the attentional focus (by a distracting task) limits the number of items that can access the second stage, which alleviates the amount of interference and reduces the probability of an AB. This may also be due to the positive affective states that are induced during the task performance that pulls away resources from the central RSVP task, resulting in reduced interference (Olivers and Nieuwenhuis, 2005). These findings and findings from other studies (Lou, 2001; Leopold et al., 2002; Li et al., 2002; Suzuki and Grabowcky, 2003; Dijksterhuis et al., 2006; Kanai and Verstraten, 2006) have been interpreted to indicate that attention and consciousness can oppose each other (Koch and Tsuchiya, 2007). It is important that these findings are critically analyzed to see whether such a conclusion is warranted.

In a study with afterimages (Suzuki and Grabowcky, 2003), two overlapping triangles with different colors were shown. Participants were asked to attend to one of the triangles and not attend to the other triangle. Afterimages were obtained with both the overlapping triangles (attended as well as unattended) but they differed in their onset as well as offset. The afterimage for the unattended triangle appeared first followed by the afterimage of the attended triangle. In addition,

the afterimage of the attended triangle is weaker and disappeared earlier than the unattended triangle. Attention does not produce early or stronger afterimages. In another experiment, participants either attended to the afterimage inducer or to a stream of letters. They found once again that afterimages appeared later when the afterimage inducer is attended compared to when the letter stream was attended. Suzuki and Grabowcky (2003) have argued that these findings imply that attention facilitates adaptation of polarity-independent processes rather than polarity-dependent processes in the visual system. It has been argued that polarity-independent processes affect the visibility of afterimages whereas polarity-dependent processes affect the formation of afterimages. Koch and Tsuchiya (2007) have argued that the results of Suzuki and Grabowcky (2003) imply that attention and awareness can oppose each other. Based on the interpretations of Suzuki and Grabowcky (2003), it is clear that it cannot be concluded that attention and awareness oppose each other. Weakened afterimages of attended stimuli might be necessary for clear awareness of subsequently attended stimuli. This can be likened to the phenomenon of inhibition of return in which previously attended locations are inhibited. In this case, the afterimage of previously attended stimulus is inhibited or weakened which should have a facilitatory effect on the awareness of the next to be attended stimulus.

Leopold et al. (2002) showed that a bistable figure like a Necker cube is stabilized (only one percept is continuously seen) if the figure is presented periodically with a blank screen between successive presentations of the figure (intermittent presentation). Attention has been shown to play a critical role in perceptual stabilization. Distracting attention during the blank interval interfered with the process of perceptual stabilization with stimuli in which the direction of motion was ambiguous (Kanai and Verstraten, 2006). They have argued that the implicit memory for the percept that is necessary for stabilization utilizes attentional resources. When subjects had to perform a dual task, stabilization decreased. Koch and Tsuchiya (2007) argue that this decrease in perceptual stabilization due to withdrawal of attention due to a dual task is evidence for opposite effects due to awareness and attention. Similar to the situation with afterimages, it is important to consider carefully the effects of attention and consciousness when multiple percepts are involved. Is stabilization evidence for enhancement or reduction of effects of awareness? Without understanding the role of ''stabilization'' which has been demonstrated with bistable stimuli in awareness or its interaction with attention, it is too early to conclude that these findings show that the effects of attention and awareness are opposite.

The role of attention was evaluated using a dual task which used a relatively difficult visual search task in which observers had to search for an odd element in an array of five randomly rotated Ls or Ts as well as a scene/object categorization task in conjunction with a primary task. It has been found that the gist of a scene is processed quickly. Apparently simple tasks like differentiating between rotated letter stimuli were more difficult to perform rather than categorization of objects present in natural scenes like animal vs. non-animals and vehicle vs. non-vehicles (Li et al., 2002). Given that less attention is available for the secondary task, it is expected that performance will decrease irrespective of the secondary task. The fact that supposedly complex categorizations like animals or vehicles can be performed quickly has been taken to indicate that categorization involving meaningful stimuli can occur with almost no attention. This has led to the argument that attention is not necessary for consciousness. It is to be noted that performance difference alone under conditions of less attention can be used to argue that it is not necessary for consciousness. In fact, studies have shown that performance in certain tasks can actually be better under conditions of less attention (Yeshurun and Carrasco, 1998).

Experiments on decision making have also been used to investigate the role of top-down attention (Dijksterhuis et al., 2006). In two experiments, participants chose (experiment 1) and rated (experiment 2) the best car among 4 cars in which the cars were characterized by either 4 or 12 negative or positive attributes. One had 75% positive attributes, two others had 50% positive attributes, and the remaining car had 25% positive

attributes. Each attribute was presented for 8 s. The participants were asked to make a decision either after 4 min of thinking about the cars or after 4 min of a distractor task in which they had to solve an anagram. When the number of attributes were small, performance was better when participants consciously as well as voluntarily (with top-down attention) thought about the cars compared to when they performed the distractor task. However, when the number of attributes increased (beyond the working memory capacity) to 12, the performance (choosing the right car) was better with the distractor task compared to when they thought about the cars. The study indicates that thinking without deliberate attention results in better decisions. These results have been used to argue that not only attention and awareness are different but also can oppose each other (Koch and Tsuchiya, 2007). The with-deliberate attention and without-deliberate attention conditions showing performance differences can be linked to the implicit vs. explicit dichotomy explored in the implicit learning literature (Dienes, this volume; Reber, 1980).

A particular problem with the results that show that attention may not be necessarily related to consciousness pertains to the nature of the attentional processes that may be involved in these studies. It is not clear whether the top-down attention involved in these studies is the same since these studies employ different stimuli, tasks, and actions. Top-down attention can refer to any set of processes in which the participant voluntarily chooses to do a particular behavior.

A particular way to characterize top-down attention would be in terms of two types of top-down attention, focused attention and diffuse attention. These need not be two different types but can also be two ends of a continuum in which the focus varies. Under certain conditions, diffuse attention enables processing of certain attributes than focused attention. Diffused attention will be better at larger spatial scales than smaller spatial scales. One of the factors that need to be taken into account is that participants are well aware that they need to report the gist or report the second letter in an experiment on attentional blink. When stimuli are expected, it is quite possible that top-down attention acts through instructions to form a set and performance in these tasks cannot be thought as independent of top-down attention. Hence, the findings that indicate better performance for target T2 in an attentional blink study with the dual task or positive affect conditions (Olivers and Nieuwenhuis, 2005) can be better explained by a differential attentional strategy (focused or diffused). In addition, it is also not clear why a positive affect picture should be associated with near lack of top-down attention (Olivers and Nieuwenhuis, 2005). A better explanation would be say that another dual task or positive affect pictures make attention more diffused. Positive affect has been found to modulate selective attention (Fenske and Eastwood, 2003; Dreisbach and Goschke, 2004) producing increased cognitive flexibility and a diffused mental state. Thus positive affective valence may contribute to ameliorating the attentional blink thus increasing detection performance. Under conditions when a large number of items need to be considered or are competing for attention, diffused attention may work better than focused attention. What distinguishes this claim from Koch and Tsuchiya (2007) is that while they claim that these imply lack of top-down attention, we claim that only focused attention is reduced. It is to be noted that participants are still attending to both the tasks (dual task as well as target detection in attentional blink) and attentional blink is not completely removed. Only a 10% increase in performance is obtained.

In the decision making study (Dijksterhuis et al., 2006), both consciousness and top-down attention was present during deliberation. It is not clear whether thoughts about the car were intermittently present during the distractor task. Similar to the dual task/AB study, rather than interpreting that there was no top-down attention, a better interpretation would be that more diffused attention condition resulted in better performance when the number of attributes exceeded the capacity of working memory. Similarly when the number of attributes was within working memory capacity, focused top-down attention resulted in better performance. In fact, this interpretation actually suggests that using diffused attention can be a

better strategy when capacity limitations are exceeded. Diffused attention may underlie creative solutions for difficult problems where deliberate thinking using focused top-down attention may not provide a solution (Dietrich, 2004).

In addition to different types of top-down attention, the relationship between attention and consciousness can be examined by focusing on different types of awareness. For example, two types of awareness are proposed; perceptual awareness of stimuli without attention and access awareness of stimuli (needed for reporting) with attention (Block, 2005). Another proposal for two types of awareness in which attention play a different role is by Iwasaki (1993) who proposed two different types: object consciousness and background consciousness. Object consciousness is associated with spatial attention. Background consciousness is associated with global scene analysis and possibly can be linked to ''fringe consciousness'' (Mangan, 1993; Norman, 2002). So, it is also possible to associate two types of consciousness, one with focused attention and another with diffused attention or no top-down attention. In this view consciousness is still closely tied to attention and attention may be necessary for consciousness. The evidence for arguing that attention is not necessary for consciousness and may even oppose is not strong. Future studies may show consciousness with the complete lack of attention but even then the relationship between attention and consciousness needs to be framed considering the different types of attention and consciousness.

It is also important to understand the role of instructions in studies involving top-down attention and consciousness. For example, in a divided attention task, the participant knows which stimuli are to be expected (through instructions), can employ top-down attentional processes (through the formation of an attentional set). More meaningful stimuli will have a higher priority and hence top-down processes through the use of voluntary attention may benefit the processing of such stimuli. One pertinent issue that has rarely been discussed is the role of attention and awareness in actions. So far, the research on elucidating the relationship between attention and consciousness has focused on primarily visual stimuli and reportability as the evidence for awareness. Verbal report is a special case of actions performed by humans and findings on consciousness from verbal report may not generalize to those from other types of actions. This is supported by findings from Marcel (1993) in which different types of responding like manual, verbal, eye-blinks are used in a visual task. Performance differed with the type of response measure used indicating that more care is needed to disentangle the role of attention and consciousness in performing a given task. Performance differences with different types of actions may also be due to conscious factors as well as unconscious (automatic) factors underlying visual performance and actions. Given these considerations, what does research on automatic or voluntary actions have to say about the link between attention and consciousness? Attention is deemed to play a large role in response selection and action planning. Does the possibility of conscious actions exist without the involvement of attention?

Concluding remarks

Is attention and consciousness the same? The preceding discussion shows that the question may not have a simple answer. Given the fact that the term ''attention'' may refer to multiple processes, the interesting question would be explore the relationship between different attentional processes and consciousness-related processes. It is quite plausible that some attentional processes may be tightly linked to consciousness and may be necessary for consciousness and others are not at all related to consciousness. Further research in cognitive science will throw light on the relationship between these two critical concepts that are important in understanding the mind.

References

Baars, B.J. (1997) In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press, Oxford, England.

Baars, B.J. (2005) Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Prog. Brain Res., 150: 45–53.

Beilock, S.L., Carr, T.H., MacMahon, C. and Starkes, J.L. (2002) When paying attention becomes counterproductive: impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. J. Exp. Psychol. Appl., 8: 6–16.

Block, N. (2005) Two neural correlates of consciousness. Trends Cogn. Sci., 9: 46–52.

Braun, J. and Julesz, B. (1998) Withdrawing attention at little or no cost: detection and discrimination tasks. Percept. Psychophys., 60: 1–23.

Broadbent, D.E. (1958) Perception and Communication. Pergamon Press, London.

Carrasco, M., Ling, S. and Read, S. (2004) Attention alters appearance. Nat. Neurosci., 7: 308–313.

Cartwright-Finch, U. and Lavie, N. (2006) The role of perceptual load in inattentional blindness. Cognition, 102: 321–340.

Chun, M.M. and Potter, M.C. (1995) A two-stage model for multiple detection in RSVP. J. Exp. Psychol. Hum. Percept. Perform., 21: 109–127.

Crick, F. and Koch, C. (2003) A framework for consciousness. Nat. Neurosci., 6: 119–126.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J. and Sergent, C. (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci., 10: 204–211.

Dennett, D. (1991) Consciousness explained. Little Brown, Boston, MA.

Deutsch, J.A. and Deutsch, D. (1963) Attention: some theoretical considerations. Psychol. Rev., 70: 80–90.

Dietrich, A. (2004) Cognitive neuroscience of creativity. Psychon. Bull. Rev., 11: 1011–1026.

Dijksterhuis, A., Bos, M.W., Nordgren, L.F. and van Baaren, R.B. (2006) On making the right choice: the deliberation without-attention effect. Science, 311: 1005–1007.

Dreisbach, G. and Goschke, T. (2004) How positive affect modulates cognitive control: reduced perseveration at the cost of increased distractibility. J. Exp. Psychol. Learn. Mem. Cogn., 30: 343–353.

Driver, J., Davis, G., Russell, C., Turatto, M. and Freeman, E. (2001) Segmentation, attention and phenomenal visual objects. Cognition, 80: 61–95.

Eriksen, C.W. and Yeh, Y. (1985) Allocation of attention in visual field. J. Exp. Psychol. Hum. Percept. Perform., 11: 583–597.

Fenske, M.J. and Eastwood, J.D. (2003) Modulation of focused attention by faces expressing emotion: evidence from flanker tasks. Emotion, 3: 327–343.

Fernandez-Duque, D. and Thornton, I.M. (2000) Change detection without awareness: do explicit reports underestimate the representation of change in visual system? Vis. Cogn., 7: 323–344.

Grimes, J. (1996) On the failure to detect changes in scenes across saccades. In: Akins K. (Ed.), Vancouver Studies in Cognitive Science, Vol. 5. Perception. Oxford University Press, New York, pp. 89–110.

Hardcastle, V.G. (1997) Attention versus consciousness: a distinction with a difference. Cogn. Stud. Bull. Jpn. Cogn. Sci. Soc., 4: 56–66.

He, S., Cavanagh, P. and Intriligator, J. (1996) Attentional resolution and the locus of visual awareness. Nature, 383: 334–337.

Iwasaki, S. (1993) Spatial attention and two modes of visual consciousness. Cognition, 49: 211–233.

James, W. (1890/1950) Principles of Psychology, Vol. 1. Macmillan & Co., London.

Jiang, Y., Costello, P., Fang, F., Huang, M. and He, S. (2006) A gender- and sexual-orientation dependent spatial attentional effect of invisible images. Proc. Natl. Acad. Sci., 103: 17048–17052.

Kanai, R. and Verstraten, F.A. (2006) Attentional modulation of perceptual stabilization. Proc. R. Soc. Lond. B Biol. Sci., 273: 1217–1222.

Kentridge, R.W., Heywood, C.A. and Weiskrantz, L. (2004) Spatial attention speeds discrimination without awareness in blindsight. Neuropsychologia, 42: 831–835.

Koch, C. (2004) The Quest for Consciousness: A Neurobiological Approach. Roberts and Company Publishers, Englewood, Colorado.

Koch, C. and Tsuchiya, N. (2007) Attention and consciousness: two distinct brain processes. Trends Cogn. Sci., 11: 16–22.

LaBerge, D. (1995) Attentional Processing. Harvard University Press, Cambridge, MA.

Lamme, V.A.F. (2003) Why visual awareness and attention are different? Trends Cogn. Sci., 7: 12–18.

Landman, R., Spekreijse, H. and Lamme, V.A.F. (2003) Large capacity storage of integrated objects before change blindness. Vis. Res., 43: 149–164.

Lavie, N. (2006) The role of perceptual load in visual awareness. Brain Res., 1080: 91–100.

Leopold, D.A., Wilke, M., Maier, A. and Logothetis, N.K. (2002) Stable perception of visually ambiguous patterns. Nat. Neurosci., 5: 605–609.

Li, F.F., van Rullen, R., Koch, C. and Perona, P. (2002) Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci., 99: 9596–9601.

Lou, L. (2001) Effects of voluntary attention on structured afterimages. Perception, 30: 1439–1448.

Mack, A. and Rock, I. (1998) Inattentional Blindness. MIT Press, Cambridge, MA.

Mangan, B. (1993) Taking phenomenology seriously: the "fringe" and its implications for cognitive research. Conscious. Cogn., 2: 89–108.

Marcel, A.J. (1993) Slippage in the unity of consciousness. In: Bock G. and Marsh J. (Eds.), Experimental and Theoretical Studies of Consciousness. Wiley, Chichester, England, pp. 168–179.

Muller, H.J. and Rabbitt, P.M.A. (1989) Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. J. Exp. Psychol. Hum. Percept. Perform., 15: 315–330.

Naaz, F. and Srinivasan, N. (2005). Storage capacity and change detection of color in multi-element displays. Symposium on Emerging Trends in Neurosciences and XXIII Annual

Meeting of Indian Academy of Neurosciences, NIMHANS, Bangalore, India.
Naccache, L., Blandin, E. and Dehaene, S. (2002) Unconscious masked priming depends on temporal attention. Psychol. Sci., 13: 416–424.
Neuman, O. (1987) Beyond capacity: a functional view of attention. In: Heuer H. and Sanders A.F. (Eds.), Perspectives on Perception and Action. Lawrence Erlbaum Associates Inc., Hillsdale, NJ, pp. 361–394.
Norman, E. (2002) Subcategories of "fringe consciousness" and their related nonconscious contexts. Psyche, 8(15), http://psyche.cs.monash.edu.au/v8/psyche-8-15-norman.html
Olivers, C.N. and Nieuwenhuis, S. (2005) The beneficial effect of concurrent task-irrelevant mental activity on temporal attention. Psychol. Sci., 16: 265–269.
Posner, M.I. (1980) Orienting of attention. Q. J. Exp. Psychol., 32: 3–25.
Posner, M.I. (1994) Attention: the mechanisms of consciousness. Proc. Natl. Acad. Sci., 91: 7398–7403.
Posner, M.I. and Cohen, Y.A. (1984) Components of visual orienting. In: Bouma H. and Bouwhuis D.G. (Eds.), Attention and Performance X. Erlbaum, Hillsdale, NJ, pp. 531–556.
Raymond, J.E., Shapiro, K.L. and Arnell, K.M. (1992) Temporary suppression of visual processing in an RSVP task: an attentional blink? J. Exp. Psychol. Hum. Percept. Perform., 18: 849–860.
Rensink, R.A. (2002) Change detection. Ann. Rev. Psychol., 53: 245–277.
Rensink, R.A. (2004) Visual sensing without seeing. Psychol. Sci., 15: 27–32.
Rensink, R.A., O'Regan, J.K. and Clark, J.J. (1997) To see or not to see: the need for attention to perceive changes in scenes. Psychol. Sci., 8: 368–373.
Sergent, C., Baillard, S. and Dehaene, S. (2005) Timing of the brain events underlying access to consciousness during the attentional blink. Nat. Neurosci., 8: 1391–1400.
Simons, D.J. and Levin, D.T. (1997) Change blindness. Trends Cogn. Sci., 1: 261–267.
Simons, D.J. and Levin, D.T. (1998) Failure to detect changes to people in a real-world interaction. Psychon. Bull. Rev., 5: 644–649.
Suzuki, S. and Grabowcky, M. (2003) Attention during adaptation weakens negative afterimages. J. Exp. Psychol. Hum. Percept. Perform., 29: 793–807.
Treisman, A. (1960) Contextual cues in selective listening. Q. J. Exp. Psychol., 12: 242–248.
Treisman, A. and Gelade, G. (1980) A feature-integration theory of attention. Cogn. Psychol., 12: 97–136.
Williams, P. and Simons, D.J. (2000) Detecting changes in novel 3D objects: effects of change magnitude, spatiotemporal continuity, and stimulus familiarity. Vis. Cogn., 7: 297–322.
Woodman, G.F. and Luck, S.J. (2003) Dissociations among attention, perception, and awareness during object-substitution masking. Psychol. Sci., 14: 605–611.
Yeshurun, Y. and Carrasco, M. (1998) Attention improves or impairs visual performance by enhancing spatial resolution. Nature, 396: 72–75.
Zeki, S. (2003) The disunity of consciousness. Trends Cogn. Sci., 7: 214–218.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 7

Computational studies of consciousness

Igor Aleksander[1,*] and Helen Morton[2]

[1]*Department of Electrical and Electronic Engineering Imperial College, Room 615, South Kensington Campus, London SW7 2BT, UK*
[2]*School of Human Science and Law, Brunel University, Uxbridge UB8 3PH, UK*

Abstract: In this chapter we present a computational architecture intended to add clarity to the concept of consciousness. We briefly review some of the motivations of work done in this area in various institutes around the world and looks closely at our own work which specifically includes phenomenology, the sense of a self in a perceptual world. This breaks consciousness into five axioms: presence, imagination, attention, volition and emotions. It develops plausible mechanisms of each and how they interact to give a single sensation. An abstract architecture, the kernel architecture, is introduced as a starting point for building computational models. It is shown that through this architecture it is possible to discuss puzzling aspects of consciousness, for example are animals conscious? What happens when we dream? What goes on when we experience an illusion? This paper is intended to elucidate and update some concepts introduced in Aleksander (2005).

Keywords: brain modelling; consciousness; neural architectures

Introduction and overview

This paper reports on recent efforts among computer scientists to create models of systems said to be conscious in some way. The professional philosopher might balk at this as the link between computational material and consciousness is seen as a difficult and puzzling topic which needs to be understood in a philosophical way before it can be studied computationally. The philosophy followed in our approach is that the process of computational modelling becomes part of the philosophical process of understanding. Philosophers particularly discourage notions of making machines that are claimed to *be* conscious *like* ourselves. This is not on the agenda of computational studies of consciousness. The process is directed towards the creation of virtual computational entities that become conscious of virtual worlds and to clarify what it is for this to happen.

To explain, think of a computational study of hurricanes rather than consciousness. This proceeds by first extracting the physical features that support a real hurricane in a real world and creating, in the computer a virtual hurricane in a virtual world. As an aside, the word 'virtual' refers to a mechanism that is simulated on a computer and may be studied for its own properties with little or no influence from the host computer. For example, most computers are equipped with a 'calculator' button, which when pressed, makes a calculator appear on the screen which can then be used as one might use a pocket calculator. So, the purpose of a virtual hurricane in a virtual world is

*Corresponding author. Tel.: +44(0)20 7594 6176;
Fax: +44(0)20 7594 6274; E-mail: i.aleksander@imperial.ac.uk

DOI: 10.1016/S0079-6123(07)68007-8

to tell us how a real hurricane will behave in a real world. Of course, the virtual hurricane will not destroy our office but it might tell us whether the one brewing up in the next town will or will not come our way.

It would be useful if we could simply replace the word 'hurricane' by the word 'consciousness' in the above procedure. That is, first extract the physical features that support real consciousness in a real brain and create, in the computer, a virtual consciousness that is conscious of a virtual world. But 'extracting the physical characteristics of consciousness' is contentious and touches on many dissonant beliefs about the relationship between the physical and the experiential. Here we move forward by hypothesising a relationship between the physical and the experiential and checking the hypothesis in a *virtual* world. We are not aiming to design a machine that will be 'conscious of a real world in the way we are'. Of course transferring the computation to a robot might allow the replacement of the virtual world by a real one. But the consciousness will not be 'like our own'. This is a meaningless phrase and it is sufficient, in order to address the concept, for robot consciousness to stand in relation to the existence of the robot, its needs and tasks in the way that *our* consciousness stands in relation to our needs and tasks. There is likely to be a major difference between our needs and tasks and robot needs and tasks, without damaging the discussion on the role and mechanisms of consciousness.

This paper discusses the very first step of extracting important mechanisms necessary for the building of a virtual machine. Recognising that consciousness is many things, it breaks consciousness down into five distinct 'axioms' to facilitate the design process, as will be seen. We first take a look at the current nature of the paradigm of 'Machine Consciousness' before returning to our own axiomatic approach giving an illustration through a virtual system, and discussion of some important puzzles related to the mechanisms of consciousness: why does vision play tricks sometimes? How does one check for the presence of consciousness? Are animals conscious? Is there higher order consciousness? What happens in unconscious moments?

Some history and those involved

At the International Conference on Artificial Neural Networks at Brighton in 1992, the two conference chairs, John Taylor and one of the authors (IA) independently suggested that the future challenge for neural network researchers was the discovery of the neural correlates of consciousness. But it was 9 years later, in May 2001, at a small workshop sponsored by the Swartz foundation at the Cold Spring Harbour Laboratories, that the paradigm of machine consciousness appeared to have been established among international workers. Organised by Koch, Chalmers, Goodman and Holland, a mixed group of neurologists, computer scientists and philosophers were reported to have concluded:

> The only (near) universal consensus at the workshop was that, in principle, one day computers or robots could be conscious. In other words, that we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans. (http://www.swartzneuro.org/abstracts/2001_summary.asp)

This led to the publication in 2003 of a special issue of the *Journal of Consciousness Studies* edited by Holland (2003). Here, several of the current researchers state their positions. Specifically, Franklin (2003) of the University of Memphis describes the Intelligent Distribution Agent based on Baars' Global Workspace model. This is important work as it concentrates on creating a system (for billeting sailors) the users of which believe that they are dealing with an entity (accessible by e-mail) that is aware of their needs. This is typical of what may now be seen as a 'functional' approach to models which create systems that appear to be conscious through their behaviour. These may be contrasted with 'phenomenological' approaches that concern mechanisms that may be needed to generate *internal sensations* as indicated in a paper by Aleksander

and Dunmall (2003). This approach is elaborated and extended in this paper.

Since 2003, several conferences, mainly the yearly meetings of the Association for the Scientific Study of Consciousness (ASSC) and Artificial Intelligence and Simulation of Behaviour (AISB) have created streams for the modelling of consciousness. The major issues debated at these conferences are the design of conscious machines (Haikonen, 2005; Holland, 2006), machines with imagination (Shanahan, 2005) and the generation of a synthetic phenomenology (Gamez, 2006; Aleksander and Morton, 2006). A second special edition of the *Journal of Consciousness Studies* devoted to machine consciousness is being compiled (Clowes et al., 2007).

The aims of computational models of consciousness

Returning to the extraction of the essential physical features of consciousness, in the work reported here the agenda is one of hypothesising such physical features using an axiomatic decomposition. This then leads to the desired virtual organism conscious of a virtual world and an assessment of what this tells us about a real organism that is conscious of a real world.

Why bother with machine models of consciousness? We are motivated by four reasons:

1. To understand in a *constructive* way (i.e., understanding by building) what it might be for any organism to be conscious.
2. To be able to discuss formally some puzzling features of consciousness (unstable vision, illusions, tests for being conscious, unconsciousness …).
3. To encourage those who work with conscious organisms to face the complexity of the brain in a formal way.
4. To ask if implemented systems have new uses.

There is considerable debate in the contemporary philosophical world as to whether consciousness, originally a philosophical concept, is to be addressed by science at all. Or is it, perhaps, an abstraction like beauty, or altruism that is best served through philosophical discussion rather than being analysed through science. In this paper we side with those who see consciousness as a definite product of the brain and seek to reconcile what is known of the neurology of the brain with the five axiomatic features that are explained in the next section. In particular we address and argue against the notion of an unassailable 'hard' problem, which, in order to prove the independence of consciousness from the physical, claims that a brain may become totally bereft of its consciousness (not just in sleep) without any physically measurable change in its neurology.

The technological basis of our work is that of neural architectures studied with a computationally efficient model of the neuron. This is elaborated in appendix. As mentioned, our narrative for computational studies of consciousness starts with a truly introspective decomposition of different aspects of consciousness into five major axiomatic sensations (presence, imagination, attention, planning and emotions) so as to make the task for the designer of computational models easier. These axioms are mapped into mechanisms that form an interlocking ensemble. Such an ensemble is called a Kernel Architecture (KA) as it forms the basis of many computational models designed to date. Also in developing the mechanisms that support the axioms, we introduce a central mechanism: *depiction.* This involves not only sensory pathways in the brain, but also signals from the musculature of the body.

Then, to illustrate our computational methodology and the role of the KA, we introduce work done on a digital model of visual awareness. This leads to discussions of some of the puzzling aspects of consciousness: volition, animal consciousness, the possible causes of dreams and the causes of Freud's concept of the unconscious. In the conclusion we assess how far our four motivational points mentioned at the beginning of this introduction have been moved forward.

The challenge of phenomenology: is there a hard problem?

The word 'phenomenology' is generally used for philosophical concerns that include, if not begin

with, personal, internal feelings of being conscious. Importantly from a modelling standpoint this implies that introspection will be the starting point for computational studies. This distinguishes phenomenology from other approaches, where one is concerned with what it is for an object to be conscious. One should also distinguish 'a phenomenon' from other constructs such as 'qualia' which relate to sensational primitives such as 'redness' or 'the sweet smell of a rose'. In general, phenomenologists like to extend the definition beyond the directly sensory to more compositional structures of experience such as enjoying a game of tennis or the experience of having dated a new person. The reader is referred to Aleksander and Morton (2006) for a more detailed account of the concept and its relevance to the machine modelling of consciousness.

The implication for computational studies is to establish a link between experienced conscious states, and the observable states of some underlying physical computational mechanism. First, the word 'computational' needs to be addressed. Conventionally this suggests a process with a defined algorithm that runs on a von Neumann type of architecture. Here we do not mean this at all. Computation is taken to be the state development of an architecture that is controlled by and in touch with a world. The prototype model for such an architecture is the living brain. This is an evolved system the fitness function of which is to allow the organism to internalise and best use in the maximum detail possible the reality of the world and the organism's own potential in this world. In computational terms this leads us to the following central assertion:

Assertion 1. To include phenomenology in a computational model of consciousness it is necessary to abstract the physical/informational properties of the brain that are hypothesised to determine conscious states and study these as structures that are virtual on a conventional computer architecture in order to check the validity of the hypotheses.

In other words, the algorithms that create the virtual object of study are totally unimportant. What matters is the virtual architecture which is brain-like in broad essence, but linked to the first person phenomenology, the axiomatic characteristics of which are made explicit below.

Here, brief mention must be made on how the above assertion relates to what Chalmers (1996) calls 'the hard problem'. He allows that while the states of mind we actually sense may (and do) have a correlation with physical events that may be measured in the brain, these physical events do not, in a logically guaranteed way, lead to the presence of phenomenological personal events. In other words, this opens the possibility of finding 'zombies': organisms that are entirely physically equivalent to conscious ones except for the lack of phenomenology. Therefore (according to Chalmers), while work on the physical may be very interesting, it does not lead to an understanding of the phenomenological. We take the view that going from a conscious object and arguing that one could remove its consciousness with no physical intervention is unhelpful and assumes ab initio that consciousness is free from physical process. To understand phenomenology it is necessary to *start* with an organism that undoubtedly *is* phenomenologically conscious. Then one can attempt to develop the mechanistic underpinnings of such an organism. To develop such a theory for a device that may or may not be conscious then becomes a quest that, in our opinion, does not lead to an understanding. But the only organism that we each know is undoubtedly conscious, is our very self, and it is for this reason that the axioms developed below are based on introspection.

Depiction: the key mechanism for conscious representation

Before addressing the axioms we need to establish an important point about a form of representation we call a depiction. Depiction is a direct representation of where elements *of* the world *are* in the world which is encoded by the efforts of the mechanism to attend to such elements. An example of depiction appears in Fig. 1 (reproduced from Aleksander, 2005).

There is a vast literature that indicates the existence of cells that react to sensory signals only if addressed by the appropriate muscular activity

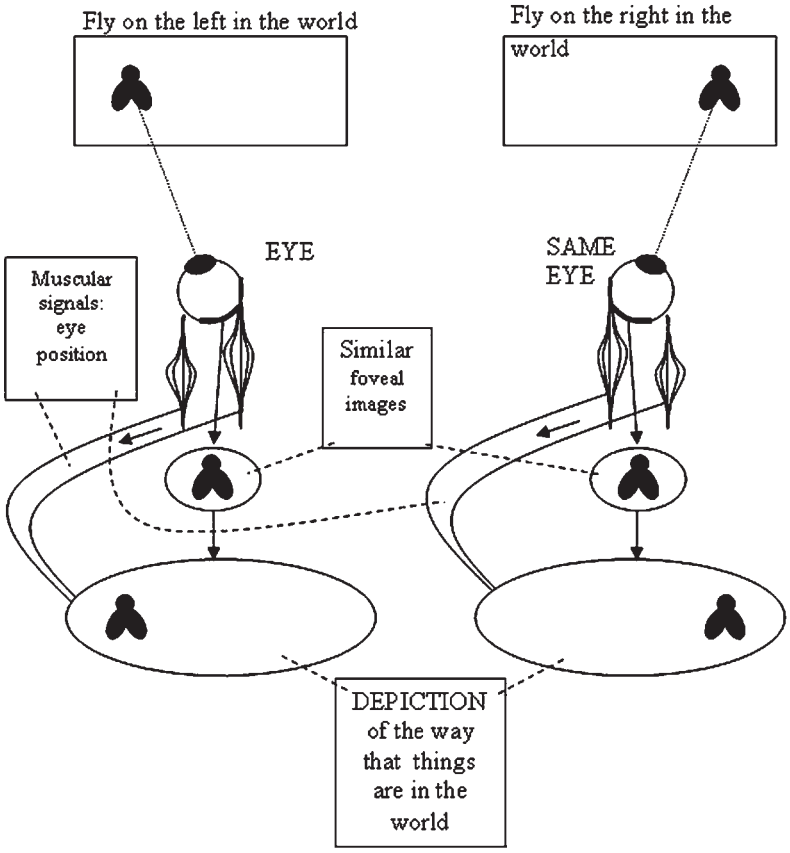


Fig. 1. An example of the way that visual and muscular signals are used to create a representation of where things are.

(see Aleksander, 2005, for references). This involves a vast variety of input from musculature including head, body and even arm movement. That is, if one points at an object, this helps to depict that object in the brain. So if there is an overall mechanism that is essential for a machine to support conscious sensation it is depiction. This leads to the second assertion.

Assertion 2. The parts of a mechanism that sustain conscious experience can only do so if they are the product of a depictive process.

But this needs elaboration as is done in the next section.

Current perspectives on five axioms

Following the introspective route, we identify how being conscious may be broken down into distinct experiences. Five of these were identified some years ago (Aleksander and Dunmall, 2003), and the list has not changed. Clearly being conscious of being at a concert at the Albert Hall in London is a different experience from waking up in the morning and trying to work out what to do during the day. This perspective leads to an intuitive way of *dividing conscious experience* into the following five axioms (stated in the first person to underscore their introspective nature):

1. *Presence: I feel that I am an entity in world that is outside of me.*

This is the most fundamental conscious experience one can have: the feeling of being an empowered, active agent in a real sensory world. Although the visual sensory modality is often used to illustrate this sensation, hearing and touch too are implicated. Lack of one or two of these can be compensated by the remaining modalities.

2. *Imagination: I can recall previous sensory experience as a more or less degraded version of that experience. Driven by language I can imagine experiences I never had.*

Sometimes this is called memory. But that is an understatement of the sensation which allows us to have a rich imagery of the world as once experienced through Axiom 1. It also allows us to create scenarios that might have been experienced or even ones that are not close to reality (say the world of Harry Potter).

3. *Attention: I am selectively conscious of the world outside of me and can select sensory events I wish to imagine.*

Attention is a vast and important topic. Sometimes it feels as if what we choose to experience is automatic. For example, a bright flash will attract our attention and then we might experience the firework patterns that go with it. A sound bang can do the same. At other times the target of our attention seems to be selected more purposefully. For example, if on seeing a motorcar, we wish to identify the make of the car, we purposefully shift our gaze to the point on the bonnet, the wheels or the rear where an emblem is expected to be found.

4. *Volition (previously called Planning): I can imagine the results of taking actions and select an action I wish to take.*

When several options for action are available, it seems possible to us to imagine taking the actions in succession and to predict the ensuing results even if they may not be very clear. For example, we may be choosing between going to a restaurant we know and thinking about what dishes are available or going to a previously untried restaurant.

5. *Emotion: I evaluate events and the expected results of actions according to criteria usually called emotions.*

Following on from the last example, a decision about which restaurant will be chosen leads to evaluations. The first might be a steak house, but the thought of eating steak may be accompanied by the pleasure of the experience, while also producing a negative feeling about the intake of cholesterol. The thought of the unknown restaurant may provoke both a fear of the unknown and the excitement of a new adventure. The involvement of emotional values is in clear evidence.

The mechanistic use of the axioms: the kernel architecture

So far we have merely identified what seem to be some important ways in which we distinguish between distinct elements of the first person experience of being conscious. This is merely the first but important step towards a more ambitious aim: to have an understanding of distinct mechanisms that may be necessary to support these distinct sensations. Once such distinct mechanisms have been evaluated, the next question asks how they can interlock to provide a sensation of a unified consciousness. To help us with the latter notion, we suggest a KA (Fig. 2) with which we can comment on the individual mechanisms. The KA is an abstract physical structure composed of neural areas. It is neural (at least, cellular) because that is the only way that the need for detailed depiction is satisfied. It is intended to facilitate discussion rather than something that will essentially be built as a model. The word 'Kernel' is due to the fact that it appears *as part* of many models that have been implemented in the past as is seen below.

1. *Presence*

The key necessary mechanism here is *depiction* which has been defined separately earlier. That is, the output of a neuron that is going to be representative of where an element of the sensory world is located (with respect to some 'at rest' body coordinates) must not only react to that element as a sensory signal, but must also be indexed by signals that position the element with respect to some frame of reference in the world. Such an indexed neuron becomes intimately connected with the event in the world while its position in the 'brain' network become less important.

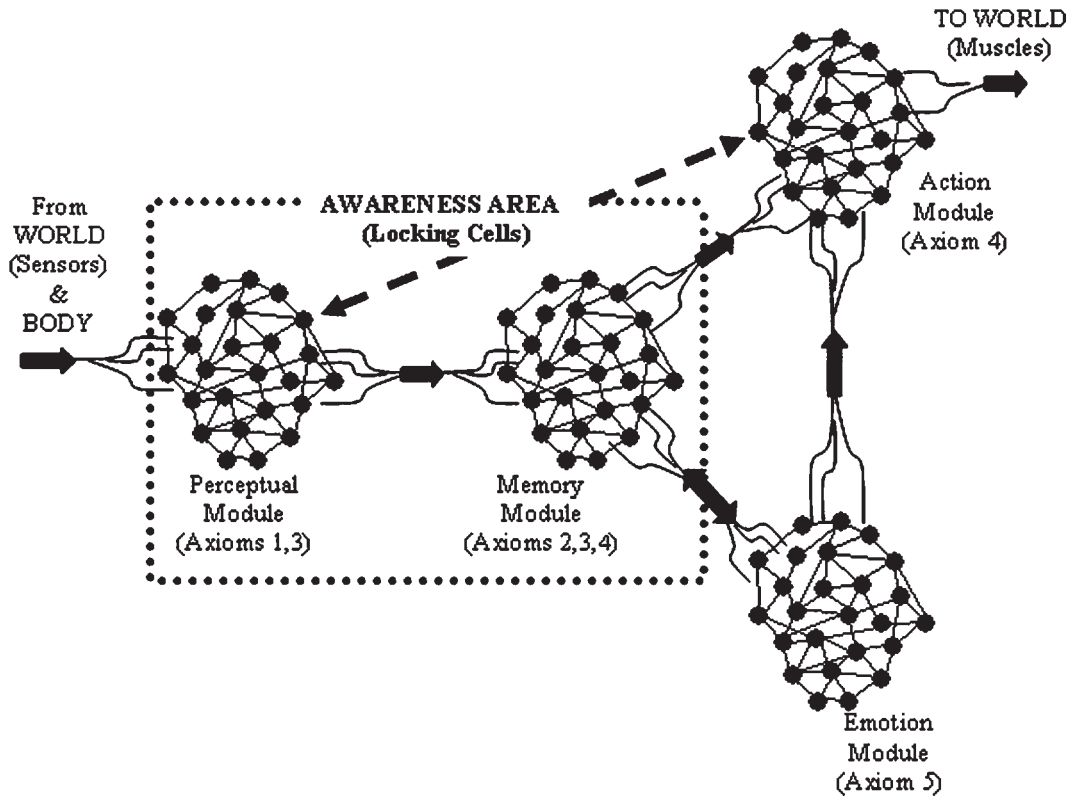


Fig. 2. The Kernel Architecture.

For example, it is known that an elemental event in the world such as a small moving green triangle stimulates representational neurons in different parts of the visual cortex: a colour area, a motion area and a shape area. How this *integrates* to provide a coherent sensation has been a persistent problem for neurophysiologists. This is the 'binding' problem. In accepting that neurons become depictive, that is, indexed by world events, the binding problem virtually disappears. The three separate neurons, being related to the same world event will depict colour, motion and shape but as they place it in the same place in depictive space, we submit that the sensation is one of overlap.

In the KA it is the perceptual module that implements the support mechanisms for Axiom 1. It is noted that depiction is obtained through the dotted connection to the action module which is the path for the necessary muscular signals.

2. *Imagination*

The simplest imaginational act is to look at an object in one side of the room (say the Left), then switch attention to another object somewhere else in the room (say the Right), after which, with eyes closed, it becomes possible to recall the object on the left (say A), then the one on the right, say B. The ability to recall suggests the presence of a dynamic neural system (neural automaton in automata theory) for which A and B are different attractors. In the KA this is labelled 'The Memory Module'. During perception, the depiction in the perceptual module (say, Ai) acts as input to the memory module. A version of this input Ai' also becomes a state of the memory module through a process known as Iconic learning. In simplified terms, iconic learning is the transfer of input pattern Ai to the state variables of the memory module to become Ai' through the following

learning sequence. The left of the equations below is the input to the memory module and the right is the resulting or taught state. The format therefore is (*state, input* $\rightarrow$*next state*).

$$(R, Ai(t)) \rightarrow Ai'(t+1) \qquad (1)$$

where R is a random initial state of the memory module

$$Ai'(t+1), Ai(t)) \rightarrow Ai'(t+2) \qquad (2)$$

We note that starting in some other time varying random state R' and random input, say Ri, a transition to Ai' is likely (due to the 'nearest neighbour' nature of generalisation of the neurons — Ai' being the only neighbour) at which point Eq. (2) will ensure that the memory module remains in that state. If the training procedure is repeated (randomising the state) for input Bi, then a similar attractor is created for Bi'. Starting from some arbitrary state, the system can fall into either attractor. It requires sensory input to switch from one attractor to the other. In implementations extensive use is made of noise bursts to encourage ranging over remembered events.

It should be noted that Ai and Ai' need not be exactly the same, but are expected to be similar much as if we look at something and close our eyes, the memory is less vivid than the perception. Vividness depends on the properties of the network. It should further be noted that sequences of Ai can lead to sequence attractor trajectories in the memory network. So, in general terms, the interaction of the memory and perception networks in the KA form a system of laying down experience. This experience has a state structure (i.e., state graphs) in the state space of the memory automaton.

3. *Attention*

The lowest level of muscular attention (eye movement etc.) is not shown in the KA. This is largely the unconscious process that causes eyes (or hands, say in the case of the blind) to move to salient parts of the world as mediated by sensory input. In the case of vision in the brain, the key mechanism is that of the superior colliculus which moves the eyes on the basis of its own representation of the retinal image. However, the above example of looking left and right to remember what is on the left and right is an instance of purposeful attention. How does the mechanism of 'thinking' about what is on the left and what is on the right work in the KA?

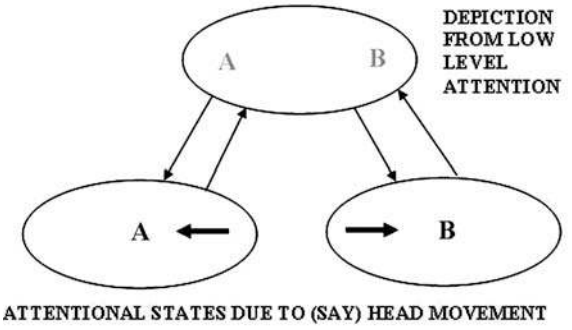


Fig. 3. Attentional states.

Simplifying things to the extreme, Fig. 3 shows the states of the perceptual automaton entered as a result of muscular action. The top state showing both A and B is a somewhat hazy depiction that is created by unconscious attentional eye movement, but one that requires focus to identify the objects it contains. The arrow leading to the A state is caused by purposeful movement, say, of the head (what causes it is not important here). Being depictive, the neurons that record this focus-on-A state will also be indexed by the left head movement. This has just been indicated as a thick arrow. Similar events take place on the B side. This exact state structure, including the motion sensations, may be transferred to the memory automaton, which due to the generalisation mentioned above, can explore it without input from the perception module. Were the system to be endowed with language, the memory automaton has all the necessary information to drive a statement that says: "if I attend to the left object, I remember an A". This, then, is an example of inner attention.

4. *Volition and 5. Emotion*

While these axioms involve two separate mechanisms, it is helpful to see them as they work together. All four modules of the KA are implicated and it may be best to look at Fig. 4.

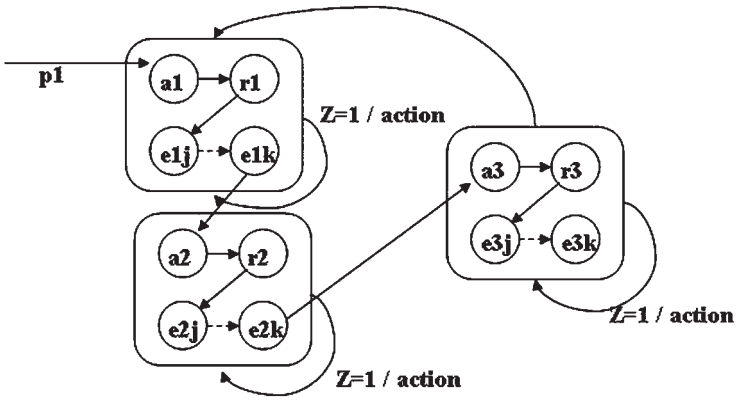


Fig. 4. A volitional state structure: cycling takes place until 'wantedness' exceeds a threshold.

This represents state activity as seen from the perspective of the memory module and we refer to a very simple scenario of being in a restaurant, and looking at the menu.

The state of the Axiom 1 module is labelled *p1* and it represents a depiction of the menu. This acts as an input to the memory module where it causes first a memory of the first item *a1* (say a pizza). This, in turn, recalls the experience of eating the pizza (*r1*) and the emotions that accompanied the experienced (iconically absorbed from previous experience in the emotion module (*e1j,e1k* …) where these may be akin to pleasure in the taste of the food, guilt (for eating fattening food) …. etc. Having dealt with the pizza, the memory module repeats the exercise for other items on the menu in some kind of succession. It is noted that no action need occur during this contemplative activity. What is changing however, is some cumulative 'wantedness' value in the emotion module related to each of the felt emotions. This value could be positive or negative, but it is assumed that the emotion module is capable of accumulating a wantedness value for each of the contemplated food items. If this value exceeds some threshold, it triggers a signal Z from 0 to 1 and this does two things: first it discontinues the search through the item and causes the wanted event to be held in the memory module. Second, it activates the action module that causes whatever action needs to be performed (calling the waiter, say) to obtain the wanted food.

Of course it is quite possible that the Z parameter is never triggered. This, in the strongest version, would be described as a pathological 'freezing' of the organism. Weaker versions might be described just as 'not being able to make up my mind'. However, it is quite possible, certainly in principle, for there to be some source of noise which is additive to the wantedness value and which causes Z to be triggered. Introspectively this corresponds to a feeling of 'Oh blast! I'll make an arbitrary choice'. From this can follow discussions about the nature of free will which is revisited below among topics related to the puzzling aspects of consciousness.

The kernel architecture and visual awareness

We have recently discussed the role of the KA in models of visual awareness (Aleksander and Morton, 2006), but here we wish to add some insights. For completeness we show where the KA fits into a model of visual awareness (Fig. 5).

The experiment is designed to show that the depictive hypothesis underlying the first four axioms forms a coherent dynamic system. Figure 5 is typical of the experimentation we carry out with 'Neural Representation Modeller' software in which architectures of neural networks can be created (see Aleksander et al., 2001). In this particular experiment an image was represented in a World, over which an eye extracted pictorial information. Every box in the figure is a neural

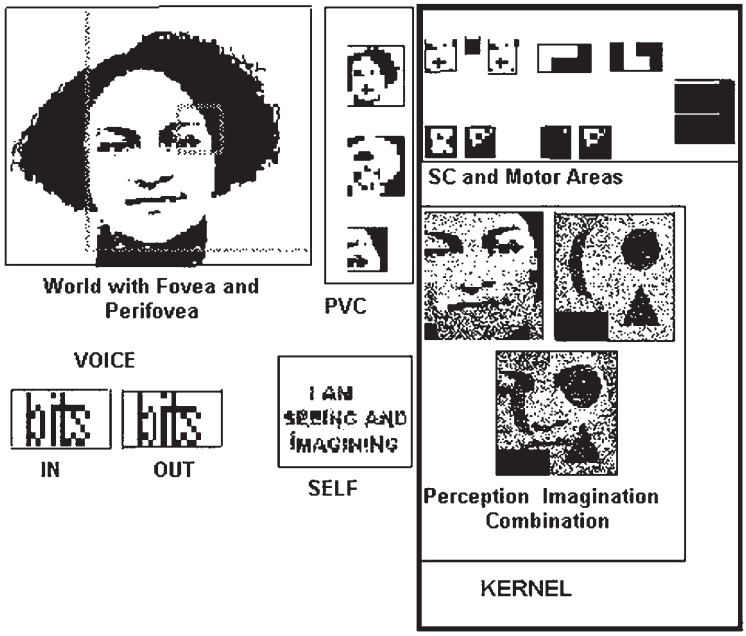


Fig. 5. A virtual neuromodel for visual awareness.

area and every dot in such a box is the state of a neuron. The KA is further supported by a primary visual cortex (PVC) which extracts the foveal information. There are also input and output interfaces.

The KA appears as several neural areas in the Action automaton. One of these is SC, a rough model of the brain's superior colliculus which has a low-level representation of the perifoveal field but drives the 'eyes' to major areas of saliencey (spatial or temporal frequency). Another (two horizontal lines) is the interpreted signal which drives the eye "muscles" (one line horizontally and the other vertically). There is also a 'SELF' box which is a neural system that measures the state of most of the activity in the system and maps it into a small set of linguistic statements, all of which start with the word I (I am seeing, I am imagining ….).

The perception and imagination modules are shown as well as the combination that comes into awareness as a result of the assumption that both are 'depictive'. In the experiment we investigate the connection balances that allow various conditions such as seeing and imagining, just seeing or just imagining to take place. But the point made here is that the presence of the KA gives us the licence to say that there is a *prima facie* case for saying that the virtual system model is potentially conscious of its visual world. At least it provides a challenge for a skeptic to argue why this may not be true.

Unstable vision

Sometimes it is instructive in the analysis of visual awareness mechanisms to look at situations where visual awareness plays tricks. One such instance is the well-known 'Necker cube' shown in Fig. 6.

This is known as a 'reversible' figure where sometimes feature A is in front of B and then the situation reverses. Carefully conducted tests show that most individuals, once the reversal is experienced can neither stop it nor control its frequency (on average about 70 reversals a minute). We also noted that while eye position seems to follow reversals it is not responsible for them. Fixating the gaze does not stop reversals as many researchers have found. This 170-year-old problem has had many suggested solutions. A vast literature shows

that they fall mainly in three camps: eye gaze control, habituation in visual pathways and (Gregory, 1997) ambiguities in 'addressing' memory in a rule-based system. None of these appear to conform comfortably with measured data. The KA has enabled us to suggest an additional mechanism as seen in Fig. 7.

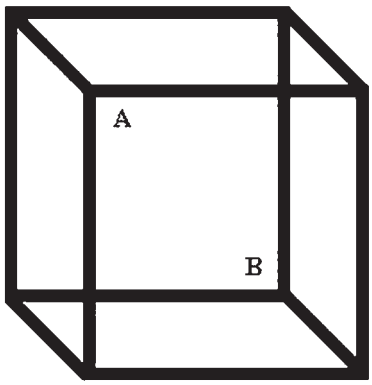


Fig. 6. The Necker cube.

It is known that in the visual system there are two major pathways between input and eventual muscular action: the ventral pathway and the dorsal one. The ventral path is known to carry signals from visual input that identify what is seen (the 'what' of vision). The dorsal pathway (according to Milner and Goodale, 2006) is not so much a 'where' of vision as thought previously, but a control input to action areas of the brain. Experiments done with participants with lesioned ventral pathways, show that such participants are able to generate appropriate actions in response to visual input (such as something approaching) without reporting any awareness of the visual event. This is called 'blindsight' and indicates that the dorsal pathway does not contribute to conscious sensation. Consequently, in the KA of Fig. 7

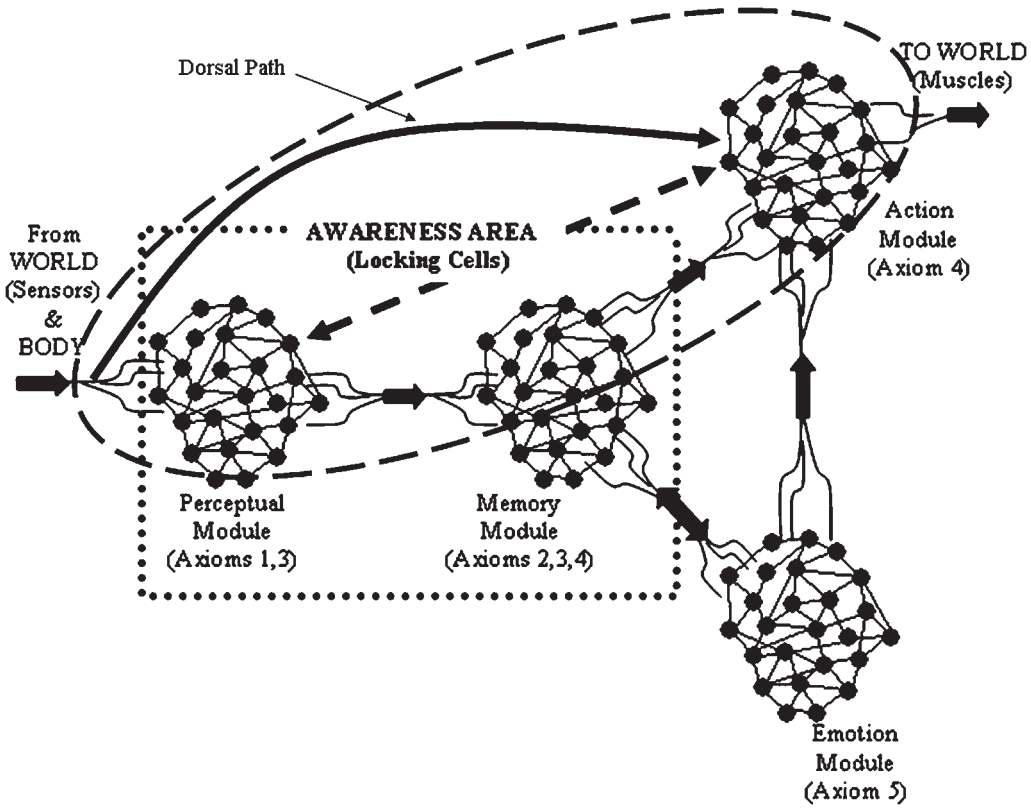


Fig. 7. The unconscious dorsal pathway responsible for ambiguous messages to action machinery.

the dorsal pathway bypasses the awareness area from retina to the action module. According to axiomatic theory, this means that events in this pathway are not depicted and hence are not conscious.

We are suggesting here that the degree of puzzlement of Necker cube illusion is due to the fact that the ambiguity lies in the ventral, unconscious path. That is, the visual cue unconsciously causes the action module to prepare action to touch the cube in two different ways. However, the action module also feeds positioning signals back to the depictive areas of the architecture. This suggests that the depictive mechanisms might be affected and as the instability occurs in the dorsal path, it modifies the perceived position of the cube edges. In Fig. 8 we show an implementation of the structure in Fig. 7. This includes a simulation of the result of the action module on the position of the fingers and how this could influence the perception of the cube.

Two metastable sets of windows of the system are shown indicating the two finger positions driven by the dorsal path. Note that these can be vetoed from becoming actual finger movements while the intention of positioning the fingers remains. The other window shows the related visual sensation where one of the faces of the cube is noisy due to the muscular depictive feedback. It indicates the face the thumb might touch even if no touching takes place. This hypothetical situation might also explain why the perception largely cannot be controlled, as the normal conscious attention control via the memory module and the action module is short-circuited by the dorsal pathway.

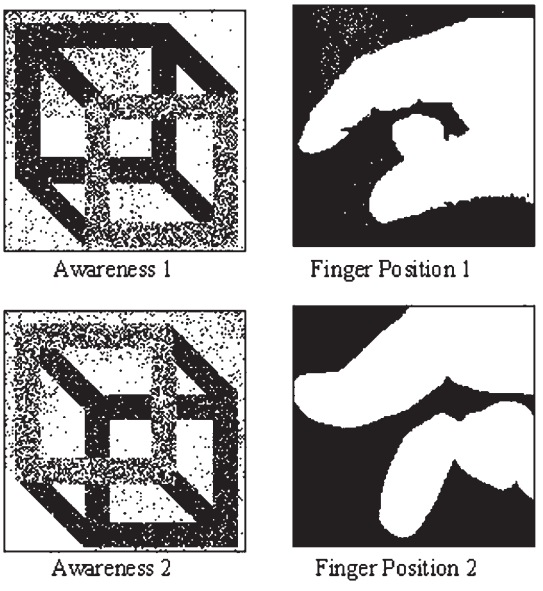


Fig. 8. Imagined grabbing the Necker cube in different ways and the effect of this on perception.

Is an organism conscious?

Here we begin to look at the KA and the axioms as tools with which to enter important discussions about consciousness. The first is as a set of tests of consciousness.

Assertion 3. There is a prima facie case for an organism to be conscious if it has a system that is isomorphic with the Kernel Architecture hence support the five axiomatic mechanisms.

Before applying this, it is most important to stress that this is a *necessary* condition for consciousness to be supported. It is not sufficient — the KA does not disappear if the organism becomes unconscious by going to sleep. It also provides no qualifying assessment of what the organism is conscious of. Clearly the virtual machine in Fig. 5 and a conscious human might in abstract principle share some mechanical properties that support both conscious acts, but such acts might be completely different. So we learn to separate the content of consciousness from the mechanisms that are necessary for there to be any content at all. Therefore, the test suggested by Assertion 3 looks for the presence of some form of consciousness without looking for classifying levels for such consciousness.

Animal consciousness

The above argument allows us to approach a basic question: are animals conscious? This question arises usually in the context of animal rights and ethics. We are not intending to make judgements

about ethics, but merely show how the architectural approach of this paper clarifies some issues. First, the axiomatic/kernel method aligns closely with the work of Crick and Koch (1998) in the sense that they have distinguished between input sections of the visual system where consciousness does *not* arise (due to lack of depiction) from 'exrastriate' areas where they argued consciousness could arise due to the link to the motor regions of the brain that makes depiction possible. Of importance is the fact that the brain maps used for the Crick/Koch work were those of a macaque monkey. It would then be hard to argue that if an animal's brain tells us about the neural correlates of human consciousness, that such an animal is not conscious. This implies that finding KAs in brains is concomitant with being conscious as this is what KAs do. It is of some interest that KAs may clearly be seen in invertebrates such as bees, but perhaps not in amoebae.

Questions of animal consciousness carry considerable controversy. It is believed by some that no matter how much one might attribute consciousness to animals (they are cute, bark at the right moment etc.) this attribution is without evidence. One of the starting points of this paper is that one cannot look for consciousness in behaviour. The introspection we each have of our own consciousness can only be attributed to animals if they are found to be endowed with the same physical support which we argue we need for our sense of consciousness. So, finding KAs may be valid in attributing consciousness where behaviour is not.

Higher order thought

Another argument sometime used against animal consciousness, is that of Higher Order Thought — HOT (Rosenthal, 1993). According to this, a sensation becomes conscious only if it is accompanied by a higher level awareness that the perception is happening. That is, in HOT theory, an organism has conscious thoughts only if they occur at these two levels. Some have argued that animals do not need this higher layer and are therefore not conscious.

To think clearly about what the KA suggests regarding HOT, it is necessary to introduce a way in which the KA might relate to language. Without going into detail, it is possible to imagine that the KA acts as a driver for some language generating machinery. In Fig. 5, the 'self' describing box is such an arrangement. Therefore, a statement such as 'I see a horse' implies that some decoder has caused the language machinery to compound an 'I see' statement (driven by the presence of activity in the visual awareness areas) with a decoding of the *content* of the two awareness areas. But any statement starting with 'I am conscious' is a statement of overall activity in the model irrespective of content. Followed by 'that I am seeing a horse', simply adds decodings of activity type ('I see' as opposed to, say, 'I hear') as well as the content of the awareness areas ('a horse'). According to axiomatic theory all the activities in the KA must all be present during a conscious act. A difference between the human and the animal is language and the human ability to choose how we use it. Taking language decoding away from a system makes it unable to make these differing pronouncements, but does not remove the KA activity hence consciousness.

In summary therefore, axiomatic theory and the KA suggest that statements about HOT do not imply a higher order of thought at all, they merely reflect a sophisticated linguistic ability which may indicate the presence of sophisticated content of consciousness but implies no special superiority in the presence of consciousness.

The kernel architecture, experience, sleep and dreaming

To give a highly non-rigorous account of how the KA helps in creating hypotheses about dreaming, we need to begin by taking note of the laying down of experience in the memory module as indicated by Eqs. (1) and (2) enunciated earlier in this paper. These show how an 'attractor' has been formed for state Ai' which is the stored experience for a world state Ai. Now assume that Ai' is experienced in the presence of Ai at which point a new world event Bi occurs. The next iconically formed event in the

memory module is Bi' and so on through Ci, ….. Adopting a simplified version of the previous style of showing learning equations we have

$$Ai' Bi \rightarrow Bi' \tag{3}$$

$$Bi' \; Bi \rightarrow Bi' \tag{4}$$

$$Ci' \; Bi \rightarrow CBi' \quad \ldots \text{ etc.} \tag{5}$$

Say now that the system has an arbitrary input and is in state Ai'. Due to the generalisation 'nearest neighbour' rule of the neurons there is equal probability the next is either Ai' (Eq. (2)) or Bi'. There is no return to Ai' so a change of state will then be to Ci etc. Therefore, it can be said that waking periods are about laying down memories of sensory inputs as state chains which can be followed during arbitrary sensory input.

But the state spaces of even the small neural areas are an exponent of two number of neurons. With some million of cells likely to be in the memory area of a living brain, this space is immense with respect to the memory trajectories of states that have been laid down. Also at the end of a thinking day, the system might be left in the last occurring thought just before going to sleep. We argue that a resetting process is in place. Theoretically this has to do with the fact that left inputless, or with some noisy input, an automaton with largely unformed transitions between states will seek to enter a set of states within which the probability for re-entrance to a state of the set is higher than entering states outside this set (Kaufmann (1969) talks of the extreme case of this 'stability at the edge of chaos' which occurs for two-input random-function cells). Here we call this the 'rest' set of states of the network or the 'rest region'. Two things happen in sleep in terms of the KA. First, the only modules that are not immobilised are the memory and the emotion modules. Immobilisation of the rest is seen as an evolved energy-saving property. Second, as we have postulated in Aleksander (2005), deep sleep virtually destroys the generalisation of the neurons (in addition to paralysing the periphery). But, it is known that the 'depth' of sleep, hence, this generalisation oscillates between near-awakedness and deep sleep four or five times each night. So, at times, with the periphery paralysed, the memory module has the opportunity of drifting towards some rest state.

The beneficial effect of this is that periods of awakedness during which experiences are laid down have a constrained set of states from which to start. This obviates the problem of making the experiences irretrievable (needle in the haystack problem) on awakening and of laying down the same experience twice in totally different parts of the state space. Figure 9 is a sketch of what we have been saying.

Here the rest region is shown as a set of states in the centre of the diagram. Each empty circle is a meaningless state, while a black-filled one is a meaningful, depictive state laid down during waking hours. The thick lines are the transitions between meaningful states created during waking hours as described. The thin lines indicate what might be happening during sleep, while the memory module is trying to return to the rest-state. It is not bound to the trajectories that occur when the perception module is active, but could enter meaningful states and trace some experience trajectory in a hap-hazard way. If woken up in one of these meaningful trajectories, the organism might report a bizarre story involving meaningful visited states in a non-meaningful sequence. This is what would then be called a dream.

Therefore the KA indicates that sleep is beneficial in resetting the memory module and making meaningful memories accessible on becoming awake, while dreaming is a by-product of this resetting process. In Aleksander (2005) attention was drawn to the fact that state trajectories could exist that have become detached from the rest state. Without going into detail, it is here that a discussion of Freud's 'unconscious' may begin, including the fact that such areas of state space may be reached in dreams.

Application to machines

Looking at the stated aims of doing computational studies of consciousness, we mentioned earlier that one needs to ask whether consciousness adds to

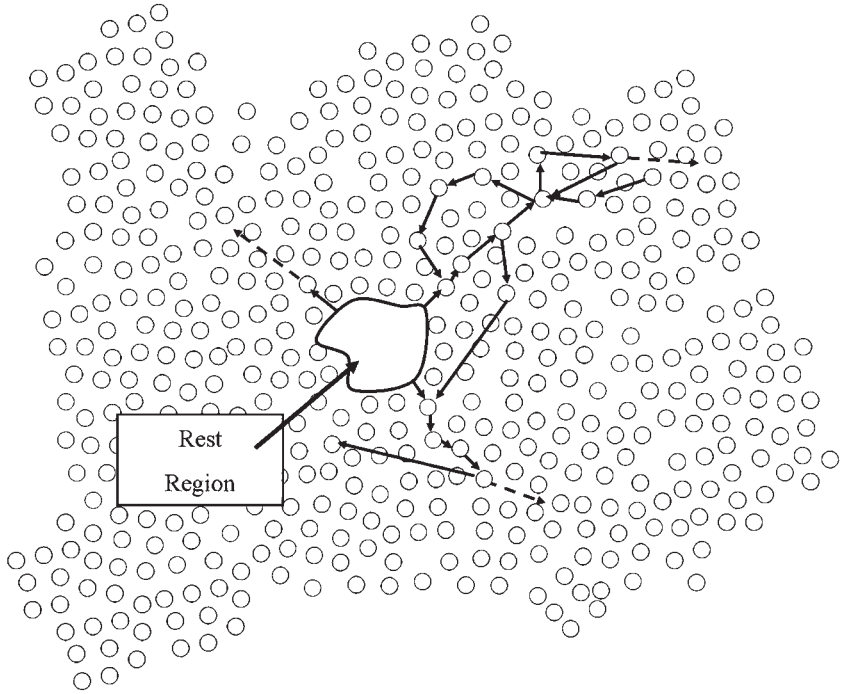


Fig. 9. State space of the memory automaton. Laid down experiences (thick), return to the rest region (thin) causing dreaming.

the performance of a constructed artefact. Shifting the KA into a robot is certainly interesting research (see Holland, 2006). However, it is not clear at the moment whether best performance can be achieved by conscious or conventionally programmed systems. Certainly conscious machines have a promise of far greater adaptation to complex worlds and greater autonomy than anything that can be achieved at the moment with direct computation. But there are many years of research and development ahead, before a clear view on this can be obtained.

Summary and conclusions

We have argued that studying the philosophical theme of consciousness through computer modelling adds a distinct point of view to this ancient pursuit. The approach has been thoroughly materialistic in the sense that the consciousness we know is assumed to arise from the machinery of the brain and that computational work can develop through what is known about the brain as a machine. Central to the computational study has been the notion of a virtual machine which is not only brain-like in character, but allows one to define a class of brain-like machines that have the property of being conscious in common. The paper has concentrated on a phenomenological approach that includes an explicit representation of personal sensation rather than alternative computational approaches in which the behaviour of the machine is a measure of the presence of consciousness.

The process starts with an introspective decomposition of consciousness into elements, or axioms of, presence in the world, imagined existence, choice of experience through attention, volition and emotion. It was argued that to keep a translation of these into computational schemes, a particular essential property of the computational system should be ‘depiction’, a mix of processing sensory information with muscular information that encodes source information about the sensory information. The five axioms were discussed as first-person sensations and then

investigated for their mechanistic implications through the use of a partly abstract computational architecture, the KA. This consists of four interlinked neural modules, two of which contribute directly to conscious sensation through being depictive. This covers the first two axioms while combinations of the modules cover the other three. Casting the explanation of this axiomatic approach in the context of the KA is the main contribution of this paper.

The application of the KA concept lies in it being the starting point for implementation. Here we have reviewed work done on visual awareness where the KA is computed using a neural simulation package (NRM) and embedded in eye-movement machinery to show that a depictive awareness can indeed be created in the overlap between imaginative and perceptual mechanisms. This model has also been helpful in addressing an age-old discussion regarding the unstable perception that arises in observing the Necker cube. A new theory for this has been based on the KA as embedded in a model of the visual system that includes the unconscious dorsal path. It was possible to demonstrate using the visual awareness model that this 'action' path may be where the instability takes place.

The latter part of the paper addresses some general puzzles that arise in the context of consciousness. It proposes that the discovery of KA-like mechanisms in an organism may be taken as evidence consciousness in that organism. This leads to the strong suggestion that consciousness is present in animals, as even bees show KA-like structures. Through this it was possible to suggest that theories of 'higher order thought' do not demand more than KA structures, they merely point to the sophistication of what KA structures do in humans. This is further evidence that, while animals may not indulge in 'higher order' speculations, this does not mean that they are devoid of consciousness. To end the paper, we argue that physical changes in sleep, bring about a 'resetting' action in the KA which allows a coherent laying down of new experiences in memory and, as a side effect, brings about the phenomenon of dreaming through accidental entry into meaningful memory states.

Will conscious machines supplant non-conscious ones? While this seems unlikely at the moment, introducing the KA into robot controllers makes for interesting research. The state-of-the-art at the moment, as demonstrated by this paper is that computational discourse and analysis is best suited to enrich philosophical and neurophysiological discussion on the nature of consciousness which will add clarity in the future to what it is for a robot to be conscious.

Appendix: the digital neuron

The above implementations, some of which contain over 1 million neurons, would not be possible without a simplification of the neuron function. At its simplest a neuron must record a wanted response to a training input and be able to generalise this to similar inputs. The algorithm for the neuron used in our implementations is:

A neuron maps a complete set of patterns $\{\boldsymbol{X}\}$ *in to* an output set $\{\boldsymbol{Z}\}$

Training on binary vectors: $X_1 \Rightarrow Z_1$, … $X_n \Rightarrow Z_n$

Testing on unknown input: $X_u \Rightarrow Z_j$

Where j is given by the Hamming-nearest X_1, … X_n to X_u

Within a variable 'spreading' limit S (variable generalisation)

If no clear nearest neighbour is found, Z_j is randomly selected.

References

Aleksander, I. (2005) The World in My Mind My Mind in the World: Key Mechanisms. Imprint Academic, Exeter.

Aleksander, I. and Dunmall, B. (2003) Axioms and tests for minimal consciousness in agents. J. Conscious. Stud., 10(4–5): 7–19.

Aleksander, I. and Morton, H. (2006) On architectures for synthetic phenomenology. Proceedings of AISB Symposium on Next Generation Machine Consciousness. Bristol, UK.

Aleksander, I., Morton, H. and Dunmall, B. (2001) Seeing is believing: depictive neuromodelling of visual awareness. Proc IWANN 2001, Granada. Heidelberg, Springer Verlag, Vol. LNCS 2084/2001: pp. 675–683.

Chalmers, D.J. (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, New York.

Clowes, R., Torrance, S. and Chrisley, R. (Eds.) (2007) Machine consciousness: embodiment and imagination. J. Conscious. Stud., 14(7).

Crick, F. and Koch, C. (1998) Consciousness and neuroscience. Cereb. Cortex, 8(2): 97–107.

Franklin, S. (2003) IDA, a conscious artefact? J. Conscious. Stud., 10(4–5): 47–66.

Gamez, D. (2006) The XML approach to synthetic phenomenology. Proceedings of AISB Symposium on Next Generation Machine Consciousness. Bristol, UK.

Gregory, R. (1997) Knowledge in perception and illusion. Philos. Trans. R. Soc. B, 352(1358): 1121–1127.

Haikonen, P. (2005) Artificial minds and conscious machines. In: Davies D. (Ed.), Visions of Mind Architectures for Cognition and Affect. ISP, Hull.

Holland, O. (2003) Editorial introduction. J. Conscious. Stud., 10: 4–5.

Holland, O. (2006) Artificial consciousness and the simulation of behaviour. Proceedings of AISB Symposium on Next Generation Machine Consciousness. Bristol, UK.

Kaufmann, SA. (1969) Metabolic stability and epigenesist in randomly constructed genetic nets. J. Theor. Biol., 22: 437–467.

Milner, D. and Goodale, M. (2006) The Visual Brain in Action (2nd ed.). Oxford Psychology Paperback, Oxford, UK.

Rosenthal, D. (1993) Thinking that one thinks. In: Davies M. and Humphreys G. (Eds.), Consciousness. Blackwell, Oxford, pp. 197–223.

Shanahan, M.P. (2005) Consciousness, emotion and imagination: a brain-inspired architecture for cognitive robotics. AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness. Bristol, UK, pp. 26–35.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 8

Identification of neuroanatomical substrates of set-shifting ability: evidence from patients with focal brain lesions

Pritha Mukhopadhyay[1,*], Aparna Dutt[1], Shyamal Kumar Das[2], Arindam Basu[3], Avijit Hazra[4], Tapan Dhibar[5] and Trishit Roy[2]

[1]*Department of Psychology, University College of Science and Technology, 92 APC Road, Calcutta 700009, India*
[2]*Department of Neuromedicine, Bangur Institute of Neuroscience and Psychiatry, Calcutta 700025, India*
[3]*Department of ENT, Guru Teg Bahadur Medical Centre, 11, D.L. Khan Road, Calcutta 700027, India*
[4]*Department of Pharmacology, Institute of Postgraduate Medical Education and Research, 244 B, A.J.C. Bose Road, Calcutta 700020, India*
[5]*Department of Neuroradiology, Bangur Institute of Neuroscience and Psychiatry, Calcutta, India*

Abstract: This work concerns the investigation of executive functions in patients with focal brain lesion. In order to identify the underlying substrates for executive functions, 54 patients with focal cortical ($n = 30$), subcortical ($n = 13$) and cerebellar damage ($n = 10$) (M = 9; F = 1) in the age range of 24–65 years with a minimum of Class V education have been investigated. The patients were admitted to the Department of Neuromedicine of Bangur Institute of Neurology, Calcutta. Each patient with focal lesion was matched with a healthy normal subject controlling for age and education. The socio-economic background was also taken into consideration. Controls were selected from the families of other patients admitted to the institution and also from individuals who volunteered to act as controls. Here too, rigid criteria have been followed to select the normals. Mini Mental State Examination (MMSE) and General Health Questionnaire (GHQ) were administered to screen out the neurological and psychiatric abnormalities in selection of normal control and Wisconsin Card Sorting Test (WCST) was administered to find out the executive function, in terms of set-shifting ability. Since standard anatomical groupings can obscure more specific brain–behavior relations, group-comparison design does not always allow determination of the effective lesion responsible for a particular deficit (Godefroy et al., 1998). The Classification and Regression Tree (CART) analysis has been used to determine the brain–behavior relationships. The result reveals that the frontal lobes are essential determinants of set-shifting capacity. However, for optimal execution of set-shifting function, the frontal lobes require participation of other cortical, subcortical and cerebellar regions. The result has been discussed in the light of the existing theories and research reports.

Keywords: executive function; set-shifting ability; perseverative error

*Corresponding author. Tel.: +91-33-25598022;
E-mail: pritha_m2@yahoo.com

DOI: 10.1016/S0079-6123(07)68008-X

Introduction

'Executive function' refers to 'higher' function of the prefrontal cortex (Milner, 1964; Luria, 1966; Stuss and Benson, 1986) and traditionally has been used synonymously with the term 'frontal skills'. The term 'executive skills' implies activities that help to adapt to one's environment. It is one's ability to plan, initiate, organize, sequence, monitor and shift cognitive set to achieve a goal. The function called 'cognitive set shifting' means ones ability to shift from one perceptual attribute or thought to another on the basis of feedback from changing environments. Cognitive set shifting is also known as 'attentional set shifting'.

The Wisconsin Card Sorting Test (WCST) has been considered as a key measure in the diagnosis of frontal lobe dysfunction (Milner, 1963, 1964; Drewe, 1974; Arnett et al., 1994). Cognitive set shifting is most often substantiated by the WCST. Successful performance on WCST requires certain capacities: cognitive flexibility, cognitive set persistence, concept identification, hypothesis generation and the ability to use response feedback information. Hence, set-shifting capacity is not a single unitary function and it's measure, the WCST, is a complex problem-solving task that requires multiple cognitive processes (Anderson et al., 1991; Dehaene and Changeux, 1991) most probably involving of brain regions other than the frontal lobes.

Evidence from neuropsychological, electrophysiological and functional neuroimaging research supports the role of frontal lobes as a critical node in the neuroanatomical network underlying set-shifting capacity. Different regions within the frontal lobes have been reported to underlie successful performance on the WCST (Milner, 1963; Drewe, 1974; Rogers et al., 2000; Monchi et al., 2001; Nagahama et al., 2001; Konishi et al., 2002). Reports are however contradictory regarding the lateralization of set-shifting function within the frontal lobes. While left dorsolateral prefrontal dominance (DLPFC) on WCST performance correlates with certain neuropsychological and neuroimaging studies (Milner, 1971; Berman and Weinberger, 1990; Grafman et al., 1990; Goldstein et al., 2004), right hemispheric lateralization has also been reported (Robinson et al., 1980; Owen et al., 1993; Haut et al., 1996; Stuss et al., 2000). Few studies however, did not indicate any lateralizing effects (Nelson, 1976; Bornstein, 1986).

WCST scores correlate with frontal lobe dysfunction in a substantial proportion of patients with non-frontal cerebral damage, though clinical and experimental research also justifies the assumption regarding the involvement of other regions of the brain in executive function. Hermann and Wyler (1988) and Drake et al. (2000) in their studies on patients with temporal lobe epilepsy, confirmed that damage to brain regions outside the prefrontal cortex may result in frontal lobe dysfunction as measured by performance on the WCST. Though laterality effect with poorer performance on WCST was associated with non-dominant temporal lobe pathology (Corcoran and Upton, 1993), no difference was observed on WCST performance between left and right temporal lobe pathology by other researchers (Trenerry and Jack, 1994; Hermann and Wyler, 1998; Martin et al., 2000).

Involvement of other non-frontal regions of the brain including inferior parietal lobule, the visual association and the inferior temporal cortices along with the DLPFC during WCST performance gets support from functional neuroimaging studies (Berman et al., 1995; Rogers et al., 2000).

The role of cortical–subcortical circuits in 'frontal' functions (Evarts et al., 1984; Cummings, 1993) is also significant. Subcortical structures like the basal ganglia have been implicated in WCST performance in both lesion and functional neuroimaging studies (Rogers et al., 2000; Monchi et al., 2001; Swainson and Robbins, 2001). Deficit in WCST performance has been observed following right basal ganglia stroke (Swainson and Robbins, 2001), dorsal and ventral caudate nucleus lesions (Mendez et al., 1989) and thalamic lesions (Ghika-Shchmid and Bogousslavsky, 2000; Annoni et al., 2003).

The role of the cerebellum in executive function has been suggested by Schmahmann and Sherman (1998) who identified the 'cerebellar cognitive affective syndrome' characterized by impairment of executive and other cognitive functions in patients with cerebellar lesions. Other lesion and

neuroimaging studies too have suggested the role of the cerebellum in executive function (Berman et al., 1995; Nagahama et al., 1996; Le et al., 1998; Levisohn et al., 2000).

It is apparent therefore that a unified neuroanatomical substrate for cognitive dysfunction is yet to emerge and there remains the need to explore the neural substrates underlying executive function. The present study aims to identify the neuroanatomical substrates underlying set-shifting ability in patients with focal brain lesions at various sites.

Methodology

The study included 54 patients of either sex with focal cortical ($n=31$), subcortical ($n=13$) and cerebellar damage ($n=10$). They were in the age range of 16–65 years and admitted to the Department of Neuromedicine of Bangur Institute of Neuroscience and Psychiatry, Calcutta. Ethical clearance for the study was obtained from the institutional ethics committee.

For inclusion, the patients had to be right handed, with a minimum education up to the fifth standard and suffering from their first cerebral, subcortical or cerebellar insult due to stroke or tumor. They had to have computerized tomography (CT) or magnetic resonance imaging (MRI) scans showing a single unilateral lesion involving a maximum of two anatomically distinct cerebral structures (in case of cerebral lesions) or restricted either to the basal ganglia or the thalamus (in case of subcortical lesion) or one of the cerebellar hemispheres (in case of cerebellar lesion).

Those with a history of past stroke, head injury, epilepsy, central nervous system infection, metabolic encephalopathies, neurodegenerative disorders like Alzheimer's disease or Parkinsonism, primary aphasia, significant motor weakness or major psychiatric disorders were excluded. Also excluded were subjects on sedative, neuroleptic or anticonvulsant medication and those with a history of alcohol or substance abuse. Significant disease of major organ systems such as hepatic, renal or pulmonary disease was a further exclusion criterion.

Each patient with a focal lesion was matched with a healthy control subject for age and education. The socioeconomic background was also taken into consideration. Controls were selected from families of other patients and also from individuals who volunteered. The control subjects had no history of neurological and psychiatric disorders, with Mini Mental State Examination (MMSE) score above 25 and a General Health Questionnaire (GHQ) score compatible with the cut-off point for normality. It was also ensured that they were not on any drugs known to affect the central nervous system function and had no history of alcohol or substance abuse.

The following tests/questionnaires were applied:

1. Wisconsin Card Sorting Test (Heaton et al., 1993).
2. Mini Mental State Examination (Folstein et al., 1975).
3. General Health Questionnaire (Goldberg and Hiller, 1979).
4. Edinburgh Handedness Inventory (Oldfield, 1971).

Majority of the patients were admitted to the Department of Neuromedicine of Bangur Institute of Neuroscience and Psychiatry, Calcutta. Clinical characteristics were assessed through careful history and neurological examination by the attending neurologist. Neuroimaging findings were interpreted by the neurologist and a neuroradiologist and a consensus opinion taken. Thereafter, informed consent was obtained from the patient. Each control subject also underwent neurological examination by the neurologist. Both patients and controls were screened for their handedness with the Edinburgh Handedness Inventory and screened for cognitive dysfunction with the MMSE.

The WCST was employed according to standardized procedures (Heaton et al., 1993) to assess executive function in terms of set-shifting abilities. The test measures the ability to change strategy in response to altered feedback about performance. In this test, the subject was asked to match a series of response cards to one of the four stimulus cards. The WCST response and stimulus cards vary along three dimensions: color of the elements,

form of elements and number of the elements on each card. The subject was not told explicitly as to which dimension correctly represented the sorting principle, but after each attempted match of a response card to a stimulus card, the examiner informed the subject whether or not his/her preceding response was correct. By generating hypotheses about the correct sorting principle, and testing these hypotheses through trial and error, the subject was able to determine whether he/she is to sort the cards according to color, form or number. Following 10 consecutive correct responses ('completion of a category'), the examiner covertly changed the sorting principle. The test was continued until the subject completed 6 categories or until all the 128 cards were used, whichever came first.

In the present study, performance on the WCST was evaluated by (a) the number of categories completed and by two measures of perseverative tendencies, (b) percent perseverative responses and (c) percent perseverative errors.

When the subject persisted in responding to a stimulus characteristic (color, form or number) that was previously in operation but which had been changed subsequently, the response was considered to match the 'perseverated-to' principle and was scored as perseverative. Once a perseverated-to-principle has been established and is operative, responses that match the perseverated-to-principle are scored as perseverative regardless of whether they are correct or incorrect. There are three rules for detecting perseveration, the details of which have been outlined in the WCST manual (Heaton et al., 1993). Perseverative responses include both errors and correct responses satisfying the perseverated-to-principle according to one of the three rules of perseveration. The total number of perseverative responses is divided by the total number of trials administered and multiplied by hundred to get the 'percent perseverative response' score. Perseverative errors are those responses which are incorrect and also satisfy the perseverated-to-principle according to one of the three rules of perseveration. The total number of perseverative errors is divided by the total number of trials administered and multiplied by 100 to get the 'percent perseverative error' score. The raw percent perseverative response and percent perseverative error scores are transformed to normalized standard scores. Standard scores on WCST range from ≤ 54 to ≥ 107.

Statistical analysis

This was carried out using SPSS version 11.0 software. A p value less than 0.05 was used to determine significance. Since age and education had a normal distribution in the study population, comparison between groups with respect to these two variables was assessed by parametric tests. Non-parametric tests were chosen for neuropsychological variables as no assumption regarding their distribution was made and the sample size was relatively small.

Standard anatomical groupings can obscure more specific brain–behavior relations. Group comparison design does not allow determination of the effective lesion responsible for a particular deficit (Godefroy et al., 1998). The classification and regression tree (CART) statistical procedure (Breiman et al., 1984) used in several brain–behavior relationship studies (Stuss et al., 2000; Alexander et al., 2003) has been found to be more successful in the determination of the nature of brain–behavior relationships.

We employed the R software for statistical analysis (2007) for CART analysis. The CART procedure, as a supervised learning algorithm, runs a recursive regression tree-like structure searching for the best classifier of a specific outcome variable based on a set of predictor variables. In this study, CART was employed defining lesion sites as the outcomes variables and WCST test scores as the predictor variables. We used the test scores to find a set of classifiers to differentiate between lesions. The purpose of CART was thus to enable prediction of lesion sites based on test scores.

Results

Comparison between cortical lesion patient group and control

The impairment in the cortical lesion group compared to control is evident from Table 1. Performance was significantly impaired in all the

Table 1. Comparison of cortical lesion patients and matched normal controls on the Wisconsin Card Sorting Test

Test parameter	Normal controls ($n=31$)	Cortical lesion patients ($n=31$)	p value
Number of categories completed	5.61±0.80	2.65±1.64	<0.01
Perseverative responses	102.35±14.42	77.16±16.85	<0.01
Perseverative errors	100.74±12.90	73.16±19.12	<0.01

Note: Values represent mean±standard deviation.

Table 2. Comparison of subcortical lesion patients and matched normal controls on the Wisconsin Card Sorting Test

Test parameter	Normal controls ($n=13$)	Subcortical lesion patients ($n=13$)	p value
Number of categories completed	5.42±1.4	3.23±1.24	<0.01
Perseverative responses	100.75±11.17	84.08±9.40	<0.01
Perseverative errors	99.39±9.79	83.85±10.11	<0.01

Note: Values represent mean±standard deviation.

Table 3. Comparison of cerebellar lesion patients and matched normal controls on the Wisconsin Card Sorting Test

Test parameter	Normal controls ($n=10$)	Cerebellar lesion patients ($n=10$)	p value
Number of categories completed	5.3±0.82	2.90±1.91	<0.01
Perseverative responses	96±14	84.00±15.83	NS
Perseverative errors	96.7±14.89	84.00±13.30	<0.05

Note: Values represent mean±standard deviation; NS, not significant.

variables of WCST — the patients achieved significantly less number of categories, showed more perseverative responses and made more perseverative errors.

Comparison between subcortical lesion patient group and control

This is depicted in Table 2. Patients with subcortical lesions also performed poorly compared to controls, achieving significantly fewer categories, showing more percent perseverative responses and making more percent perseverative errors.

Comparison between cerebellar lesion patient group and control

Comparison of WCST performance of patients with cerebellar lesions and control subjects is presented in Table 3. However, here there was no significant difference in percent perseverative responses, although these patients made significantly more percent perseverative errors and achieved significantly lesser number of categories than controls.

CART analysis on percent perseverative responses

On the basis of percent perseverative responses (Fig. 1), the CART procedure initially provided the first split on the percent perseverative responses score at 80. Those who had scores equal to or greater than 80 were then classified on the basis of whether the score was >86.5. CART identified left cerebellar lesions (LCB) as scores with more than 86.5, and left temporoparietal lesions (LTP) as having scores between 80 and 86.5. Those who had scores less than 80 were classified into left frontal (LF) and right thalamic (RTH) lesion groups on the basis of scores at less than 72.5 for LF lesions.

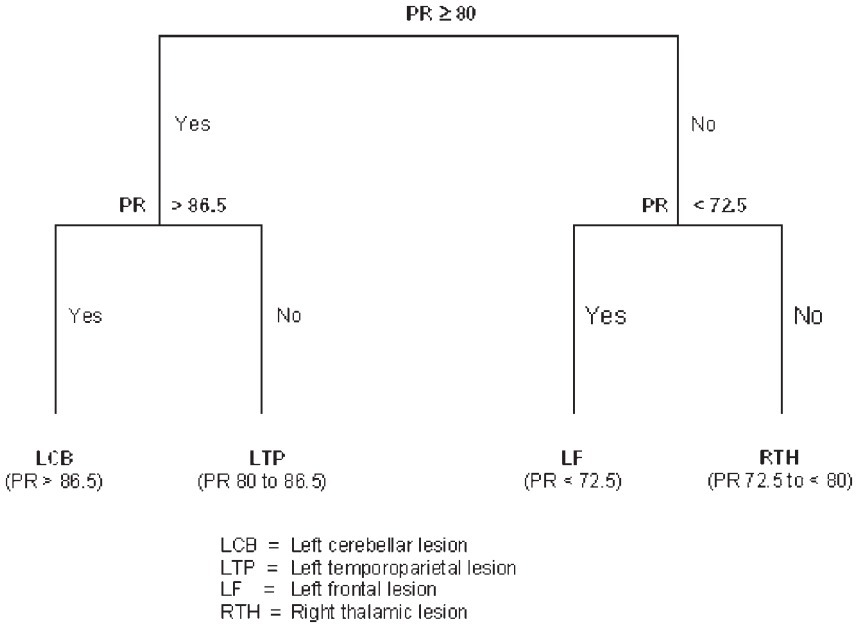


Fig. 1. Classification and regression tree (CART) analysis output for perseverative response scores (PR).

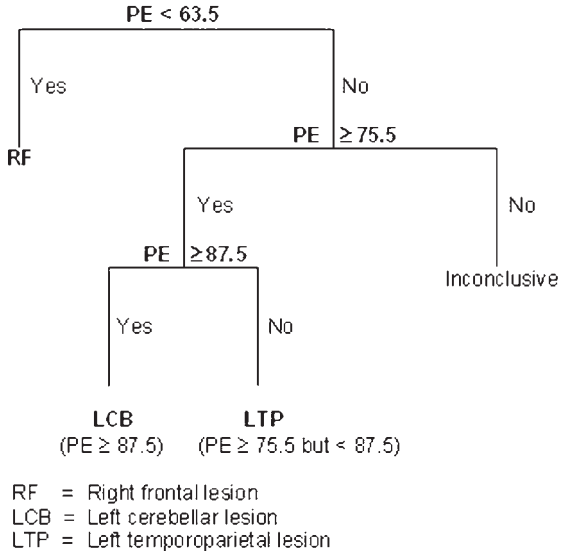


Fig. 2. Classification and regression tree (CART) analysis output for perseverative error scores (PE).

CART analysis on percent perseverative errors

On the basis of percent perseverative error scores (Fig. 2), the CART procedure classified right frontal (RF) from the rest of the lesions on the basis of a score less than 63.5. LTP lesions scored between 75.5 to less than 87.5. Scores equal to or greater than 87.5 distinguished LCB lesions.

Discussion

In this study, patients with focal brain lesions, irrespective of the lesion site, performed poorly on the WCST parameters of category completion, percent perseverative responses and percent perseverative errors, in contrast to their healthy counterparts. Poor function on the aforesaid variables perhaps restricted the patients from operating on the feedback provided by the examiner. Utilization of feedback from the examiner is the basis on which the performer takes decision to shift set from one category to other based on color, form and number on WCST task. Poor utilization of feedback thereby revealed the impairment in set-shifting capacity in each of the three lesion groups in terms of the reduced numbers of categories achieved by them.

The results of CART analysis in the present study revealed the following pattern based on perseverative response score: LF < RTH < LTP < LCB. Indirectly this indicates that lesions in these four anatomical zones are involved in perseverative response, with frontal lesions perhaps being the most damaging and cerebellar lesions the least. We did not find authoritative references on influence of LF lesions on perseverative responses. Possible involvement of other regions suggests that the frontal lobe is not the only site to explain perseverative response. Rather, at least these four regions may form part of the cortico–subcortical–cerebellar network that can explain perseverative responses.

Involvement of different regions of the brain in perseverative response can be explained by a concept of executive function that postulates set shifting to be a multifaceted rather than a unitary entity. Since set shifting involves multiple cognitive components, it may be presumed that each of these independent components are subserved by different anatomical circuits in the brain and not by the frontal lobes alone. Other recent studies have also indicated that patients with non-frontal lesions exhibit significant difficulty on the WCST (Drake et al., 2000; Ghika-Shchmid and Bogousslavsky, 2000; Levisohn et al., 2000; Annoni et al., 2003).

The association of LF lesions with perseverative responses could be attributed to the requirement of a WCST task to involve verbal mediation for its optimal performance. The findings of Berman et al. (1995) and Nagahama et al. (1996) suggest that although WCST is a visual task and essential visual processes are mediated by non-verbal systems in the right prefrontal regions, set shifting in WCST could require participation of the verbal systems in the left hemisphere also. The involvement of the LTP region further substantiates the reason put forward for explaining left hemispheric participation in perseverative response.

Our observation suggests perseverative responses following RTH lesion. However, this is at variance with previous studies where the most pronounced impact on executive function has been observed in bilateral and left unilateral thalamic infarctions (Rousseaux, 1994). On the other hand Van der Werf et al. (1999) have reported a case of RTH involvement. On the WCST, a 44-year-old patient with right lacunar thalamic infarction was unable to sort according to color and therefore was unable to achieve a single category. Their findings correlated with hypoperfusion of the right frontal cortex on brain single photon emission CT, suggesting that dysexecutive symptoms resulted from disconnection of the prefrontal cortex from the brainstem activating nuclei due to the strategic localization of the right thalamic infarction.

The propensity for patients with cortical or subcortical lesions to make perseverative errors in comparison to healthy controls corroborates previous reports (Drewe, 1974; Owen et al., 1993; Stuss et al., 2000; Goldstein et al., 2004) on this aspect of executive function. The cerebellar lesion group also made perseverative errors. Levisohn et al. (2000) made a similar observation in children who underwent resection of cerebellar tumors. The authors attributed the greater perseverative errors in such children to their difficulty in shifting attention. The role of cerebellum in shifting attention has also been suggested by Akshoomoff and Courchesne (1994) in a functional MRI study that revealed cerebellar activation during non-spatial shifting of selective attention. The authors suggested that the cerebellar cortex participates in the ''rapid sequential changes and adjustment of neural activity to proceed from one condition to another'', while the neocerebellar participation in attention arises from the need to predict, prepare for and adjust to imminent information acquisition, analysis or action (Allen et al., 1997).

On the basis of perseverative error score, CART could identify RF, LTP and LCB as three anatomical sites (with the pattern RF < LTP < LCB) where lesions could result in variations in perseverative error scores. Of the three, those with RF lesions were distinguished from the other sites by the lowest scores, suggesting that these patients may have difficulty in employing effective mechanisms to inhibit previously learnt contextual rules. Earlier studies have also indicated perseverative errors as the main signs of frontal dysfunction (Robinson et al., 1980; Owen et al., 1993; Haut et al., 1996; Barcelo and Knight, 2002). Some studies indicate the left hemispheric effect (Drewe, 1974;

Goldstein et al., 2004). Stuss et al. (2000) have demonstrated that patients with either left or right focal frontal lesion are impaired on the 'perseveration to preceding criterion' score in the WCST, but the right prefrontal group was impaired more severely than the left prefrontal group. The 'perseveration to preceding criterion' score is similar to perseverative errors as scored by Heaton et al. (1993) and as followed in the present work. Previous studies have suggested that the tendency of RF lesion patients to make more perseverative errors than the LF lesion patients may be reflective of the greater sustained attention and monitoring role of the right frontal lobe (Sandson and Albert, 1987; Wilkins et al., 1987). Functional neuroimaging studies also suggest RF predominance in the process of attentional set shifting (Monchi et al., 2001; Nagahama et al., 2001).

Observations from this study thus suggest that perseveration is possibly a bilateral phenomena resulting from the loss of integrity of the frontal lobes along with non-frontal cortical, subcortical and cerebellar regions. The involvement of non-frontal cortical regions gets support from positron emission tomography findings of Berman et al. (1995) who reported activation of a complex network of regions involving inferior parietal lobule, visual association cortices and the inferolateral temporal cortices in addition to the prefrontal cortex during the performance of WCST in normal subjects. The same study also showed activation of portions of the cerebellum during WCST.

The significant difficulty of frontal patients to inhibit previous incorrect responses which results in perseverative behavior is possibly due to interference in the presumed inhibitory role of the dopaminergic pathways in the prefrontal cortex. Roberts et al. (1994) have hypothesized that attentional set shifting is mediated by a balanced interaction of prefrontal and striatal dopaminergic activity, with enhanced shifting following depressed prefrontal dopaminergic function and impaired shifting following elevated prefrontal dopaminergic or depressed striatal dopaminergic function. The prefrontal cortex mediated inhibitory control of cortical and subcortical regions gains support from the physiological viewpoint through event-related potential (ERP) studies (Stuss and Knight, 2002).

The discordance in frontal lobe function or frontal-like performance following lesions in any part of the brain can be explained from the cognitive viewpoint by the dynamic filtering theory (Shimamura, 2000) which suggests dynamic interplay of selecting, maintaining, updating and rerouting between the prefrontal cortex and regions in the posterior cortex through feedforward and feedback activations. The prefrontal cortex orchestrates these signals by maintaining certain activations and inhibiting others. As such, the prefrontal cortex refines cortical activity by increasing signal to noise ratio. Possibly, this modulation could extend beyond the cortex to the subcortical and cerebellar regions.

Conclusion

The frontal lobes are the essential determinants of set-shifting capacity. The present findings are in conformity with reports of neuropsychological studies (Milner, 1963, 1964; Drewe, 1974; Janowsky et al., 1989; Arnett et al., 1994) which indicate that set-shifting ability as measured by WCST is predominantly affected by frontal lobe lesions. The present observations are also in line with more objective data obtained through functional neuroimaging studies which have confirmed the involvement of the prefrontal cortex in the performance of the WCST (Nagahama et al., 1996, 2001; Rogers et al., 2000; Monchi et al., 2001; Konishi et al., 2002). However, the frontal lobes need to be in tune with other cortical, subcortical and cerebellar regions for optimal execution of a complex function like set-shifting ability.

References

Akshoomoff, N.A. and Courchesne, E. (1994) ERP evidence for a shifting attention deficit in patients with damage to the cerebellum. J. Cogn. Neurosci., 6: 388–399.

Alexander, M.P., Stuss, D.T. and Fansabedian, N. (2003) California Verbal Learning Test: performance by patients with focal frontal and non-frontal lesions. Brain, 126: 1493–1503.

Allen, G., Buxton, R.B., Wong, E.C. and Courschesne, E. (1997) Attentional activation of the cerebellum independent of motor movement. Science, 275: 1940–1943.

Anderson, S.W., Damasio, H., Jones, R.D. and Tranel, D. (1991) Wisconsin Card Sorting Test performance as a measure of frontal lobe damage. J. Clin. Exp. Neuropsychol., 13: 909–922.

Annoni, J.M., Gramigna, S., Staub, F., Carota, A., Maeder, P. and Bogousslavsky, J. (2003) Chronic cognitive impairment following laterothalamic infarcts. Arch. Neurol., 60: 1439–1443.

Arnett, P.A., Rao, S.M., Bernardin, L., Grafman, J., Yetkin, F.Z. and Lobeck, L. (1994) Relationship between frontal lobe lesions and Wisconsin Card Sorting Test performance in patients with multiple sclerosis. Neurology, 44: 420–425.

Barcelo, F. and Knight, R.T. (2002) Both random and perseverative errors underlie WCST deficits in prefrontal patients. Neuropsychologia, 40: 349–356.

Berman, K.F. and Weinberger, D.R. (1990) Lateralisation of cortical function during cognitive tasks: regional cerebral blood flow studies of normal individuals and patients with schizophrenia. J. Neurol. Neurosurg. Psychiatry, 53: 150–160.

Berman, K.F., Ostrem, J.L., Randolph, C., Gold, J., Goldberg, T.E. Coppola, R. et al. (1995) Physiological activation of a cortical network during performance of the Wisconsin Card Sorting Test: a positron emission tomography study. Neuropsychologia, 33: 1027–1046.

Bornstein, R.A. (1986) Contributions of various neuropsychological measures to detection of frontal lobe impairment. Int. J. Clin. Neuropsychol., 8: 18–22.

Breiman, L., Friedman, J.H., Oshen, R.A. and Store, C.J. (1984) Classification and regression trees. Wadsworth International Group, Belmont, CA.

Corcoran, R. and Upton, D. (1993) A role for the hippocampus in card sorting. Cortex, 29: 293–304.

Cummings, J.L. (1993) Frontal-subcortical circuits and human behaviour. Arch. Neurol., 50: 873–880.

Dehaene, S. and Changeux, J.-P. (1991) The Wisconsin Card Sorting Test: theoretical analysis and modeling in a neural network. Cereb. Cortex, 1: 62–79.

Drake, M., Allegri, R.F. and Thompson, A. (2000) Executive cognitive alteration of prefrontal type in patients with mesial temporal lobe epilepsy. Medicina (Buenos Aires), 60: 453–456.

Drewe, E.A. (1974) The effect of type and area of brain lesion on Wisconsin Card Sorting Test performance. Cortex, 10: 159–170.

Evarts, E.V., Kimura, M., Wurtz, R.H. and Hikosaka, O. (1984) Behavioral correlates activity in basal ganglia neurons. Trends Neurosci., 7: 447–453.

Folstein, M.F., Folstein, S.E. and McHugh, P.R. (1975) 'Mini-mental state': a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res., 12: 189–198.

Ghika-Shchmid, F. and Bogousslavsky, J. (2000) The acute behavioural syndrome of anterior thalamic infarction: a prospective study of 12 cases. Ann. Neurol., 48: 220–226.

Godefroy, O., Duhamel, A., Leclerc, X., Saint Michel, T., Henon, H. and Leys, D. (1998) Brain–behaviour relationships: some models and related statistical procedures for the study of brain-damaged patients. Brain, 121: 1545–1556.

Goldberg, D.P. and Hiller, V.E. (1979) A scaled version of the General Health Questionnaire. Psychol. Med., 9: 139–146.

Goldstein, B., Obrzut, J.E., John, C., Ledakis, G. and Armstrong, C.L. (2004) The impact of frontal and non-frontal brain tumor on Wisconsin Card Sorting Test performance. Brain Cogn., 54: 110–116.

Grafman, J., Jonas, B. and Salazar, A. (1990) Wisconsin Card Sorting Performance based on location and size of anatomical lesion in Vietnam veterans with penetrating head injury. Percept. Mot. Skills, 71: 1120–1122.

Haut, M.W., Cahill, J., Cutlip, W.D., Stevenson, J.M., Makela, E.H. and Bloomfield, S.M. (1996) On the nature of Wisconsin Card Sorting Test performance in schizophrenia. Psychiatry Res., 65: 15–22.

Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G. and Curtiff, G. (1993) Wisconsin Card Sorting Test Manual. Revised and Expanded. Psychological Assessment Resources, Inc., Odessa, Bethesda, FL, USA.

Hermann, B.P. and Wyler, A.R. (1988) Effects of anterior temporal lobectomy on language function: a controlled study. Ann. Neurol., 23: 585–588.

Janowsky, J.S., Shimamura, A.P., Kritchevsky, M. and Squire, L.R. (1989) Cognitive impairment following frontal lobe damage and its reference to human amnesia. Behav. Neurosci., 103: 548–560.

Konishi, S., Hayashi, T., Uchida, I., Kikyo, H., Takahashi, E. and Miyashita, Y. (2002) Hemispheric asymmetry in human lateral prefrontal cortex during cognitive set shifting. Proc. Natl. Acad. Sci. U.S.A., 99: 7803–7808.

Le, T.H., Pardo, J.V. and Hu, X. (1998) T-fMRI study of nonspatial shifting of selective attention: cerebellar and parietal contributions. J. Neurophysiol., 79: 1535–1548.

Levisohn, L., Cronin-Golomb, A. and Schmahmann, J.D. (2000) Neuropsychological consequences of cerebellar tumor resection in children: cerebellar cognitive affective syndrome in paediatric population. Brain, 123: 1041–1050.

Luria, A.R. (1966) Higher cortical functions in man (transl. B. Haigh). Basic Books, New York.

Martin, R.C., Sawrie, S.M., Gilliam, F.G., Palmer, C.A., Faught, E. Morawetz, R.B. et al. (2000) Wisconsin Card Sorting Performance in patients with temporal lobe epilepsy: clinical and neuroanatomical correlates. Epilepsia, 41: 1626–1632.

Mendez, M.F., Adams, N.L. and Lewandowski, K.S. (1989) Neurobehavioural changes associated with caudate lesions. Neurology, 39: 349–354.

Milner, B. (1963) Effect of different brain lesions on card sorting. The role of the frontal lobes. Arch. Neurol., 9: 100–110.

Milner, B. (1964) Some effects of frontal lobotomy in man. In: Warren J.A. and Akert G. (Eds.), The Frontal Granular Cortex and Behavior. McGraw-Hill, New York, pp. 313–334.

Milner, B. (1971) Interhemispheric differences in the localization of psychological processes in man. Br. Med. Bull., 27: 272–277.

Monchi, O., Petrides, M. and Petre, V. (2001) Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. J. Neurosci., 21: 7733–7741.

Nagahama, Y., Fukuyama, H., Yamauchi, H., Matsuzaki, S., Konishi, J. Shibasaki, H. et al. (1996) Cerebral activation during performance of a card sorting test. Brain, 119: 1667–1675.

Nagahama, Y., Okada, T., Katsumi, Y., Hayashi, T., Yamauchi, H. Oyanagi, C. et al. (2001) Dissociable mechanisms of attentional control within the human prefrontal cortex. Cereb. Cortex, 11: 85–92.

Nelson, H.E. (1976) A modified card sorting test sensitive to frontal lobe defects. Cortex, 12: 313–324.

Oldfield, R.C. (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia, 9: 97–113.

Owen, A.M., Roberts, A.C. and Hodges, J.R. (1993) Contrasting mechanisms of impaired attentional set shifting in patients with frontal lobe or Parkinson's disease. Brain, 116: 1159–1175.

R software for statistical analysis [homepage on the Internet] [cited 2007 Feb. 26]. Available from: http://www.r-project.org

Roberts, A.C., De Salvia, M.A., Wilkinson, L.S., Collins, P., Mur, J.L. Everitt, B.J. et al. (1994) 6-hydroxydopamine lesions of the prefrontal cortex enhanced performance on an analog of the Wisconsin Card Sorting test: possible interactions with subcortical dopamine. J. Neurosci., 14: 2531–2544.

Robinson, A.L., Heaton, R.K., Lehman, R.A. and Stilson, D.W. (1980) The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. J. Consult. Clin. Psychol., 48: 605–614.

Rogers, R.D., Andrews, T.C. Grasby, P.M. et al. (2000) Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. J. Cogn. Neurosci., 12: 142–162.

Rousseaux, M. (1994) Amnesias following limited thalamic infarctions. In: Delacour J. (Ed.), The Memory System of the Brain (Adv. Series Neurosci. Vol. 4). World Scientific, Singapore, pp. 241–277.

Sandson, J. and Albert, M.L. (1987) Perseveration in behavioural neurology. Neurology, 37: 1736–1741.

Schmahmann, J.D. and Sherman, J.C. (1998) The cerebellar cognitive affective syndrome. Brain, 121: 561–579.

Shimamura, A. (2000) The role of prefrontal cortex in dynamic filtering. Psychobiology, 28: 207–218.

Stuss, D.T. and Benson, D.F. (1986) The Frontal Lobes. Raven Press, New York.

Stuss, D.T. and Knight, R.T. (2002) Principles of Frontal Lobe Function. Oxford University Press, New York.

Stuss, D.T., Levine, B., Alexander, M.P., Hong, J., Palumbo, C. Hamer, L. et al. (2000) Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia, 38: 388–402.

Swainson, R. and Robbins, T.W. (2001) Rule-abstraction deficits following a basal ganglia lesion. Neurocase, 7: 433–443.

Trenerry, M. and Jack, Jr., C.R. (1994) Wisconsin Card Sorting Test performance before and after temporal lobectomy. J. Epilepsy, 7: 313–317.

Van der Werf, Y.D., Weerts, J.G.E., Jolles, J., Witter, M.P., Lindeboom, J. and Scheltens, P. (1999) Neuropsychological correlates of a right unilateral lacunar thalamic infarction. J. Neurol. Neurosurg. Psychiatry, 66: 36–42.

Wilkins, A.J., Shalllice, T. and McCarthy, R. (1987) Frontal lesions and sustained attention. Neuropsychologia, 25: 359–365.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 9

Thinking is believing

Rajesh Kasturirangan*

National Institute of Advanced Study, Indian Institute of Science Campus, Bangalore 560012, India

Abstract: Philosophers as well lay people often think of beliefs as psychological states with dubious epistemic properties. Beliefs are conceptualized as unregulated conceptual structures, for the most part hypothetical and often fanciful or deluded. Thinking and reasoning on the other hand are seen as rational activities regulated by rules and governed by norms. Computational modeling of the mind has focused on rule-governed behavior, ultimately trying to reduce them to rules of logic. What if thinking is less like reasoning and more like believing? I argue that the classical model of thought as rational is mistaken and that thinking is fundamentally constituted by believing. This new approach forces us to re-evaluate classical epistemic concepts like ''truth'', ''justification'' etc. Furthermore, if thinking is believing, then it is not clear how thoughts can be modeled computationally. We need new mathematical ideas to model thought, ideas that are quite different from traditional logic-based mathematical structures.

Keywords: thoughts; beliefs; stories; mathematical modeling

Introduction

Normally, when your mother comes and tells you that your friend is at the door, you believe her. In fact, not only do you believe her, you feel like you *know* that your friend is at the door. In other words, ordinary linguistic communication is often seen by us as a source of knowledge. Philosophers mostly disagree with this situation — they think that hearsay is too unstable a source of knowledge. If you did not have a means of testing your beliefs, then how can you possibly call that belief knowledge? Strangely enough, such a stringent criterion for knowledge would disqualify beliefs that we, as scientists, would consider the most secure. For example, most of us have not done a Michelson–Morley type experiment, nor are we about to do so anytime soon. Therefore, by the above stringent criterion, we do not know that the speed of light is a constant. In fact, according to this criterion, *Einstein did not know* that the speed of light is a constant since he never did the relevant experiments. I think we should find such skepticism about the role of language in knowledge problematic.

So, what does the philosopher claim as being a good source of knowledge? Most philosophers, East and West would agree that perception is a reputable source of knowledge, which is indicated in the classical motto ''seeing is believing''. While there is good reason to doubt the veridicality of perceptions (since there are so many perceptual illusions), most philosophers would agree that perception gives us as good knowledge of the external world as we are going to get. Apart from perception, reason is also seen as a solid source of

*Corresponding author. Tel.: +91-80-22185000;
Fax: +91-80-23606634; E-mail: rkasturi@gmail.com

DOI: 10.1016/S0079-6123(07)68009-1

knowledge. There are two broad kinds of reasoning, *deductive* and *inductive*. Despite the efforts of Godel and company, we tend to think that deductive reasoning is extremely secure, as long as we start from sound axioms. Induction is another matter altogether. Suppose I see some water in front of my house in the morning and then, on going out on the street, I notice water everywhere, I might conclude that it rained heavily the previous night. However, it might be that the water mains broke and the flooding was a result of the water leak, not rainfall. On these grounds, induction may be doubted as a source of knowledge. At best, the philosopher would say, induction gives you beliefs that have some justification, but they are not indubitable. If you thought that induction caused too many problems, the situation regarding beliefs is actually much worse. In my opinion, most patterns of belief propagation are not based on logic at all, whether deductive or inductive. What do I mean? Take the following two sentences:

(1) When Ramesh was driving to his house he remembered that he was supposed to be at a meeting and he immediately turned around and drove back to his office.
(2) When Ramesh was driving to his house he got stuck in a traffic jam behind an old red Maruti.

The first statement has an inferential form, where the belief "I must be at the office, but I am at my house, therefore I should drive back", but what about the second? It only has a temporal order, and there is no sense in which "there is an old Maruti" follows logically from "John was caught in a traffic jam". Yet, most of our oral and written communication is of this form, for the second sentence has a recognizable narrative structure and sounds plausible enough to our ears, unlike the third sentence below.

(3) When Ramesh was driving to his house, he turned into a dragon and ate a couple of pedestrians.

What makes the first two sentences acceptable in ordinary discourse, while not the third? In other words, what must our thought patterns be, so that we feel like we learn something when a friend tells us a story beginning with sentence (2)?

In this paper, I want to take seriously the claims that

(a) Our thoughts are more like beliefs, in that they can rarely be defined and captured in logical form and furthermore, have emotional components that are outside what we normally term "rational".
(b) Nevertheless, these thoughts/beliefs of ours usually constitute knowledge, knowledge that is fallible to be sure, but knowledge nonetheless.
(c) Our thought patterns have a narrative structure which is far more general than the inferential or deductive structures normally studied by logicians.
(d) Nevertheless, these narrative structures implicit in thought patterns can be studied formally using mathematical tools.

In the subsequent sections, I will make each one of the above claims as explicit as possible and then conclude with a short expose of the mathematical tools needed to model the patterns involved in narrative analysis.

Preliminaries

Physicists often start their investigations by doing a qualitative analysis of the basic features of the problem, typically involving simple calculations. The purpose of these calculations is to give the researcher a sense for the lay of the land, to reveal the basic features of the terrain, before he or she embarks upon a more complicated analysis. Sometimes this kind of qualitative analysis goes by the name of "dimensional analysis". My goal in this paper is to do the same for beliefs, i.e., to lay out the basic dimensions involved in modeling beliefs and then show how these principles can be implemented in simple back of the envelope type calculations that can tell us quite a bit about the underlying structure of beliefs. We could take as a provisional hypothesis, the following statement:

Hypothesis: Beliefs evolve in the interactions between the individual, community and society at large.

Therefore, the natural question is: what are the forces that influence the dynamics of belief change over the multiple scales of individual, community and society? The case of research in language is instructive. For example, we do not want to believe that beliefs are purely learnt. Without general psychological mechanisms, it is hard to see how beliefs could be acquired at all by individuals (Fodor, 1975). On the other hand, we do not think beliefs are entirely in our brains, otherwise how would different cultures have different beliefs (Lakoff, 1987; Varela et al., 1991)? So the million dollar questions remain

(a) What cognitive mechanisms underlie our ability to form, manipulate and evaluate beliefs?
(b) How do these mechanisms interact with community and cultural factors and give rise to the belief systems that characterize particular social groups?

This paper is an attempt to lay out a methodology to approach these two questions. For the theorist, the domain of beliefs presents an embarrassment of riches. One could say that the road to heaven is paved with various theories of beliefs. We have *quantitative* rational models of beliefs (as studied in economics, political science and sociology), we have normative accounts of beliefs that we *should be having* (the province of philosophy) and then, there are the *descriptive* approaches to modeling beliefs as seen in the various psychological, sociological, anthropological and historical accounts of the *origins* of beliefs. My goal is to combine the advantages of the two approaches in one package, namely, the quantitative power of the rational models and the etiological insight of the psychological and historical models and to ask "can we develop a formal framework to study the origins and dynamics of beliefs"? The main contribution of this paper is to answer the above question in the affirmative.

Normally, when people study beliefs quantitatively, they are interesting either in evaluating the rationality of the beliefs relative to the actions in the world they represent, or study how beliefs influence the choices that people make. Similarly, beliefs change, i.e., how people shift from believing one thing to another is studied in terms of a rational model of beliefs, whether it be Bayesian or logical. While these are valid topics for study, my goal is not to adopt a rational framework for modeling beliefs. Instead, I will be concentrating on a structural approach, which focuses on the underlying patterns that allow us to have beliefs in the first place.

Suppose I ask you "Have you seen the novel I was reading yesterday"? and you reply "In the living room", which is a typical "daily-life" exchange involving our beliefs about the nature of the world. How do I evaluate your statement? When you say it, I have no way of knowing it to be true. Of course, I could check the truth of the statement by going to the living room, but even in order to make that effort, I have to evaluate the plausibility of your statement. What principles do I use for that purpose? As one might imagine, answering these questions requires an analysis of the underlying structure of beliefs.

From the structural perspective, the first step is to delineate some of the regularities underlying beliefs. The next step is to develop a formal scheme in which these regularities can be stated and modeled. The third step is to extract some generative principles that explain those regularities. All ready, at this stage, we should see some interesting explanations of the original phenomena, which cannot be derived in a direct way from the phenomena themselves. Only after this is done do we go to the last step, which is to put numbers on the models, in order to extract detailed quantitative predictions. In this paper, I will only perform the first three steps, leaving the fourth step for the future.

A methodological note

While one can study beliefs using various kinds of data, I have chosen to concentrate on belief formation and the principles behind them based on what one would call typical, daily life interactions — the conversations we have, the stories we tell and the usual contexts (work, family etc.) in which we live.

So, what are the phenomena that one should focus on, while trying to model beliefs in the

structural mode? While I cannot claim to have a complete analysis, here are four that I find crucial

(1) Beliefs are asymmetric. There is a difference between articulating a belief and accepting it, as there is between uttering a sentence and understanding it. The way beliefs are articulated is fundamentally underdetermined. Suppose a friend told me that New Orleans is suffering from severe flooding. Even if I was an evangelical Christian, I could reply in several ways: that it was an act of god, and we should accept it for what it is, or say it was an act of god and we should stop leading sinful lives. Certainly, there are people who believe one or the other. It is hard, if not impossible to tell how people will string their beliefs together. Even having the same starting point does not ensure that we will end up at the same point.

(2) On the other hand, beliefs are fundamentally social. Beliefs are meant to be shared — which is to say that a person holding a belief has a reasonable expectation that when he communicates the belief to another person, that person can comprehend the belief and evaluate its content within the cultural norms of the community that they belong to. A conjecture in advanced mathematics is a belief only if the recipient of the conjecture can understand the statement. In other words, there are implicit constraints on what constitutes a belief in a given social context. So beliefs have to be coherent and acceptable.

(3) Beliefs are judged by norms of *acceptability* (Bach and Harnish, 1979) that are more general than the usual constraints like predictive power or explanatory adequacy often used in rational models of beliefs. For example, in a Christian community, the recent flooding in New Orleans can be acceptably explained as an act of god, which a scientist may view as being un-predictive. In other words, acceptability is relative to the social context in which the belief is being stated.[1] Note that the same Christian who explains the floods as an act of god might still prefer to go to a hospital when ill rather than to a faith healer.

(4) Beliefs are nested within several contexts, each with its own domain of application, though some contexts may trump others. For example, if someone asks me, ''why is this laptop in your living room'', I might say, ''because I like writing on the couch'', which in turn is predicated on my having bought a laptop from Dell, the existence of computer technology and so on. A given belief may well be articulated and expanded in several contexts simultaneously as one often does, say as a ''life story''.[2]

Any theory of beliefs has to take these four constraints in mind. Furthermore, one may need a domain in which all four aspects of beliefs and their interaction with each other can be studied in detail. Is there such a domain? I believe that stories are a good domain for studying the structure of beliefs, for two reasons: first, beliefs are usually communicated in narrative form and secondly, there has been a lot of work by people such as Campbell (1972) and others in trying to encapsulate the structure of stories. In the next few sections I will argue for using stories as models for beliefs, introduce the prototypical structure of a story informally, model its structure formally and outline a few formal operations underlying the narrative flow.

[1]The acceptability based norms are related to the earlier methodological remarks. Of course, a scientific hypothesis is evaluated by the scientific community in ways that are different from our daily beliefs, but for my purposes, the scientific community is a highly specialized community with a process of belief formation and evaluation that are markedly different. Of course, we could say that informal conversations and meetings play a much larger role in belief formation within the scientific community as well, they are just not part of the public record of science.

[2]For example, on being asked, ''why is this laptop in the living room?'' I may say ''Because I like writing on the couch. I find that it makes me think more informally than if I was writing on a table. In grade school and middle school I was always forced to write on a chair and desk and I can't stand that anymore.''

Beliefs and stories

Why is a belief like a story? Before we go any further, let me first say what I mean by the term ''belief''.

Definition 1. A belief is a *stable*, *communicable* attitude about the nature of the world that is *acceptable* to the community with whom the belief is shared.

Of course, not all members of a community have to share the same attitude. A famous example of a lack of shared beliefs is the old story of the blind men and the elephant, where each man, after touching one of the elephant's parts believes that the elephant is a snake or a tree or some other object, not realizing that they are basing their inference on partial exposure to the facts. That said, until the blind men talk to each other and arrive at a common opinion of the nature of the beast in front of them, they do not have a shared belief. One could say that the blind men are having a ''conversation'', each one with his point of view, but they are not telling a common story. Beliefs are like stories in that they are special ''conversations''.

A typical conversation involves alternating points of view. For example, suppose I am home and I ask my wife ''Have you seen the book I was reading yesterday''? and she replies ''which one''? The two of us do not share the same information (I know what book it was while she does not) and until we can deliberate and come to a position of shared knowledge, we can have a conversation but we cannot share our beliefs about the location of the book. As one can see, this situation is exactly the same as the blind men and the elephant. In a story on the other hand, there is a uniform point of view, namely, the narrator's perspective. Indeed, in the story of the blind men and the elephant, the reader is in on the joke because he, unlike the blind men, can see the whole elephant because the narrator has provided a consistent point of view. More formally, I could say the following:

Definition 2. A *conversation* is a sequence of communications, $\mathbf{C}_i$, where each communication is a pair (S_i, P_i) where S_i is the semantic content of the communication and P_i is the point of view of the communication.

In general, a conversation can have multiple points of view, one point of view for each speaker (or more if a speaker changes his point of view during the conversation). A story or a belief — in the shared sense of the Definition 1 above — is a conversation where everybody shares the *same* point of view.

Definition 3. A *story* or a *belief* is a conversation $\mathbf{C}_i$ such that $P_i = P_j$ for all i, j.

A belief or a story is a special conversation that admits an invariant point of view, or one could say that a story or a belief is a highly *non-accidental* conversation. A person listening to a conversation will quickly conclude when the conversation is articulating a common belief. Given that beliefs are like stories, one should expect that the structure of stories illuminates the structure of beliefs.

Stories

One of the oldest maxims of storytelling, one that can be found in every book on writing fiction or screenplays (McKee, 1997) is the following.

Conflict drives stories

In order to understand this maxim, let us start with a simple narrative. Everyday you walk to your office at 8:00 AM in the morning. You take the elevator to the fifth floor, walk to your office door, take your keys out, open the door, boot up your computer and you are ready to go. Now think of a slightly alternative situation.

Suppose you are walking to your office in the morning with the intention of doing some work. You take the elevator to the fifth floor, walk to your office door, fish in your pockets for your keys out, when you realize that you left them on your mantelpiece at home.

The first one of the two narratives is reminiscent of ''boy meets girl, they fall in love, get married

and live happily ever after'', while the second one is more like ''boy meets girl, they fall in love but their families forbid the match''. It is also clear that the first narrative is complete while the second one is not in that the second one invites a further articulation of events until the door is opened.

What does this mean for a theory of stories? Here are three observations

(a) Both narratives start with a non-accidental event, i.e., Adam meets Eve and not some other possible mate.
(b) Both narratives take two agents (me and the door, Adam and Eve) and incorporate them into a new frame or context.
(c) However, the first narrative stops there, i.e., no other events occur in the new frame that pertains to the original non-accidental event. The second narrative on the other hand involves a new force (family, lack of keys) that ''impacts'' the stability of the original non-accidental configuration.

Given these intuitions, one may hazard the following hypothesis

Hypothesis. A story is a sequence of stable events such that

(i) Each event is a non-accidental mapping of two contexts.
(ii) Each event ''forces'' the previous one.

Of course, stories are not just about stringing non-accidental events together. As we all know, there are millions of stories, since each culture has its own myths, fables, heroic sagas and what not (Leeming, 2001). However, some scholars have pointed out how stories the world over share similar themes and forms, even going to the extent, as Booker (2005) has in his recent book, to say that there are only seven basic plots. For our purpose, we need not even get to the magic number seven, for even the most general fact about stories is enough to stimulate the theorist of beliefs. As we saw in the case of opening a door (or failing to do so), a story at its most basic has the following structure:

$$\text{Story begins} \rightarrow \text{Life going on as usual} \rightarrow \text{Conflict intervenes} \rightarrow \text{Conflict resolved} \rightarrow \text{End}$$

For example, considers the typical ''hero's journey'' which can be caricatured as ''The hero is sitting comfortably in his village smoking a cigar, when a demon comes threatening to kill everybody. The hero takes up arms, kills the demon and comes home to a heroic welcome. End''. Or you have the ''romantic tragedy'' where ''Boy meets girl, they fall in love, their families object, they commit suicide''. From a structural perspective we can model the story logic as follows

$$\begin{gathered}\text{Story begins} \rightarrow \text{Protagonist 1 in Context 1,} \\ \text{Protagonist 2 in Context 2} \rightarrow \text{Conflict} \\ \text{between the two protagonists forces Context 3} \\ \rightarrow \text{Conflict resolved in Context 3, story} \\ \text{descends back to Contexts 1 and 2} \\ \text{(possibly to a different state)} \rightarrow \text{End}\end{gathered}$$

In other words, stories are about moving in and out of nested contexts. Furthermore, each new context has a force that destabilizes the event in the previous context. The theorist needs to model these nested contexts in a manner that takes into account the acceptability and the social character of beliefs? I believe that certain tools from cognitive linguistics (in particular, Talmy's (2000) analysis of force dynamics) can be a useful part of a formal approach to modeling the dynamics of narrative. Here are the four components of the formal model in my system

(a) A Superstructure of causality, i.e., a catalog of the types of forces that are possible. An old but still useful classification, usually attributed to Aristotle, classifies causes into four types: Material (what the object is made of), Formal (what it is, say a cup), Efficient (who made it) and Final (why it was made).
(b) A representation of how causal forces interact in the belief world, which would be something like Talmy's force dynamics, where agents are classified as having a

positive or negative force valuation with respect to each other.[3]

(c) A theory of how new contexts are formed from old ones.

(d) A theory of acceptability of beliefs that shows how beliefs are evaluated by shared societal norms.

Now, let us see how to introduce these formal elements into the structural description of stories. Given these four elements of a narrative, what is the structure of a story? For example, Romeo and Juliet goes as follows: boy meets girl, they fall in love, families intervene, tragedy ensues and then the families come together as a result. Structurally, one can identify several elements of this story schema

(a) Story starts with the principal agents separated — they do not exist within a single frame or context.

(b) The agents come together. Their coming together indexes a new frame which subsumes the earlier two. Furthermore, each agent assigns a positive or negative value to the other (e.g., friend or foe). A new frame/context is formed. Now, one can think of two sub-possibilities:

  (b1) The valuations of the two agents are respected, the story comes to a natural conclusion. For example, if the two characters come together, fall in love and get married. End of story. Or, bad man enters town, sheriff gets into a gunfight and shoots him down. End of story.

  (b2) The new context enables forces that change the natural valuation. For example, in Romeo and Juliet, the intervention of the two families causes the tragedy. Here the outcome is different from what one would expect as the default.[4]

(c) In case of (b2), the story keeps moving until no new forces are introduced and the situation is resolved. What is interesting in both the $a \to b$ and $b \to c$ transitions is that an unexpected event moves the story along. This will be crucial for us later — that the dynamics of stories involve non-accidental events.

To summarize, we can see three structural principles at play — one principle that indexes into a new frame/context when an unexpected event happens, another principle that forms the new context out of two (or more) old ones and the third that indexes an unexpected force in the new context. Let us see how to model these three aspects formally.

Partial order of contexts

The simplest model of a story, as we have seen, is a linear sequence that moves forward to a higher context by mapping two contexts (that are related non-accidentally) together. We can model the linear story as follows:

Definition 4. A partial ordering is a relation, Γ, that is reflexive, anti-symmetric and transitive.

Definition 5. Let Γ and Δ be two partial orders. Then the mapped partial order $\Gamma \cdot \Delta$ is the smallest partial order that dominates the two.

Note that if Γ and Δ are two partial orders, there are natural inclusion maps $\Gamma \to \Gamma \cdot \Delta$ and $\Delta \to \Gamma \cdot \Delta$ as well as natural projection maps $\Gamma \cdot \Delta \to \Gamma$ and $\Gamma \cdot \Delta \to \Delta$.

Definition 6. Let Γ_1 and Γ_2 be two partial orderings of closed directed line segments.

A Story is a sequence of inclusions and projections $\Gamma_1, \Gamma_2 \to \Gamma_1 \cdot \Gamma_2 \to \Gamma_1, \Gamma_2$.

Note that there is no unique partial order that contains the two partial orders that we start with. The minimal partial order is uniquely determined only if the context is fully defined, which it rarely is. In general there are multiple minimal mappings, which is not a problem really, since as we saw earlier, beliefs are asymmetric — we want the

[3] We can see how society by changing the valuation of one agent for another can strongly influence the structure of beliefs.

[4] As mentioned earlier, the asymmetry of beliefs/stories means that they are fundamentally underdetermined. Therefore, the new context is not *fully* determined by the contexts that feed into it, which leaves open the possibility of unexpected forces.

formal framework to be underdetermined since the underlying beliefs are as well.

The partial orderings show how the contexts are arranged relative to each other, but they do not tell us how the story moves within a context. For this we need a theory of force dynamics. To start with, one can get away with a relatively simple model of agent–agent interactions.

Suppose we have two agents, x and y, each with its own positive or negative valuation of the other. Let us call these valuations the source valuation. After their interaction (which may invoke other agents, like in Romeo and Juliet) the new context ends up with a target valuation of the two agents with respect to each other. This target valuation has four possible qualitative outcomes, assuming that the source valuation has two choices (positive, negative) to start with, some of which are default and the others are non-accidental. For example, if the two agents have a negative valuation of each other, but the interaction leads to them having a positive valuation of the other, it is non-accidental. Lets us see how this plays out in full, formally.

Definition 7. An agent–agent interaction is a map F: $R^2 \to (+,-)$.

Here is the matrix of expected/default outcomes.

$A(+) \leftrightarrow B(+) \to AB$ (they come together);

$A(-) \leftrightarrow B(-) \to A$ if $\text{val}(A) > \text{val}(B); \to B$ if $\text{val}(B) > \text{val}(A)$ (one beats the other);

$A(+) \leftrightarrow B(-)$ or $A(-) \leftrightarrow B(+)$, indeterminate. (There is no natural solution.)

What does this mean though? It means that the story ends in that context if one of the following happens

(a) If the two agents value each other positively, they come together to form one unit.
(b) If the two agents value each other negatively, they fight it out and the stronger one of the two wins.
(c) The agents are of mixed opinions and they quit, i.e., they decide not to play the game.

Given this default valuation rule, we can also state the non-accidental interaction in which context shift happens as follows.

Principle of context change

A story moves from one context to another when the target valuation does not have the expected sign.

The story moves on to the next frame if one of the following happens

(a) The two agents value each other positively but they do not form a unit.
(b) The two agents value each other negatively and the weaker one wins.
(c) The two agents have mixed opinions but they get to have the same valuation, positive or negative.

A story (or a belief), as it develops, invokes new characters (agents) and new forces, but it can also loop back and change the mutual valuation of old characters (agents) and contexts. Enemies can become friends and vice versa. The above scheme is capable of generating infinitely long stories that in principle can loop back.[5] But, the fact that the story or belief can go on forever does not mean it will. Its unfolding depends on the acceptability of new context shifts each time that happens. In the next section, I try to address this issue.

Acceptability and such

Now, what we want to encode in our framework is that people, for the most part, do not find beliefs that are too far away or inconsistent with their own beliefs acceptable. How to model this? Similarity metrics that measure how close a belief is to another are hard to come by. Furthermore, how to compare beliefs that are not that similar to each other? For example, if I believe in a vengeful God, I might believe that Hurricane Katrina was a

[5] A fascinating question is whether stories are finite structures that lead to an infinite inferential leap. For example, the series of negative and positive valuations in a given story form a tree structure. As the story progresses, and loops come into play, the shape of the tree approximates a fractal structure. A natural question is whether the limiting fractal tree structure is inferred by the listener? In other words, is a listener, on hearing a story, capable of inferring an infinite recursive structure?

result of the sins of the people of New Orleans. How is the second belief like the first? Indeed the second belief is more like an inference from the first than anything else. So what we need is a theory of inference built around notions of acceptability. For that task, partial orders are more than adequate.

From an intuitive point of view, partial orders are easy to understand. Just imagine a family tree of grandchildren, children, parents, grandparents etc. Consider the relation "DESCENDENT OF". Then each person in the family tree is a part of a sequence of members A, B, C, D such that A is a DESCENDENT OF B, B is a DESCENDENT OF C and so on. Note that if A is a DESCENDENT OF B and B is a DESCENDENT OF C then A is a DESCENDENT of C. In this way, an entire family can be organized according to the ordering imposed by the relationship "DESCENDENT OF". Note that in a family, not every body is related to one another by the DESCENDENT OF relation, for example, siblings are not descendents of each other. From a mathematical point of view, partial orderings are nothing more than a generalization of a family tree. To be precise

Definition 8. Let S be a set and a, b, c etc. be its elements. Then a partial ordering, $\leqslant$, is a relation on S, that follows the following rules:

(a) Reflexivity, i.e., for all elements a belonging to S, $a \leqslant a$.
(b) Antisymmetry, i.e., for all elements a and b belonging to S, if $a \leqslant b$ and $b \leqslant a$ then $a = b$.
(c) Transitivity, i.e., if $a \leqslant b$ and $b \leqslant c$ then $a \leqslant c$.

Here is how one can use partial ordering to define a metric on a collection of objects:

Definition 9. Let Γ be a partial ordering and let $d(a_{max}, a_{min})$ be the number of steps from a maximal element, a_{max}, to a minimal element, a_{min}. The depth of the partial ordering is Maximum $(d(a_{max}, a_{min})$, over all pairs $(a_{max}, a_{min}))$.

For example, in a family tree going from my great grandparents (maximal elements) to my children (minimal elements) the depth of the family tree is 4. Now, coming back to my measure of beliefs as encoded in stories. Each protagonist, Γ_i, in a story, S, is an element in a partial order and that partial order has known some depth (let us call that $\text{Depth}(\Gamma_i)$). We can define the relative depth of a story as follows

Definition 10. Let Γ_1 and Γ_2 be the two protagonists of the story with their own individual partial orders. Then the relative depth of the story is the $\text{Depth}(\Gamma_1 \cdot \Gamma_2) - \min(\text{Depth}(\Gamma_1), \text{Depth}(\Gamma_2))$.

In simple terms, the relative depth measures the number of additional contexts it takes to enfold both the protagonists of the story. This gives an easy measure of acceptability, which is that people would like to keep the number of new contexts down, i.e., if you have to invoke a large number of intermediate contexts to wrap up the story, then it is not believable. In more abstract terms, we can state the restriction on the number of new contexts in terms of *locality*, where by locality, I mean the following: a new belief should be a small perturbation of old beliefs. In the case of partial orders, the locality condition can be stated as follows:

Locality condition: Given two partial orders, the mapped partial order is acceptable only if the depth of the new partial order is a small perturbation of the original two.

So the locality condition gives us a principle by which we can study the dynamics of beliefs. In the extreme situation, we have the following condition:

The depth 1 condition: A story is acceptable only if the context jump is of depth 1.

Suppose we go back to the earlier example. I am working on my laptop in the living room. You come in and ask me "why is the computer here" to which I reply "I like working on the laptop in the living room". So far, there is nothing new. A default context has been set up, of answering the question in the proximate spatio-temporal context. But if you say "but why the PC"? there is a conflict, which cannot be resolved in the default context. What do I do? I can say: "Because my Mac is being repaired" or "I have switched to PC's" or "It is my brother's laptop, I am loading some music in it for him". There is no *correct* answer, but I cannot say: "Because computers

have been invented'' except as a wisecrack, since it is not a minimal answer. The question arises, whether the locality principle applies to other examples of belief change, more real world kind of examples, so to speak. I believe so, but the full analysis will have to wait for another day.

Summary

In this paper, I outlined a few steps towards formulating a structural theory of beliefs, organized around the idea that beliefs are like stories. The fundamental lesson for belief change is the following: new beliefs come from an acceptable mixture of old beliefs, which maps neatly on to the old idea that stories have to a conflict as a driving force. From a structural perspective, we could say belief change occurs when a new context is induced by two old contexts. Finally, one can combine the contextuality and acceptability conditions into one locality condition for partial orders. This locality condition predicts when new beliefs are acceptable.

References

Bach, K. and Harnish, R. (1979) Linguistic Communication and Speech Acts. MIT Press, Cambridge, MA.

Booker, C. (2005) The Seven Basic Plots: Why We Tell Stories. Continuum International Publishing Group, London.

Campbell, J. (1972) The Hero with a Thousand Faces. Bollingen, Princeton, NJ.

Fodor, J.A. (1975) The Language of Thought. Harvester Press, Sussex.

Lakoff, G. (1987) Women, Fire and Dangerous Things. Chicago University Press, Chicago, IL.

Leeming, D. (2001) Myth: A Biography of Belief. Oxford University Press, Oxford.

McKee, R. (1997) Story: Substance, Structure, Style and the Principles of Screenwriting. Regan Books, New York.

Talmy, L. (2000) Towards a Cognitive Semantics. MIT Press, Cambridge, MA.

Varela, F., Thompson, E. and Rosch, E. (1991) The Embodied Mind. MIT Press, Cambridge, MA.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 10

The emergence of mind and brain: an evolutionary, computational, and philosophical approach

Klaus Mainzer*

Chair for Philosophy of Science, Institute of Interdisciplinary Informatics, University of Augsburg, D-86135 Augsburg, Germany

Abstract: Modern philosophy of mind cannot be understood without recent developments in computer science, artificial intelligence (AI), robotics, neuroscience, biology, linguistics, and psychology. Classical philosophy of formal languages as well as symbolic AI assume that all kinds of knowledge must explicitly be represented by formal or programming languages. This assumption is limited by recent insights into the biology of evolution and developmental psychology of the human organism. Most of our knowledge is implicit and unconscious. It is not formally represented, but embodied knowledge, which is learnt by doing and understood by bodily interacting with changing environments. That is true not only for low-level skills, but even for high-level domains of categorization, language, and abstract thinking. The embodied mind is considered an emergent capacity of the brain as a self-organizing complex system. Actually, self-organization has been a successful strategy of evolution to handle the increasing complexity of the world. Genetic programs are not sufficient and cannot prepare the organism for all kinds of complex situations in the future. Self-organization and emergence are fundamental concepts in the theory of complex dynamical systems. They are also applied in organic computing as a recent research field of computer science. Therefore, cognitive science, AI, and robotics try to model the embodied mind in an artificial evolution. The paper analyzes these approaches in the interdisciplinary framework of complex dynamical systems and discusses their philosophical impact.

Keywords: brain; mind; complex systems; nonlinear dynamics; self-organization; computational systems; artificial minds

From linear to nonlinear dynamics

The brain is a complex cellular system of 10^{11} neurons and 10^{14} synaptic connections. In order to understand and to model the emergence of its mental functions, we must study the nonlinear dynamics of complex systems. In general, a dynamical system is a time-depending multi-component system of elements with local states determining a global state of the whole system. In a planetary system, for example, the state of a planet at a certain time is determined by its position and momentum. The states can also refer to moving molecules in a gas, the excitation of neurons in a neural network, nutrition of organisms in an ecological system, supply and demand of economic markets, the behavior of social groups in human societies, routers in the complex network

*Corresponding author. Tel.: +49-821-598-5568; Fax: +49-821-598-5584; E-mail: klaus.mainzer@philuni-augsburg.de

DOI: 10.1016/S0079-6123(07)68010-8

of the internet, or units of a complex electronic equipment in a car. The dynamics of a system, that is, the change of system's states depending on time, is represented by linear or nonlinear differential equations. In the case of nonlinearity, several feedback activities take place between the elements of the system. These many-body problems correspond to nonlinear and nonintegrable equations with instabilities and sometimes chaos (Mainzer, 2007).

From a philosophical point of view, mathematical linearity means a strong concept of causality with similar causes or inputs of a dynamical system leading to similar effects or outputs: small changes in the parameters or small perturbations added to the values of the variables produce small changes in subsequent values of the variables. Further on, composite effects of linear systems can be reduced to the sum of more simple effects. Therefore, scientists have used linear equations to simplify the way in which we think about the behavior of complex systems. The principle of superposition has its roots in the concept of linearity. But, in the case of nonlinearity, similar causes lead to exponentially separating and expanding effects: small changes in the parameters or small perturbations added to the values of the variables can produce enormous changes in subsequent values of the variables because of its sensitivity to initial conditions. In this case, the whole is more than the sum of its elements.

The mathematical theory of nonlinear dynamics distinguishes different types of time-depending equations, generating different types of behavior, such as fixed points, limit cycles, and chaos. In a top-down approach of model building, we start with an assumed mathematical model of a natural or technical system and deduce its behavior by solving the corresponding dynamical equations under certain initial conditions. The solutions can be represented geometrically as trajectories in the phase space of the dynamical system and classified by different types of attractors. But, in practice, we often adopt the opposite method of a bottom-up approach. Physicists, chemists, biologists, physicians, or engineers start with data mining in an unknown field of research. They only get a finite series of measured data corresponding to time-depending events of a dynamical system. From these data they must reconstruct the behavior of the system in order to guess its type of a dynamical equation. Therefore, the bottom-up approach is called time series analysis. In many cases, we have no knowledge of the system from which the data was acquired. Time series analysis then aims to construct a black box, which take the measured data as input and provides as output a mathematical model describing the data (Small, 2005; Floridi, 2004). In practice, the realistic strategy of research is a combination of the top-down approach with model building and the bottom-up approach with time series analysis of the measured data.

In classical measurement theory, measurement error is analyzed by statistical methods, such as correlation coefficient and autocorrelation function. But these standard procedures are not able to distinguish between data from linear and nonlinear models. In nonlinear data analysis, the measured data are used in a first step to reconstruct the dynamics of the system in a phase space. Nonlinear dynamical systems generating chaos must be determined by at least three equations. For example, a three-dimensional attractor is generated in a phase space with three coordinates $x(t)$, $y(t)$, and $z(t)$, which are determined by three time-depending nonlinear differential equations. But, in practice, it is often difficult to distinguish several variables of a system. Nevertheless, if only one variable can be measured, an attractor with a finite number of dimensions can be reconstructed from the measured time series with great similarity to the original attractor of the system. We must only assume that we can also measure the derivative of that variable, and further higher order derivatives up to some finite level d. Then, if the dimension of the system is less than d, we have enough information to completely describe the system with d differential or difference equations. Measuring d derivatives is equivalent to measuring the system at d different time intervals. Therefore, according to Takens' theorem, the measured time series of a variable can be embedded in a reconstructed phase space with d dimensions. The sequence of points created by embedding the measured time series is a reconstructed trajectory of the original

system, generating an attractor with great similarity to the original one of the system.

In practice, decisions about chaotic dynamics are rather difficult. How can we decide that a time series of measured data is not generated by noisy irregularity but by highly structured chaotic attractors? A chaotic attractor is determined by a trajectory in a bounded region of a phase space with aperiodic behavior and sensitive dependence on initial conditions. These criteria — determinism, boundedness, aperiodicity, and sensitivity — can be checked by several techniques of time series analysis. In the case of noise, the trajectories spread unbounded all over the phase space. A chaotic attractor is finite and always bounded in a certain region of the phase space. Aperiodicity means that the states of a dynamical system never return to their previous values. But values of states may return more or less to the vicinity of previous values. Thus, aperiodicity is a question of degree which can be studied in recurrence plots of measured points. Such plots depict how the reconstructed trajectory recurs or repeats itself. The correlation integral defines the density of points in a recurrence plot where the measured time series are closer than a certain degree of distance.

If a time series is generated by a chaotic system, the trajectory of the time series, which is reconstructed from the measurement data of embedding, has the same topological properties as the original attractor of the system, as long as the embedding dimension is large enough. Takens proved a method for finding an appropriate embedding dimension for the reconstruction of an attractor. But this method yields no procedure for finding a chaotic attractor, because its existence has been already assumed in order to determine its dimension from the measured data.

Another way to characterize chaotic dynamics is to measure the strength of their sensitive dependence on initial data. Consider two trajectories starting from nearly the same initial data. In chaotic dynamics only a tiny difference in the initial conditions can result in the two trajectories diverging with exponential speed in the phase space after a short period of time. In this case, it is difficult to calculate long-term forecasts, because the initial data can only be determined with a finite degree of precision. Tiny deviations in digits behind the decimal point of measurement data may lead to completely different forecasts. This is the reason why attempts to forecast weather fail in an unstable and chaotic situation. In principle, the wing of a butterfly may cause a global change of development. This ''butterfly effect'' can be measured by the so-called Lyapunov exponent Λ. A trajectory $x(t)$ starts with an initial state $x(0)$. If it develops exponentially fast, then it is approximately given by $|x(t)| \sim |x(0)|e^{\Lambda t}$. The exponent is smaller than zero if the trajectory is attracted by attractors, such as stable points or orbits. It is larger than zero if it is divergent and sensitive to very small perturbations of the initial data.

An attractor is typically a finite region in the phase space. Sensitivity to initial conditions means that any nearby points on the attractor in the phase space diverge from each other. They cannot, however, diverge forever, because the attractor is finite. Thus, trajectories from nearby initial points on the attractor diverge and are folded back onto the attractor, diverge and are folded back, etc. The structure of the attractor consists of many fine layers, like an exquisite pastry. The closer one looks, the more detail in the adjacent layers of the trajectories is revealed. Thus, the attractor is fractal. An attractor that is fractal is called strange. There are also chaotic systems that are only exponentially sensitive to initial conditions but not strange. Attractors can also be strange (fractal), but not chaotic. The fractal dimension of an attractor is related to the number of independent variables needed to generate the time series of the values of the variables. If d is the smallest integer greater than the fractal dimension of the attractor, then the time series can be generated by a set of d differential equations with d independent variables. For example, a strange attractor of fractal dimension 2.03 needs three nonlinear coupled differential equations to generate its trajectory.

In summary, dynamical systems can be classified by attractors with increasing complexity from fixed points, periodic and quasi-periodic up to chaotic behavior. This classification of attractors can be

characterized by different methods, such as typical patterns of time series, their power spectrum, phase portraits in a phase space, Lyapunov exponents, or fractal dimensions. A remarkable measure of complexity is the Kolmogorov–Sinai (KS) entropy, measuring the information flow in a dynamical system (Deco and Schürmann, 2000; Mainzer, 2007). A dynamical system can be considered an information-processing machine, computing a present or future state as output from an initial past state of input. Thus, the computational efforts to determine the states of a system characterize the computational complexity of a dynamical system. The transition from regular to chaotic systems corresponds to increasing computational problems, according to the computational degrees in the theory of computational complexity. In statistical mechanics, the information flow of a dynamical system describes the intrinsic evolution of statistical correlations between its past and future states. The KS-entropy is an extremely useful concept in studying the loss of predictable information in dynamical systems, according to the complexity degrees of their attractors. Actually, the KS-entropy yields a measure of the prediction uncertainty of a future state provided the whole past is known (with finite precision).

In the case of fixed points and limit cycles, oscillating or quasi-oscillating behavior, there is no uncertainty or loss of information, and the prediction of a future state can be computed from the past. Consequently, the KS-entropy is zero. In chaotic systems with sensitive dependence on the initial states, there is a finite loss of information for predictions of the future, according to the decay of correlations between the past states and the future state of prediction. The finite degree of uncertainty of a predicted state increases linearly to its number of steps in the future, given the entire past. In the case of chaos, the KS-entropy has a finite value (larger than zero). But in the case of noise, the KS-entropy becomes infinite, which means a complete loss of predicting information corresponding to the decay of all correlations (i.e., statistical independence) between the past and the noisy state of the future. The degree of uncertainty becomes infinite.

Self-organization and emergence in evolution

How can the knowledge of chaos be applied in order to control risky and unstable situations in complex systems? This question will be a challenge for modeling the brain with millions of interacting cells in nonlinear dynamics. It seems to be paradoxical that chaotic systems which are extremely sensitive to the tiniest fluctuations can be controlled. But nowadays the control of chaos has been realized in chemical, fluid, and biological systems. In technology, for example, the intrinsic instability of chaotic celestial orbits is routinely used to advantage by international space agencies who divert spacecraft to travel vast distances using only modest fuel expenditures. All techniques of chaos control make use of the fact that chaotic systems can be controlled if disturbances are countered by small and intelligently applied impulses. Just as an acrobat balances about an unstable position on a tightrope by the application of small correcting movements, a chaotic system can be stabilized about any of an infinite number of unstable states by continuous application of small corrections.

Two characteristics of chaos make the application of control techniques possible. First, chaotic systems alternatively visit small neighborhoods of an infinite number of periodic orbits. The presence of an infinite number of periodic orbits embedded within a chaotic trajectory implies the existence of an enormous variety of different behaviors within a single system. Thus, the control of chaos opens up the potential for a great flexibility in operating performance within a single system.

A second characteristic of chaos that is important for control applications is its exponential sensitivity. It follows that the state of chaotic system can be drastically altered by the application of small perturbations. Therefore, uncontrolled chaotic systems fluctuate wildly. But, on the other side, controlled chaotic systems can be directed from one state to a very different one using only very small controls. Obviously, controlling strategies require that the system state lie close to the desired state. In such a case, the system dynamics can be linearized, making control calculations rapid and effective. In chaotic systems, ergodicity

ensures that the system state will eventually wander arbitrarily close to the desired state. But in higher dimensional or slowly varying systems, the time taken for the state to move on its own from one state to another can be prohibitive. In this case, fully nonlinear control strategies have been devised that use chaotic sensitivity to steer the system state from any given initial point to a desired state. Since chaotic systems amplify control impulses exponentially, the time needed to steer such a system can be quite short. These strategies have been demonstrated both in systems in which a large effect is desired using very modest parameter expenditures (e.g., energy and fuel) and in systems in which rapid switching between states is needed (e.g., computational and communication applications).

Nonlinear dynamics does not only yield chaos and noise, but also order. The emergence of order and structures in evolution can be explained by the dynamics of attractors in complex systems (Mainzer, 2007). They result from collective patterns of interacting elements in the sense of many-body problems that cannot be reduced to the features of single elements in a complex system. Nonlinear interactions in multi-component (''complex'') systems often have synergetic effects, which can neither be traced back to single causes nor be forecasted in the long run or controlled in all its details. Again, the whole is more than the sum of its parts. This popular slogan for emergence is precisely correct in the context of nonlinearity.

The mathematical formalism of complex dynamical systems is taken from statistical mechanics. If the external conditions of a system are changed by varying certain control parameters (e.g., temperature), the system may undergo a change in its macroscopic global states at some critical point. For instance, water as a complex system of molecules changes spontaneously from a liquid to a frozen state at a critical temperature of zero Celsius. In physics, those transformations of collective states are called phase transitions. Obviously they describe a change of self-organized behavior between the interacting elements of a complex system.

According to Landau, the suitable macrovariables characterizing the change of global order are denoted as ''order parameters''. For example, the emergence of magnetization in a ferromagnet is a self-organized behavior of atomic dipoles that is modeled by a phase transition of an order parameter, the average distribution of microstates of the dipoles, when the system is annealed to the Curie point. The concept of order parameters can be generalized for phase transitions, when the system is driven away from equilibrium by increasing energy (Haken and Mikhailov, 1993). If, for example, the fluid of a stream is driven further and further away from thermal equilibrium, by increasing fluid velocity (control parameter), then fluid patterns of increasing complexity emerge from vortices of fixed points, periodic oscillations to chaotic turbulence. Roughly speaking, we may say that old structures become unstable, broken down by changing control parameters, and new patterns and attractors emerge.

More mathematically, nonlinear differential equations are employed to model the dynamics of the system. At first, we study the behavior of the elements on the microlevel in the vicinity of a critical point of instability. In a linear-stability analysis, one can distinguish stable and unstable modes which increase to the macroscopic scale, dominating the macrodynamics of the whole system. Thus, some few unstable modes become the order parameters of the whole system. From a methodological point of view, the introduction of order parameters for modeling self-organization and the emergence of new structures is a giant reduction of complexity. The study of, perhaps, billions of equations, characterizing the behavior of the elements on the microlevel, is replaced by some few equations of order parameters, characterizing the macrodynamics of the whole system. Complex dynamical systems and their phase transitions deliver a successful formalism to model self-organization and emergence. Further on, the knowledge of characteristic order parameters and critical values of control parameters open a chance to influence the whole dynamics and to create desired states of technical systems by self-organization. The formalism does not depend on special, for example, physical laws, but must be appropriately interpreted for biological and technical applications.

According to the general scheme of nonlinear dynamics, biological organisms function on many levels that have emerged step-by-step during evolution. It is a question of granulation as to how "deep" we like to lay the initial layer of microdynamics. As far as we know at least atomic dynamics influence states of living organisms. During prebiotic evolution, interacting atoms and molecules created complex biomolecules (e.g., proteins) by catalytic and autocatalytic processes which are the building blocks of cells. Interacting cells achieved complex cellular systems like organs or organisms which are elements of populations. Further on, interacting populations became elements of ecological networks as examples of complex systems. Thus, from the nonlinear dynamics at each level, there emerge new entities that are characterized by order parameters. Examples of order parameters are characteristic macroscopic features of phenotypes which are determined by the genotype of an organism on the microlevel. The macrodynamics of these order parameters determine the microdynamics of the new entities, providing the basis of macrodynamics on the following level. In principle, the dynamics of each level can be modeled by appropriate nonlinear differential equations. In this case, the succeeding hierarchical level can be mathematically derived from the previous one by a linear-stability analysis.

How can order be regulated and controlled in living organisms? This is a key question with respect to modeling in organic computing. Important examples are bio-oscillators which can be considered to be the order parameters of life. Nature abounds with rhythmic behavior that closely intertwines the physical and biological sciences. The diurnal variations in dark and light give rise to circardian physiological rhythms. But the rhythmic nature of biological process is not only controlled by external processes. It often arises from the intrinsic dynamics of complex nonlinear networks. Since all biological systems are thermodynamically open to the environment they are dissipative, that is, they give up energy to their surroundings in the form of heat. Thus, for the oscillator to remain periodic, energy must be supplied to the system in such a way as to balance the continual loss of energy due to dissipation. If a balance is maintained, then the phase space orbits become a stable limit cycle, that is, all orbits in the neighborhood of this orbit merge with it asymptotically. Such a system is called a bio-oscillator, which left to itself begins to oscillate without apparent external excitation. The self-generating or self-regulating features of bio-oscillators depend on the intrinsic nonlinearity of the biological system.

How can perturbations of such systems be used to explore and control their physiological properties? This question does not only inspire new therapeutic methods in medicine, but also technical applications in organic computing. An example of a complex cellular system is the heart which can be considered a bio-oscillator (Bassingthwaighte et al., 1994). In the simple case of an embryonic chick heart, a cardiac oscillator can be described by a system of ordinary differential equations with a single unstable steady-state and displaying an asymptotically stable limit cycle oscillation that is globally attracting. After short perturbations, the pulses return quickly to the limit cycle. The dynamics can be studied in the corresponding time series of ECG-curves. The dynamics of a mammalian heart is much more complex. The question arises if observed fluctuations are the result of the oscillations being unpredictably perturbed by the cardiac environment, or are a consequence of cardiac dynamics being given by a chaotic attractor, or both. In healthy patients, the heart rate is modulated by a complex combination of respiratory, sympathetic, and parasympathetic regulators. For unhealthy patients, the ideas of chaos control can be incorporated into therapeutic situations. Control is attempted by stimulating the heart at appropriate times. Repeated intervention prevents the rhythm from returning to the chaotic mode.

Obviously, the total and global chaos of a system is dangerous. But local chaotic fluctuations are physiologically advantageous. Sustained periodicities are often unhealthy. To maintain health, the variables of a physiological system must be able to extend over a wide range to provide flexible adaption. Healthy people have greater variability in their heart rates than those with heart disease.

Thus, local chaotic fluctuations may provide the plasticity to cope with the exigencies of an unpredictable and changing environment.

Chaotic systems can be controlled more finely and more quickly than linear systems. In linear systems, the response of the output depends linearly on the input. Small changes in a parameter of a linear system produce only small changes in the output. The variable controlling a chaotic physiological response may need to change by only a small amount to induce the desired large change in the physiological state. Moreover, a chaotic physiological system can switch very rapidly from one physiological state to another. Natural self-control and self-organization of complex physiological systems open a wide range of medical and engineering applications.

The traditional notion of health is one of homeostasis and is based on the idea that there exists an ideal state in which the body is operating in a maximally efficient way. In this opinion, illness is considered to be the deviation of the body from this state, and it is the business of the physician to assist the patient in regaining this state again. The nonlinear dynamics of biological systems suggest the replacement of homeostasis by homeodynamics allowing a more flexible view of how the systems work and making room for the concept of systems with complex responses, even to the point of inherent instability. The mammalian organism is composed of multiple nested loops of nonlinear interacting systems on the physiological level. How much greater are the possibilities for complex behavior at the psychic levels of the brain.

Self-organization and emergence of brain and mind

The coordination of the complex cellular and organic interactions in an organism needs a new kind of self-organizing control. That was made possible by the evolution of nervous systems that also enabled organisms to adapt to changing living conditions and to learn from experiences with their respective environments. The hierarchy of anatomical organizations varies over different scales of magnitude, from molecular dimensions to that of the entire central nervous system (CNS). The research perspectives on these hierarchical levels may concern questions, for example, of how signals are integrated in dendrites, how neurons interact in a network, how networks interact in a system like vision, how systems interact in the CNS, or how the CNS interacts with its environment. Each stratum may be characterized by some order parameters determining its particular structure, which is caused by complex interactions of subelements with respect to the particular level of hierarchy.

On the microlevel of the brain, there are massively many-body problems which need a reductionist strategy to get a handle with their complexity. In the case of EEG-pictures, a complex system of electrodes measures local states (electric potentials) of the brain. The whole state of a patient's brain on the microlevel is represented by local time series. In the case of, for example, petit mal epilepsy, they are characterized by typical cyclic peaks. The microscopic states determine the macroscopic electric field patterns during a cyclic period. Mathematically, the macroscopic patterns can be determined by spatial modes and order parameters, that is, the amplitude of the field waves. In the corresponding phase space, they determine a chaotic attractor characterizing petit mal epilepsy.

The neural self-organization on the cellular and subcellular level is determined by the information processing in and between neurons. Chemical transmitters can effect neural information processing with direct and indirect mechanisms of great plasticity. Long-term potentiation (LTP) of synaptic interaction is an extremely interesting topic of recent brain research. LTP seems to play an essential role for the neural self-organization of cognitive features such as, memory and learning. The information is assumed to be stored in the synaptic connections of neural cell assemblies with typical macroscopic patterns.

But while an individual neuron does not see or reason or remember, brains are able to do so. Vision, reasoning, and remembrance are understood as higher level functions. Scientists who prefer a bottom-up strategy recommend that higher level functions of the brain can be neither

addressed nor understood until particular property of each neuron and synapse is explored and explained. An important insight of the complex system approach discloses that emergent effects of the whole system are synergetic system effects which cannot be reduced to the single elements. They are results of nonlinear interactions. Therefore, the whole is more than the (linear) sum of its parts. Thus, from a methodological point of view, a purely bottom-up strategy of exploring the brain functions must fail. On the other hand, the advocates of a purely top-down strategy proclaiming that cognition is completely independent of the nervous system are caught in the old Cartesian dilemma "How does the ghost drive the machine?"

Today, we can distinguish several degrees of complexity in the CNS. The scales consider molecules, membranes, synapses, neurons, nuclei, circuits, networks, layers, maps, sensory systems, and the entire nervous system. The research perspectives on these hierarchical levels may concern questions, for example, of how signals are integrated in dendrites, how neurons interact in a network, how networks interact in a system like vision, how systems interact in the CNS, or how the CNS interacts with its environment. Each stratum may be characterized by some order parameters determining its particular structures, which is caused by complex interactions of subelements with respect to the particular level of hierarchy. Beginning at the bottom, we may distinguish the orders of ion movement, channel configurations, action potentials, potential waves, locomotion, perception, behavior, feeling, and reasoning.

The different abilities of the brain need massively parallel information processing in a complex hierarchy of neural structures and areas. We know more or less complex models of the information processing in the visual and motoric systems. Even the dynamics of the emotional system is interacting in a nonlinear feedback manner with several structures of the human brain. These complex systems produce neural maps of cell assemblies. The self-organization of somatosensoric maps is well known in the visual and motoric cortex. They can be enlarged and changed by learning procedures such as the training of an ape's hand.

Positron emission tomography (PET) pictures show macroscopic patterns of neurochemical metabolic cell assemblies in different regions of the brain which are correlated with cognitive abilities and conscious states such as looking, hearing, speaking, or thinking. Pattern formation of neural cell assemblies are even correlated with complex processes of psychic states. Perturbations of metabolic cellular interactions (e.g., cocaine) can lead to nonlinear effects initiating complex changes of behavior (e.g., addiction by drugs). These correlations of neural cell assemblies and order parameters (attractors) of cognitive and conscious states demonstrate the connection of neurobiology and cognitive psychology in recent research, depending on the standards of measuring instruments and procedures.

Many questions are still open. Thus, we can only observe that someone is thinking and feeling, but not, what he is thinking and feeling. Further on, we observe no unique substance called consciousness, but complex macrostates of the brain with different degrees of sensoric, motoric, or other kinds of attention. Consciousness means that we are not only looking, listening, speaking, hearing, feeling, thinking, etc., but we know and perceive ourselves during these cognitive processes. Our self is considered an order parameter of a state, emerging from a recursive process of multiple self-reflections, self-monitoring, and supervising of our conscious actions. Self-reflection is made possible by the so-called mirror neurons (e.g., in the Broca area) which let primates (especially humans) imitate and simulate interesting processes of their companions. Therefore, they can learn to take the perspectives of themselves and their companions in order to understand their intentions and to feel with them. The emergence of subjectivity is neuropsychologically well understood.

The brain does not only observe, map, and monitor the external world, but also internal states of the organism, especially its emotional states. Feeling means self-awareness of one's emotional states which is mainly caused by the limic system. In neuromedicine, the "Theory of Mind" (ToM) even analyzes the neural correlates of social feeling which are situated in special areas of the

neocortex. For example, people suffering from Alzheimer disease, loose their feeling of empathy and social responsibility because the correlated neural areas are destroyed. Therefore, our moral reasoning and deciding have a clear basis in brain dynamics.

From a neuropsychological point of view, the old philosophical problem of ''qualia'' is also solvable. Qualia mean properties which are consciously experienced by a person. In a thought experiment a neurobiologist is assumed to be caught in a black-white room. Theoretically, she knows everything about neural information processing of colors. But she never had a chance to experience colors. Therefore, exact knowledge says nothing about the quality of conscious experience. Qualia in that sense emerge by bodily interaction of self-conscious organisms with their environment which can be explained by the nonlinear dynamics of complex systems. Therefore, we can explain the dynamics of subjective feelings and experiences, but, of course, the actual feeling is an individual experience. In medicine, the dynamics of a certain pain can often be completely explained by a physician, although the actual feeling of pain is an individual experience of the patient.

In order to model the brain and its complex abilities, it is quite adequate to distinguish the following categories. In neuronal-level models, studies are concentrated on the dynamic and adaptive properties of each nerve cell or neuron, in order to describe the neuron as a unit. In network-level models, identical neurons are interconnected to exhibit emergent system functions. In nervous system-level models, several networks are combined to demonstrate more complex functions of sensory perception, motor functions, stability control, etc. In mental-operation-level models, the basic processes of cognition, thinking, problem solving, etc. are described.

In the complex systems approach, the microscopic level of interacting neurons should be modeled by coupled differential equations modeling the transmission of nerve impulses by each neuron. The Hodgkin–Huxley equation is an example of a nonlinear diffusion reaction equation with an exact solution of a traveling wave, giving a precise prediction of the speed and shape of the nerve impulse of electric voltage. In general, nerve impulses emerge as new dynamical entities like ring waves in BZ-reactions or fluid patterns in nonequilibrium dynamics. In short they are the ''atoms'' of the complex neural dynamics. On the macroscopic level, they generate a cell assembly whose macrodynamics is dominated by order parameters. For example, a synchronously firing cell assembly represents some visual perception of a plant which is not only the sum of its perceived pixels, but characterized by some typical macroscopic features like form, background, or foreground. On the next level, cell assemblies of several perceptions interact in a complex scenario. In this case, each cell assembly is a firing unit, generating a cell assembly of cell assemblies whose macrodynamics is characterized by some order parameters. The order parameters may represent similar properties of the perceived objects.

In this way, we get a hierarchy of emerging levels of cognition, starting with the microdynamics of firing neurons. The dynamics of each level is assumed to be characterized by differential equations with order parameters. For example, on the first level of macrodynamics, order parameters characterize a visual perception. On the following level, the observer becomes conscious of the perception. Then the cell assembly of perception is connected with the neural area that is responsible for states of consciousness. In a next step, a conscious perception can be the goal of planning activities. In this case, cell assemblies of cell assemblies are connected with neural areas in the planning cortex, and so on. They are represented by coupled nonlinear equations with firing rates of corresponding cell assemblies. Even high-level concepts like self-consciousness can be explained by self-reflections of self-reflections, connected with a personal memory which is represented in corresponding cell assemblies of the brain. Brain states emerge, persist for a small fraction of time, then disappear, and are replaced by other states. It is the flexibility and creativeness of this process that makes a brain so successful in animals for their adaption to rapidly changing and unpredictable environments.

Self-organization and emergence of computational systems

Computational systems were historically constructed on the background of Turing's theory of computability (Dennett, 1998). In his functionalism, the hardware of a computer is related to the wetware of human brain. The mind is understood as the software of a computer. Turing argued: If human mind is computable, it can be represented by a Turing program (Church's thesis) which can be computed by a universal Turing machine, that is, technically by a general purpose computer. Even if people do not believe in Turing's strong AI-thesis, they often claim classical computational cognitivism in the following sense: computational processes operate on symbolic representations referring to situations in the outside world. These formal representations should obey Tarski's correspondence theory of truth: imagine a real world situation X1 (e.g., some boxes on a table) which is encoded by a symbolic representation A1 = encode(X1) (e.g., a description of the boxes on the table). If the symbolic representation A1 is decoded, then we get the real world situation X1 as its meaning, that is, decode(A1) = X1. A real-world operation T (e.g., a manipulation of the boxes on the table by hand) should produce the same real-world result A2, whether performed in the real world or on the symbolic representation: decode(encode(T)(encode(X1))) = T(X1) = X2. Thus, there is an isomorphism between the outside situation and its formal representation. As the symbolic operations are completely determined by algorithms, the real-world processes are assumed to be completely controlled. Therefore, classical robotics operates with completely determined control mechanisms.

Symbolic representations with ontologies, categories, frames, and scripts of expert systems work along this line. But, they are restricted to a specialized knowledge base without the background knowledge of a human expert. Human experts do not rely on explicit (declarative) rule-based representations only, but also on intuition and implicit (procedural) knowledge (Dreyfus, 1982; Searle, 1983). Further on, as already Wittgenstein knew, our understanding depends on situations. The situatedness of representations is a severe problem of informatics. A robot, for example, needs a complete symbolic representation of a situation which must be updated if the robot's position is changed. Imagine that it surrounds a table with a ball and a cup on it. A formal representation in a computer language may be ON(TABLE,BALL), ON(TABLE,CUP), BEHIND(CUP,BALL), etc. Depending on the robot's position relative to the arrangement, the cup is sometimes behind the ball or not. So, the formal representation BEHIND(CUP,BALL) must always be updated in changing positions. How can the robot prevent incomplete knowledge? How can it distinguish between reality and its relative perspective? Situated agents like human beings need no symbolic representations and updating. They look, talk, and interact bodily, for example, by pointing to things. Even rational acting in sudden situations does not depend on symbolic representations and logical inferences, but on bodily interactions with a situation (e.g., looking, feeling, reacting).

Thus, we distinguish formal and embodied acting in games with more or less similarity to real life: chess, for example, is a formal game with complete representations, precisely defined states, board positions, and formal operations. Soccer is a nonformal game with skills depending on bodily interactions, without complete representations of situations and operations which are never exactly identical. According to the French philosopher Merleau-Ponty (1962), intentional human skills do not need any symbolic representation, but they are trained, learnt, and embodied by the organism. An athlete like a pole-vaulter cannot repeat her successful jump like a machine generating the same product. The embodied mind is no mystery. Modern biology, neural, and cognitive science give many insights into its origin during the evolution of life.

Organic computing applies the principles of evolution and life to technical systems. The dominating principles in the complex world of evolution are self-organization and self-control. How can they be realized in technical systems? A nice test bed for all kinds of technical systems

are computational automata. There is a precise relation between self-organization of nonlinear systems with continuous dynamics and discrete cellular automata (CA). The dynamics of nonlinear systems is given by differential equations with continuous variables and a continuous parameter of time. Sometimes, difference equations with discrete time points are sufficient. If even the continuous variables are replaced by discrete (e.g., binary) variables, we get functional schemes of automata with functional arguments as inputs and functional values as outputs. There are classes of CA modeling attractor behavior of nonlinear complex systems which is well known from self-organizing processes.

But in many cases, there is no finite program, in order to forecast the development of random patterns. In general, there are three reasons for computational limits of system dynamics. (1) A system may be undecidable in a strict logical sense. (2) Further on, a system can be deterministic, but nonlinear and chaotic. In this case, the system depends sensitively on tiny changes of initial data in the sense of the butterfly effect. Long-term forecasting is restricted, and the computational costs of forecasting increase exponentially after some few steps of future predictions. (3) Finally, a system can be stochastic and nonlinear. In this case, only probabilistic predictions are possible. Thus, pattern emergence of CA cannot be controlled in any case.

Cellular automata are only a theoretical concept of computational dynamics. In electrical engineering, information and computer science, the concept of cellular neural networks (CNNs) has recently become an influential paradigm of complexity research and is being realized in information and chip technology (Chua and Roska, 2002; Mainzer, 2007). CNNs have been made possible by the sensor revolution of the late 1990s. Cheap sensors and micro-electro-mechanical system (MEMS) arrays have become popular as artificial eyes, noses, ears, tastes, and somatosensor devices. An immense number of generic analog signals have been processed. Thus, a new kind of chip technology, similar to signal processing in natural organisms, is needed. Analog cellular computers are the technical response to the sensor revolution, mimicking the anatomy and physiology of sensory and processing organs. A CNN is their hard core, because it is an array of analog dynamic processors or cells.

In general, a CNN is a nonlinear analog circuit that processes signals in real time. It is a multicomponent system of regularly spaced identical units called cells that communicate directly with each other only through their nearest neighbors. The locality of direct connections is a natural principle which is also realized by brains and CA. Total connectivity would be energetically too expensive with the risk of information chaos. Therefore, it was selected by evolution of the brain and not applied in technology. Unlike conventional CA, CNN host processors accept and generate analog signals in continuous time with real numbers as interaction values. The dynamics of a cell's state are defined by a nonlinear differential equation (CNN state equation) with scalars for state, output, input, threshold, and coefficients, called synaptic weights, modeling the intensity of synaptic connections of the cell with the inputs and outputs of the neighbor cells. The CNN output equation connects the states of a cell with the outputs.

CNN arrays are extremely useful for standards in visual computing. Examples are CNNs that detect patterns in either binary (black-and-white) or gray-scale input images. An image consists of pixels corresponding to the cells of CNN with binary or gray scale. From the perspective of nonlinear dynamics, it is convenient to think of standard CNN state equations as a set of ordinary differential equations. Contrary to the usual CA approach with only geometric pattern formation of cells, the dynamical behavior of CNNs can be studied analytically by nonlinear equations. Numerical examples deliver CNNs with limit cycles and chaotic attractors. For technical implementations of CNNs, such as silicon chips, complete stability properties must be formulated, in order to avoid oscillations, chaotic, and noise phenomena. These results also have practical importance for image processing applications of CNNs. As brains and computers work with units in two distinct states, the conditions of bistability are studied in brain research, as well as in chip technology.

CNNs are optimal candidates to simulate local synaptic interactions of neurons generating collective macro phenomena. Hallucinations, for example, are the results of self-organizing phenomena within the visual cortex. This type of pattern perception seems to be similar to pattern formation of fluids in chemistry or aerodynamics. Pattern formation in the visual brain is due to local nonlinear coupling among cells. In the living organism, there is a spatial transformation between the pattern perception of the retina and the pattern formation within the visual cortex of the brain. First simulations of this cortico-retinal transformation by CNNs generate remarkable similarities with pattern perceptions that are well known from subjective experiences of hallucinations. Perceptions of a spiraling tunnel pattern have been reported by people who were clinically dead and later revived. The light at the end of the tunnel has sometimes been interpreted as religious experiences.

CNNs with information processing in nanoseconds and even the speed of light seem to be optimal candidates for applications in neurobionics. Obviously, there are surprising similarities between CNN architectures and, for example, the visual pathway of the brain. An appropriate CNN approach is called the ''Bionic Eye'', which involves a formal framework of vision models combined and implemented on the so-called CNN universal machine. Like a universal Turing machine, a CNN universal machine can simulate any specialized CNN and is technically constructed in chip technology. Visual illusions which have been studied in cognitive psychology can also be simulated by a universal CNN chip. The same architecture of a universal machine can not only be used to mimic the retinas of animals (e.g., of a frog, tiger salamander, rabbit, or eagle), but they can also be combined and optimized for technical applications. The combination of biological and artificial chips is no longer a science fiction-like dream of cyborgs, but a technical reality with inspiring ramifications for robotics and medicine.

In epileptology, clinical applications of CNN chips have already been envisaged. The idea is to develop a miniaturized chip device for the prediction and prevention of epileptic seizures. Nonlinear time series analysis techniques have been developed to characterize the typical EEG patterns of an epileptic seizure and to recognize the phase transitions leading to the epileptic neural states. These techniques mainly involve estimates of established criteria such as correlation dimension, Kolmogorov–Sinai-entropy, Lyapunov exponents, fractal similarity, etc. Implantable seizure predictions and prevention devices are already in use with Parkinsonian patients. In the case of epileptic processes, such a device would continuously monitor features extracted from the EEG, compute the probability of an impending seizure, and provide suitable prevention techniques. It should also possess both a high flexibility for tuning to individual patient patterns and a high efficacy to allow the estimation of these features in real time. Eventually, it should have low energy consumption and be small enough to be implemented in a miniaturized, implantable system. These requirements are optimally realized by CNNs, with their massive parallel computing power, analog information processing, and capacity for universal computing.

In complex dynamical systems of organisms monitoring and controlling are realized on hierarchical levels. Thus, we must study the nonlinear dynamics of these systems in experimental situations, in order to find appropriate order parameters and to prevent undesired emergent behavior as possible attractors. From the point of view of systems science, the challenge of organic computing is controlled emergence.

A key application is the nonlinear dynamics of brains. Brains are neural systems which allow quick adaption to changing situations during lifetime of an organism. In short they can learn, assess, and anticipate. The human brain is a complex system of neurons self-organizing in macroscopic patterns by neurochemical interactions. Perceptions, emotions, thoughts, and consciousness correspond to these patterns. Motor knowledge, for instance, is learnt in an unknown environment and stored implicitly in the distribution of synaptic weights of the neural nets. Technically, self-organization and pattern emergence can be realized by neural networks, working like brains with appropriate topologies and

learning algorithms. Neural networks are complex systems of threshold elements with firing and nonfiring states, according to learning strategies (e.g., Hebbian learning). Beside deterministic homogeneous Hopfield networks, there are so-called Boltzmann machines with stochastic network architecture of nondeterministic processor elements and a distributed knowledge representation which is described mathematically by an energy function. While Hopfield systems use a Hebbian learning strategy, Boltzmann machines favor a backpropagation strategy (Widrow–Hoff rule) with hidden neurons in a many-layered network (Mainzer, 2003).

In general, it is the aim of a learning algorithm to diminish the informatic–theoretic measure of the discrepancy between the brain's internal model of the world and the real environment via self-organization. The interest in the field of neural networks is mainly inspired by the successful technical applications of statistical mechanics and nonlinear dynamics to solid state physics, spin glass physics, chemical parallel computers, optical parallel computers, or laser systems. Other reasons are the recent development of computing resources and the level of technology which make a computational treatment of nonlinear systems more and more feasible.

A simple robot with diverse sensors (e.g., proximity, light, collision) and motor equipment can generate complex behavior by a self-organizing neural network. In the case of a collision with an obstacle, the synaptic connections between the active nodes for proximity and collision layer are reinforced by Hebbian learning: a behavioral pattern emerges, in order to avoid collisions in future (Pfeifer and Scheier, 2001). In the human organism, walking is a complex bodily self-organization, largely without central control of brain and consciousness: It is driven by the dynamical pattern of a steady periodic motion, the attractor of the motor system.

What can we learn from nature? In unknown environments, a better strategy is to define a low-level ontology, introduce redundancy — which is commonly prevalent in sensory systems, for example — and leave room for self-organization. Low-level ontologies of robots only specify systems like the body, sensory systems, motor systems, and the interactions among their components, which may be mechanical, electrical, electromagnetic, thermal, etc. According to the complex systems approach, the components are characterized by certain microstates generating the macrodynamics of the whole system.

Take a legged robot. Its legs have joints that can assume different angles, and various forces can be applied to them. Depending on the angles and the forces, the robot will be in different positions and behave in different ways. Further on, the legs have connections to one another and to other elements. If a six-legged robot lifts one of the legs, this changes the forces on all the other legs instantaneously, even though no explicit connection needs to be specified. The connections are implicit. They are enforced through the environment, because of the robot's weight, the stiffness of its body, and the surfaces on which it stands. Although these connections are elementary, they have not been made explicit by the designer. Connections may exist between elementary components that we do not even realize. Electronic components may interact via electromagnetic fields that the designer is not aware of. These connections may generate adaptive patterns of behavior with high fitness degrees (order parameter). But they can also lead to sudden instability and chaotic behavior. In our example, communication between the legs of a robot can be implicit. In general, much more is implicit in a low-level specification than in a high-level ontology. In restricted simulated agents, only what is made explicit exists, whereas in the complex real world, many forces exist and properties obtain, even if the designer does not explicitly represent them. Thus, we must study the nonlinear dynamics of these systems in experimental situations, in order to find appropriate order parameters and to prevent undesired emergent behavior as possible attractors.

But not only "low level" motor intelligence, but also "high level" cognition (e.g., categorization) can emerge from complex bodily interaction with an environment by sensory–motor coordination without internal symbolic representation. We call it "embodied cognition": an infant learns to categorize objects and to build up concepts by

touching, grasping, manipulating, feeling, tasting, hearing, and looking at things, and not by explicit symbolic representations (e.g., language). The categories are based on fuzzy patchworks of prototypes and may be improved and changed during life. We have an innate disposition to construct and apply conceptual schemes and tools.

Moreover, cognitive states of persons depend on emotions. We recognize emotional expressions of human faces with pattern recognition of neural networks and react by generating appropriate facial expressions for nonverbal communication. Emotional states are generated in the limbic system of the brain which is connected with all sensory and motoric systems of the organism. All intentional actions start with an unconscious impulse in the limbic system which can be measured before their performance. Thus, embodied intentionality is a measurable feature of the brain (Freeman, 2004). Humans use feelings to help them navigate the ontological trees of their concepts and preferences, to make decisions in the face of increasing combinational complexity. Obviously, emotions help to reduce complexity.

The embodied mind is a complex dynamical system acting and reacting in dynamically changing situations. The emergence of cognitive and emotional states is made possible by brain dynamics which can be modeled by neural networks. According to the principle of computational equivalence (Mainzer, 2007), any dynamical system can be simulated by an appropriate computational system. But, contrary to Turing's AI-thesis, that does not mean computability in every case. In complex dynamical systems, the rules of locally interacting elements (e.g., Hebb's rules of synaptic interaction) may be simple and programmed in a computer model. But their nonlinear dynamics can generate complex patterns and system states which cannot be forecast in the long run without increasing loss of computability and information. Thus, artificial minds could have their own intentionality, cognitive, and emotional states which cannot be forecast and computed like in the case of natural minds. Limitations of computability are characteristic features of complex systems.

In a complex dynamical world, decision-making and acting is only possible under conditions of bounded rationality. Bounded rationality results from limitations on our knowledge, cognitive capabilities, and time. Our perceptions are selective, our knowledge of the real world is incomplete, our mental models are simplified, our powers of deduction and inference are weak and fallible. Emotional and subconscious factors affect our behavior. Deliberation takes time and we must often make decisions before we are ready. Thus, knowledge representation must not be restricted to explicit declarations. Tacit background knowledge, change of emotional states, personal attitudes, and situations with increasing complexity are challenges of organic computing.

In a dramatic step, the complex systems approach has been enlarged from neural networks to global computer networks like the World Wide Web. It is not only a metaphor to call them global "superbrains". The internet can be considered a complex open computer network of autonomous nodes (hosts, routers, gateways, etc.), self-organizing without central mechanisms. Routers are nodes of the network determining the local path of each information packet by using local routing tables with cost metrics for neighboring routers. These buffering and resending activities of routers can cause congestions in the internet. Congested buffers behave in surprising analogy to infected people. There are nonlinear mathematical models describing true epidemic processes like malaria extension as well as the dynamics of routers. Computer networks are computational ecologies.

But complexity of global networking does not only mean increasing numbers of PCs, workstations, servers, and supercomputers interacting via data traffic in the internet. Below the complexity of a PC, low-power, cheap, and smart devices are distributed in the intelligent environments of our everyday world. Like GPS in car traffic, things in everyday life could interact telematically by sensors. The real power of the concept does not come from any one of these single devices. In the sense of complex systems, the power emerges from the collective interaction of all of them. For instance, the optimal use of energy could be considered a macroscopic order parameter of a

household realized by the self-organizing use of different household goods according to less consumption of electricity during special time-periods with cheap prices. The processors, chips, and displays of these smart devices do not need a user interface like a mouse, windows, or keyboards, but just a pleasant and effective place to get things done. Wireless computing devices on small scales become more and more invisible to the user. Ubiquitous computing enables people to live, work, use, and enjoy things directly without being aware of their computing devices.

A challenge of the automobile industry is the increasing complexity of electronic systems. If we consider the electronic cable systems of automobiles from the beginning through to today, there will be a surprising similarity to neural networks of organisms which increase in complexity during evolution. Contrary to biological evolution, electronic systems of today are rigid, compact, and flexible. In an evolutionary architecture (EvoArch) the nervous system of an automobile is divided into autonomous units (carlets) which can configurate themselves in cooperative functions, in order to solve intelligent tasks (Hofmann et al., 2002). They are the macroscopic features realized by interacting unities in a complex system. Examples are the complex functions of motor, brake, and light, wireless guide systems like GPS, smart devices for information processing, and the electronic infrastructure of entertainment. In an evolutionary electronic architecture (EvoArch), there are several ''self-x-features'' with great similarity to self-organizing organic systems in biological evolution: self-healing demands self-configuration and self-diagnosis. Self-diagnosis means error recognition and self-reflection, etc. In short: the principles of self-organizing brains are realized in the electronic hardware of cars.

Perspectives of modeling and computing with complex dynamical systems

What is the reason behind the successful interdisciplinary applications of nonlinear complex systems? This approach cannot be reduced to special natural laws of physics, although its mathematical principles were discovered and at first successfully applied in physics. Thus, it is no kind of traditional physicalism to explain the dynamics of laser, ecological populations or our brain by similar structural laws. From a formal point of view, complex systems may be multi-component systems of, for example, atoms, molecules, cells, or organisms. If certain control parameters are changed, the interactions of elements in multi-component systems may lead to new collective macroscopic properties of order or chaos which cannot be reduced to the individual elements on the microlevel.

The emergence of order and chaos depends essentially on the nonlinearity of evolutionary equations modeling the dynamics of complex systems. Further conditions come in by the specific parameters of physical, chemical, biological, psychological, or computational systems. Therefore, the formal models of nonlinear complex systems do not eliminate the requirements of specific experimental research on the different levels and scales in the different sciences. Interdisciplinary applications of nonlinear complex systems are successful if they find a clever combination of formal mathematics, computer-assisted modeling, and disciplinary research. Complexity and nonlinearity are interdisciplinary problems of current research. Thus, the formal analysis and computer-assisted modeling of complex systems must be combined with experimental and empirical research in the natural and social sciences. According to a famous quotation of Kant, we may conclude that formal models without specific disciplinary research are empty, while disciplinary research without common principles is blind.

Organic and neural computing not only aims at modeling but also at constructing self-organizing computing systems that display desired emergent behavior like organisms in natural evolution (Horn, 2001; Kephart and Chess, 2003; Müller-Schloer, 2005). Emergence refers to a property of a system that is not contained in anyone of its parts. In the sense of nonlinear dynamical systems, the whole is more than the sum of its parts. In robotics, it concerns behavior resulting from the agent–environment interaction whenever the behavior is not preprogrammed. It is thus not common

to use the term if the behavior is entirely prespecified like a trajectory of a hand that has been precalculated by a planner. Agents designed using high-level ontologies have no room for emergence, for novel behaviors. A domain or high-level ontology consists of a complete representation of the basic vocabulary, the primitives, that are going to be used in designing the system. These are the only components that can be used: everything is built on top of these basic elements. The domain ontology remains constant for an extended period of time, often for the entire life of the system. A well-known example is the bounded knowledge representation of an expert system. High-level ontologies are therefore used whenever we know precisely in what environments the systems will be used, as for traditional computational systems as well as for factory robot systems. In unknown environments, a better strategy is to define a low-level ontology and to introduce redundancy with a great variety of self-organization.

In the dynamical systems approach, we first need to specify what system we intend to model and then we have to establish the differential or difference equations. Time series analysis and further criteria of data mining help to construct the appropriate phase spaces, trajectories, and attractors (Small, 2005). In organic computing, one approach would be to model an agent and its environment separately and then to model the agent–environment interaction by making their state variables mutually dependent (Pfeifer and Scheier, 2001). The dynamical laws of the agent A and the environment E can be described by simplified schemes of differential equations $\mathrm{d}x_\mathrm{a}/\mathrm{d}t = A(x_\mathrm{a}, p_\mathrm{a})$ and $\mathrm{d}x_\mathrm{e}/\mathrm{d}t = E(x_\mathrm{e}, p_\mathrm{e})$, where x represents the state variables, such as angles of joints, body temperature, or location in space, and p parameters like thresholds, learning rates, nutrition, fuel supply and other critical features of change. Agents and environment can be coupled by defining a sensory function S and a motor function M. The environment influences the agent through S. The agent influences its environment through M. S and M constitute the agent–environment coupling, that is, $\mathrm{d}x_\mathrm{a}/\mathrm{d}t = A(x_\mathrm{a}, S(x_\mathrm{e}), p_\mathrm{a})$ and $\mathrm{d}x_\mathrm{e}/\mathrm{d}t = E(x_\mathrm{e}, M(x_\mathrm{a}), p_\mathrm{e})$, where p_a and p_e are not involved in the coupling. Examples are walking or moving robots in environments with obstacles. In this case, the basic analysis problem can be stated in the following way: given an environment dynamics E, an agent dynamics A, and sensory and motor functions S and M, explain how the agent's observed behavior is generated.

One of the controllers of the dynamics evolves when the agent's angle sensors are turned off and cannot sense the position of its legs. In this case, the activation levels of the neurons exhibit a limit cycle that causes the agent's single leg to stand and swing rhythmically. By that, it causes the robot to walk. The system's state repeatedly changes from the stance phase with the foot on the ground to the swing phase with the foot in the air and back. This example illustrates that the dynamical systems approach can be applied in a synthetic way in order to design and construct robots and their environments. But, in general, the dynamical systems approach is used in an analytical way: it starts from a given agent–environment interaction, which is formalized in terms of differential equations. The complex variety of behavior can be analyzed by solving, approximating, or simulating the equations, in order to find the attractors of dynamics. The dynamical attractors of the interacting system can be used to steer an agent or to let it self organize in a desired way.

Obviously, self-organization leads to the emergence of new phenomena on sequential levels of evolution. Nature has demonstrated that self-organization is necessary, in order to manage the increasing complexity on these evolutionary levels (Mainzer, 2005). But nonlinear dynamics can also generate chaotic behavior which cannot be predicted and controlled in the long run. In complex dynamical systems of organisms monitoring and controlling are realized on hierarchical levels. There is still no final and unified theory of organic computing. We only know parts of biological, neural, cognitive, and social systems in the framework of complex dynamical systems. But even in physics, we have no unified theory of all physical forces. Nevertheless, scientists work successfully with an incomplete patchwork of theories. In order to know more about it, we need an interdisciplinary cooperation of technical, natural,

computer, and cognitive science, and last but not the least humanities. The goal of organic and neural computing is the construction of self-organizing computing systems to be of service for the people, in order to manage a world of increasing complexity and to support a sustainable future of human infrastructure.

In summary, we conclude with following statements

- Natural evolution of brain and mind
  Natural evolution generates nervous systems working with neurochemical mechanisms. They enable organisms to learn, adapt, and change their environment autonomously. Thinking, feeling, and consciousness are considered mental states of neural dynamics evolving in the human organism and interacting with its environment (embodied mind).
- Nonlinear dynamics of complex systems
  The natural evolution of brain and mind is an example of nonlinear dynamics of complex systems. The emergence of order in complex systems is made possible by, for example, thermodynamics, genetic, and neural self-organization. Hierarchical levels of neural self-organization lead to the emergence of mental states and functions which can (in principle) be modeled by order parameters (attractors) of nonlinear dynamics.
- Artificial life and artificial intelligence
  Under different conditions, the laws of nonlinear dynamics would also have allowed variations of life different from the organisms which actually evolved on Earth. Therefore, technology can use the laws of nonlinear dynamics to find similar or new solutions of self-organizing systems in, for example, bionics, artificial intelligence, and artificial life.
- Self-organization and the emergence of human mind
  Nonlinear dynamics of self-organizing complex systems generate emerging properties and functions which often cannot be forecast in the long run. Therefore, neuropsychic systems may lead to individual feelings, perceptions (qualia), intentions (intentionality), and self-consciousness which cannot be derived from single neural activities, but by synergetic interactions of the whole system.
- Dynamical and computational systems
  According to the principle of computational equivalence, complex dynamical systems can be considered computational systems (e.g., CA, neural networks). Nevertheless, complex computational systems can lead to chaos, randomness, and undecidability. Therefore, computational systems (e.g., stochastic systems) can also simulate brain dynamics with their undetermined features.
- Self-organization and human technology
  Contrary to the Laplacian Spirit, the laws of nonlinear dynamics exclude the total computability of nature, life, and mind. But, we can understand their phase transitions in order to find conditions (control parameters) making the emergence of desired states and developments (order parameters) possible and probable (e.g., health, wellness in medicine and psychology). Empathy and responsibility are made possible by human brain dynamics.
- Superbrain and mankind
  In the age of globalization, mankind is growing together by worldwide information and communication systems. They are the self-organizing nervous systems of an emerging superbrain with increasing complexity. It is a challenge to find the conditions of global governance in the virtual networks of global organizations.
- Philosophical and religious perspectives
  In the philosophical and religious tradition of mankind (e.g., Buddhism), the emergence of a global consciousness is assumed to evolve in steps with increasing clearness from individual consciousness to the soul of the world. Nonlinear science opens avenues to follow this line from a scientific point of view.

References

Bassingthwaighte, J.B., Liebovitch, L.S. and West, B.J. (1994) Fractal Physiology. Oxford University Press, New York.

Chua, L.O. and Roska, T. (2002) Cellular Neural Networks and Visual Computing. Foundations and Applications. Cambridge University Press, Cambridge.

Deco, G. and Schürmann, B. (2000) Information Dynamics. Foundations and Applications. Springer, New York.

Dennett, C.D. (1998) Brainchildren: Essays on Designing Minds. MIT Press, Cambridge, MA.

Dreyfus, H.L. (1982) Husserl, Intentionality, and Cognitive Science. MIT Press, Cambridge, MA.

Floridi L. (2004). Philosophy of Computing and Information. Blackwell, Oxford.

Freeman, W.J. (2004) How and why brains create meaning from sensory information. Int. J. Bifurcat. Chaos, 14: 515–530.

Haken H. and Mikhailov A. (Eds.), (1993). Interdisciplinary Approaches to Nonlinear Complex Systems. Springer, Berlin.

Hofmann, P.H., Lukas, G., Wohlgemuth, F., Schneider, B., Müller, A., Schmidt, U., Keydel, M., Sakretz, R., Leboch, S., Müller, T., Klenk, U., Perans, A. and Dohmeyer, V. (2002) Evolutionäre E/E-Architektur. DaimlerChrysler, Esslingen.

Horn, P. (2001) Autonomic Computing: IBM's Perspective on the State of Information Technology. IBM, New York.

Kephart, J.O. and Chess, D.M. (2003) The Vision of Autonomic Computing. IEEE Comput. Soc., 1: 41–50.

Mainzer, K. (2003). KI — Künstliche Intelligenz. Grundlagen intelligenter Systeme, Wissenschaftliche Buchgesellschaft, Darmstadt.

Mainzer, K. (2007) Thinking in Complexity. The Computational Dynamics of Matter, Mind, and Mankind (5th enlarged ed.). Springer, New York.

Mainzer, K. (2005) Symmetry and Complexity. The Spirit and Beauty of Nonlinear Science. World Scientific Publisher, Singapore.

Merleau-Ponty, M. (1962) Phenomenology of Perception. Routledge & Kegan Paul, London.

Müller-Schloer, C. (Ed.). (2005). Schwerpunktthema: Organic Computing — Systemforschung zwischen Technik und Naturwissenschaften. it Inf. Technol., 47(4): 179–181.

Pfeifer, R. and Scheier, C. (2001) Understanding Intelligence. MIT Press, Cambridge, MA.

Searle, J.R. (1983) Intentionality. An Essay in the Philosophy of Mind. Cambridge University Press, Cambridge.

Small, M. (2005) Applied Nonlinear Time Series Analysis. Applications in Physics, Physiology and Finance. World Scientific Publisher, Singapore.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 11

Dynamic geometry, brain function modeling, and consciousness

Sisir Roy[1,*] and Rodolfo Llinás[2]

[1]*Physics and Applied Mathematics Unit, Indian Statistical Institute, Calcutta, India; College of Science, George Mason University, Fairfax, VA, USA*
[2]*Department of Neuroscience and Physiology, New York University School of Medicine, MSB 4 448, 550 First Avenue, New York, NY 10016, USA*

Abstract: Pellionisz and Llinás proposed, years ago, a geometric interpretation towards understanding brain function. This interpretation assumes that the relation between the brain and the external world is determined by the ability of the central nervous system (CNS) to construct an internal model of the external world using an interactive geometrical relationship between sensory and motor expression. This approach opened new vistas not only in brain research but also in understanding the foundations of geometry itself. The approach named tensor network theory is sufficiently rich to allow specific computational modeling and addressed the issue of prediction, based on Taylor series expansion properties of the system, at the neuronal level, as a basic property of brain function. It was actually proposed that the evolutionary realm is the backbone for the development of an internal functional space that, while being purely representational, can interact successfully with the totally different world of the so-called ''external reality''. Now if the internal space or functional space is endowed with stochastic metric tensor properties, then there will be a dynamic correspondence between events in the external world and their specification in the internal space. We shall call this dynamic geometry since the minimal time resolution of the brain (10–15 ms), associated with 40 Hz oscillations of neurons and their network dynamics, is considered to be responsible for recognizing external events and generating the concept of simultaneity. The stochastic metric tensor in dynamic geometry can be written as five-dimensional space-time where the fifth dimension is a probability space as well as a metric space. This extra dimension is considered an imbedded degree of freedom. It is worth noticing that the above-mentioned 40 Hz oscillation is present both in awake and dream states where the central difference is the inability of phase resetting in the latter. This framework of dynamic geometry makes it possible to distinguish one individual from another. In this paper we shall investigate the role of dynamic geometry in brain function modeling and the neuronal basis of consciousness.

Keywords: dynamic geometry; tensor network; consciousness; dream state; central nervous system

*Corresponding author. Tel.: +91-33-25753021;
Fax: +91-33-25776680; E-mail: sisir@isical.ac.in

DOI: 10.1016/S0079-6123(07)68011-X

> We think a given thought, then the meaning of this thought is expressed in the shape of corresponding neurophysiological process.
>
> B. Riemann

Introduction

The proposal for a geometrical interpretation of brain function by Pellionisz and Llinás (1982, 1985) and Llinás (2002) introduced an integrated approach to understanding brain function. It was originally based on the assumption that the relationship between the brain and the external world is determined by the ability of the central nervous system (CNS) to construct an internal model of the world accomplished through the interactive relationship between sensory and motor expression. In this model the evolutionary realm provides the backbone for the development of an internal functional geometry space. Almost a century before Mach (1959) investigated this issue in the context of the analysis of sensations and geometry. He emphasized at that time that without co-operation among sensory perceptions, in the sense of inductive reasoning, the understanding of a scientific geometry would be inconceivable. This is consistent with Indian geometry where inductive reasoning is the dominant approach while in the more familiar Greek geometry pure ''understanding'' (deductive reasoning) dominates.

In tensor network theory internal space is viewed as isomorphic to the external world allowing successful operational interactions between them, with the possibility of functional prediction necessary for the implementation of even the simplest motor coordination paradigm guided by brain function. Accordingly the nervous system was regarded as a sensory–motor transformation entity that handled sensory input such that well executed motor output is delivered back to the external world in a physically acceptable fashion via a plant (the body) that is totally different in kind, shape and functionality to the structure and characteristic of the external world. Consider the unlikely interaction of a living organism and a completely different entity, such as a musical instrument made of brass or wood. The relationship between animate and inanimate objects (an evolutionary event) can be perfectly coordinated despite their having different natures. Such strange fellowship was originally considered to have developed given a metric tensor transformation, via the geometry of abstract spaces. Thus, sensory input, a covariant vector, represents the properties of the external world as measured by the senses.

The covariance relates both due to the fact that the sensory apparatus measures large sets of parallel, but independent, multiple fiber input determinations, and to the fact that a given external percept must covary with the dynamic variance in the external world it measures. In the brain such covariant input must be converted into a contravariant vector representation. This internalized and transformed vector description can then be used to activate, in a coordinated, independent manner the manipulative counterpart of the motor response (a totally different functional space) that makes the object–subject interaction possible.

Before going into the details of these frameworks, the applicability of functionally unbound (in the sense of closed or recurrent) and ''smooth'' metric tensors (in an otherwise bounded and nonlinear nervous system) should be thoroughly discussed. Because the metric tensor is defined over an anatomically bounded and closed system, it is reasonable to assume that the metric tensor is structurally but not functionally bounded. The analysis of sensory perception requires differentiation. This differentiability of the functional geometry ensures the smoothness of internal space-time. Considering the smooth metric tensor it is possible to define the derivatives of the metric tensor. This is an essential step in the construction of a non-Euclidean internal space-time. The assumption of a contravariant vectorial transformation, as associated with the motor output and more importantly with intentionality as a premotor event (Llinás, 2002), must be generated from an internal description of sensory simultaneity.

Here we introduce the concept of a "stochastic metric tensor" ("stochastic space-time") as an appropriate description of CNS function. This resolves several fundamental limitations related to the original tensor network theory (Pellionisz and Llinás, 1985). We call this *dynamic geometry* that is associated with the neuronal activities of brain.

In this paper, we approach the problem of whether it is possible to find a mathematical formalism (a "model") that describes the function of the brain. Such a formalism would lead to a set of equations that could be solved, even in the sense of numerical solutions. It would also lead to the design of a type of "computing machine" which could be built in the real world (Caianiello, 1961). The question arises whether the CNS can act as a computing machine or not. Wiener (1948) began this kind of discussion in a systematic way in the middle of last century. Von Neumann (1951) discussed the issue of a computing machine and the CNS from a digital and an analog perspective.

In the following four sections we begin with a brief survey of brain function from a modern perspective. In section "Brain Function from a Modern Perspective" several fundamental issues related to functional geometry as proposed by one of the authors (RL) is critically analyzed. In section "Functional Geometry and the Central Nervous System" we introduce the concept of stochastic space-time and hence dynamic geometry. Finally, in section "Fluctuating Metric and Dynamic Geometry", some implications of the role of dynamic geometry are discussed.

Brain function from a modern perspective

Donald H. Perkel (1990) wrote "The question that immediately arises is whether the biological phenomena themselves dictate or justify the theory's mathematical structures. The alternative is that the beauty, versatility, and power of mathematical approach may have led its aficionado to find areas of application in the spirit of the proverbial small boy with a hammer, who discovers an entire world in need of pounding".

Three apparently distinct queries in brain functions can be stated as

(a) The unknown related to how the brain represents the world and how it performs transformations on this representation.
(b) The unknown related to sensorimotor coordination.
(c) The unknown related to the physical basis for perceptual organization.

The internalization of the properties of the external world is the central issue in brain research. Plato (1991) in his well-known allegory of the prisoners in the cave discussed this kind of issue long before the development of modern brain research. Following Wittgenstein (1997) we can distinguish between the world as the domain of our experience and the world as the domain of things in themselves.

So, the central issue in brain function is the internalization of the properties of the external world *into an internal functional space*. By internalization, we mean the ability of the nervous system to fracture external reality into sets of sensory messages and to simulate such reality in brain reference frames. We will study this internalization within the framework of dynamic geometry.

One of us (RL) proposed that the brain should be considered as a fundamentally closed system that is perforated and modulated by the senses. This is one of the basic issues that needs to be investigated with great rigor. The concept of open and closed systems has been discussed for many decades in the physical sciences. Developments in the study of thermodynamics helped to formalize conceptual structures for open and closed systems. The work of Prigogine and Kondepudi (1998) took a further step in this direction. The issue of an open or closed system for biological systems become one of the central issues in 20th century biophysics. Broadly speaking, an open system in physics is defined as a system that exchanges both energy and mass with the outside surroundings or its environment. In case of a biological system, the system that accepts inputs from the environment processes them and returns them to the external world as a reflex is considered an open system.

Let us first analyze this issue since it plays an important role in understanding brain function.

The brain as a closed system that is modulated by the senses

The idea of *tabula rasa* was discussed in western philosophy even in the writings of Aristotle (2004). Aristotle wrote about the inscribed tablet which is probably the first textbook of psychology in the western paradigm.

We consider the CNS as a closed system in the sense that its basic organization generates intrinsic images such as thoughts or predictions whereas inputs specify the internal states rather than "homuncular vernacles". So the CNS possesses an autonomy similar to that proposed by Maturana (2000). However, they introduced the term as an operational closure of the nervous system rather than considering it as a closed system.

von Bertalanffy (1968) devoted much time to understanding the distinction between open and closed systems in the context of living organisms. He thought that the very distinction of closed and open systems was fundamental to understanding the contrasts of the physical and living worlds. He (von Bertalnffy, 1967) discussed the issue in the perspective of thermodynamics as well as within the purview of general system theory. We will discuss these aspects shortly from the point of view of functional geometry.

In physics, systems may be generally classified into the following three categories:

(a) Isolated system: There is no transfer of matter or energy from the system to the environment.
(b) Closed system: There is no transfer of matter but exchange of energy is allowed and the system is open with respect to energy.
(c) Open system: This allows the transfer of matter as well as energy.

In thinking of the brain as a "closed system" we do so with respect to the exchange of the intrinsic recurrent interactions of neuronal circuits with the environment but which is open to the exchange of matter (at the molecular level), information, or energy. Thus, within the approach of functional geometry, the brain is viewed as embedding intrinsic degrees of freedom that are open to the exchange of sensory input that modulates the ongoing intrinsic "closed activity". This is like changing the direction of the spin of an elementary particle when the particle is placed in an external magnetic field. The CNS was also considered as a closed system by Victor Hamburger. To him, the basic organization of the CNS is responsible for the generation of internal states (Hamburger) rather than inform "homuncular vernacles". Internal neuronal properties have been extensively discussed (Bekoff, 1981).

The brain as a self-referential system

Recent in vitro recordings from neurons in the inferior olivary nucleus demonstrated, for the first time, that phase reset in neuronal intrinsic oscillations (Leznik et al., 2002) differs from typical oscillatory systems. Here, the phase reset is controlled by input parameters and does not depend on the time moment, i.e. the initial phase, when the input is received. In this sense, the phase reset is self-referential and ignores the history of the system. Makarenko and Llinás (1998) and Kazantsev et al. (2003), proposed a mathematical model using a set of non-linear differential equations to describe the self-referential phase reset (Kazantsev, 2004). The main motivation of this modeling was to implement phase resetting into a motor control system based on a dynamics theory.

40 Hz oscillations and the quanta of time

Magnetic and electric recordings from the human brain revealed the existence of coherent oscillatory activity near 40 Hz. A magneto-encephalography (MEG) system was used by Joliot et al. (1994) to test whether the 40 Hz oscillatory activity relates to the temporal binding of sensory stimuli. The results showed that the 40 Hz oscillations not only relate to primary sensory processing, but also could reflect the temporal binding underlying cognition. Experimental results have shown that there exists a time interval of 10–14 ms (corresponding to the up trajectory of 40 Hz oscillations) that is the minimum time required for the binding of sensory inputs into the cognition of any single

event. This was proposed as the cognitive ''quanta of time''.

Tensor network theory and neural networks

Neural network models concerning brain function have been developed and discussed over the last five decades by numerous authors (McCulloch and Pitts, 1943). The concept of state vector was introduced to describe both CNS input and output functions. A state vector is a point (inner product of all individual vector values) in a multidimensional space. This space also requires geometric notions like ''nearness'' or ''angle'' between points and hence the concept of ''distance'' in state space. Amari (1994), Von der Malsburg (1998), Kohonen (2001) considered various types of spaces (like parametric space by Amari, 1994) to explain the topographic maps in the brain. In their analysis, units physically near to one another in an array are more correlated in their response properties than units that are further away in the array. Most of these networks are developed to determine the statistical structure of the environment. There is another important aspect of the environment, which is not considered in most of these networks — the presence of the observer and his reaction to the observed world. Thus, motor activity needs to be considered in any biological neural computation. In this case, the role of the sensory–motor and the motor–sensory loops (completed by the presence of an environmental perceptual target) is very important in a more realistic biological computation. The external world where the organisms live is not a higher dimensional space as considered in neural networks. In fact, this external world can be described by three-dimensional Euclidean geometry. Perceptual spaces have colors and hues and specially the olfactory space (probably affine) has thousands of convergence points but no geodesics putting them together (no origin to the vector components). However, the geometry of motor action is not necessarily limited to three dimensions as there are hundreds of muscles and they act in different ways and a particular body motion involves the coordinated contraction of many muscles. Usually, the motor output is underdetermined so that different patterns of motor neuron output may give rise to the same overall forces. This resultant or overall force can be described by the simple laws of mechanics. It is well known that each muscle contains a set of proprioceptive sensors that informs the brain about the particular muscle and the force it is exerting. The sensorimotor system uses this input as primary information to close the loop in the interaction with the environment. Given the above, a mathematical model may be proposed using a set of coordinate transformations involving sensory input, motor proprioception, and motor output. We will consider such transformations in terms of the original tensor network theory (Pellionisz and Llinás, 1985) as the starting point.

Functional geometry and the central nervous system

A central question concerning present day neuroscience is that of understanding the rules for the embedding of ''universals'' into intrinsic functional space. One of the present authors (RL) addresses this internal space as a dynamic geometry. Now one needs to understand the fundamental structure of this internal space or functional geometry before going into the details of its applications. They can be stated as follows

(1) For any geometry one needs to define the ''smoothness'' property of the manifold. Thus in functional geometry the existence of derivatives associated with the sensory covariant vector and motor contravariant vector is the prerequisite of the functional manifold, and so a definition of differentiability in the functional space becomes necessary.
(2) Non-orthogonal coordinate axes have been considered to be associated with covariant and contravariant vectors based on physiological observations (Pellionisz and Graf, 1987; Graf et al., 1992). So one must consider the non-orthogonal frame of references in this type of functional space.
(3) The analysis of the contravariant character of the forces exerted by the muscles that such a geometry must accommodate includes an

overcomplete set of possible dynamic configurations associated with a non-orthogonal frame of reference leading to similar motor execution. Accordingly, the CNS must calculate an inverse solution via the mathematical indeterminacy inherent in the CNS in a manner similar to that implemented in Moore-Penrose based generalized inverse solutions (Pellionisz and Llinás, 1985).

(4) The motion of an object in the external world does not engender simultaneity of space and time in its counterpart functional space because the conduction speeds through various axons are different for a given stimulus (Pellionisz and Llinás, 1982).

Let us consider the possible geometrical structures for the CNS. To define any metric tensor one needs to address a well-defined distance function that satisfies all the axioms of metric tensors. From a global point of view, the anatomy of the brain does not present a smooth and linear representation of the external world. So a distance function must be constructed which can be defined as functionally isotropic. In fact, the multidimensional spaces of the CNS are, for the most part, not definable in well–known geometries such as Euclidean or Riemannian spaces. Indeed, if we consider, for instance the olfactory space with more than 10,000 different categories, we find that this space is defined by the chemical and biochemical properties of the odorant substance. This, in turn, can only be defined by the requirements of the organism. Accordingly, unlike Euclidean or Riemannian spaces, this olfactory space cannot be defined independently of the measurement instruments. It is well known in the tensor approach, that the metric tensor is $g_{ij} = \vec{e}_i \cdot \vec{e}_j$ where $\vec{e}_i$ and $\vec{e}_j$ are the unit vectors along the non-orthogonal axes. Here, if the angle between these vectors be unknown or uncertain, the metric will be uncertain. The weakly chaotic nature of neuronal oscillations (Makarenko and Llinás, 1998) may give rise to the uncertainty in the angle. This gives rise to the stochastic nature of the functional space since the metric will be stochastic. Now the question is whether we can construct a distance function associated with this kind of stochastic space which will be the proper function for the above type of functional space.

From neurophysiological observations (Makarenko and Llinás, 1998; Leznik et al., 2002; Llinás, 1988) it is clear that local chaotic oscillations may give rise to synchronous oscillations on a larger scale. Starting with the local chaotic behavior, which causes indeterminacy of the angles and hence the stochastic nature of the metric, the neuronal architecture at the global level (for which there exists synchronous oscillations) will support a smooth metric structure at this level.

Fluctuating metric and dynamic geometry

Recent neurophysiological observations indicate that quanta of time exist for both motor execution (Llinás, 1991) and sensory perception (Joliot et al., 1994). The latter is of the order of 10–14 ms and is associated with gamma band ($\approx$40 Hz) oscillatory activity in the brain (Joliot et al., 1994). Again, the delay in conduction speeds along different axons and the integration time for individual neuronal elements in the circuit are both of the same order of magnitude as the temporal quanta. So, in spite of such delays, the concept of simultaneity of the external event will be considered valid for functional space, i.e. as an operational definition of simultaneity.

In the present context, the operational definition has been used to study one event at a certain place and particular instant of time in the external world as cogitated by the brain. Here, the simultaneity is between the event in the external world and the event in the internal world. Broadly speaking, an operational definition specifies the type of observations that are relevant to making decisions about the applicability of the defined terms in a particular situation. Ingraham (1962) considered this type of operational definition in the context of stochastic space-time, relativity, and quantum field theory. Now, if the internal space or functional space is endowed with stochastic metric tensor properties, then there will be a dynamic correspondence between the event in the external world and the event in the internal space. We shall call this correspondence "Dynamic Geometry" since

the time resolution of the brain that is associated with gamma band oscillations is considered to be responsible for recognizing external events and generating the concept of simultaneity. It is evident that we are using an instrument which has well-defined finite resolution and hence we can only consider operational definitions as working hypotheses. Moreover, in our view there is no separation between the instrument and observer, and the internal space is determined by the dynamics associated with the instrument-observer continuum. This situation is different from the present understanding of the conceptual framework of the geometry and probes in physics. The pure understanding of geometry as considered by western science in the tradition of Greek thinking is not possible; at least in the context of the brain.

The stochastic metric tensor in dynamic geometry can be written as $g_{ij}=h_{ij}(x,\xi)$ where x is the four space-time coordinates (i.e., three space and one time of Minkowski space) and ξ is a fifth dimension (Roy, 1998). This fifth dimension is a probability space as well as a metric space. It satisfies the following condition

$$\int \rho(\xi)\mathrm{d}\xi = 1$$

where $\rho(\xi)$ is the probability distribution function. This extra dimension is an embedded degree of freedom. Takano (1967) considered this type of stochastic metric in their description of the interaction of elementary particles and in tackling the divergence problem in quantum field theory.

The square of the distance function can be written as

$$\mathrm{d}s^2 = h_{ij}(x,\xi)\mathrm{d}x^i\mathrm{d}x^j$$

After taking suitable averaging one gets the usual distance function in Minkowski space.

It should be noted that for Minkowski, space-time is considered to be a continuous parameter. On the other hand, in the brain, time is discrete because of the existence of quanta of time (i.e., a time granularity that simplifies brain function) (Llinás, 1991; Welsh and Llinas, 1997). As soon as we introduce the concept of probability in the geometry, i.e. stochasticity in the metric tensor (due to averaging at a certain scale, a) continuous space and time may be defined (Menger, 1945, 1951). Menger introduced the concept of probability in geometry by analyzing the issue of non-transitivity in the following manner.

The conceptual foundation of non-transitivity has been discussed elaborately by Karl Menger. In fact Poincare also discussed this problem as follows. If we take quantity A is equal to B and B is equal to C, then A is equal to C. This is true in mathematics. However, this is not true in physical world. In the physical world when we say A is equal to B, we mean A is indistinguishable from B. If we follow the language of Weber and Fechner, A may lie within the threshold of B and B lies in the threshold of C but A may not lie within the threshold of C. Poincare expressed this as "The raw result of experience" as

$$\mathrm{A} = \mathrm{B},\ \ \mathrm{B} = \mathrm{C},\ \ \mathrm{A} < \mathrm{C}$$

i.e., physical equality is not a transitive relation. The same two objects sometimes can be made distinguishable and sometimes as identical. For example, two simultaneous irritations on the same spot on the skin of a blindfolded man may sometimes appear to be one and sometimes as two distinct sensations. Karl Menger introduced the concept of probability in geometry to understand this type of distinction, i.e. transitive relation in mathematics and intransitive relations among physical and physiological quantities. Instead of simple yes or no, a number (a quantity with certain probability) is introduced lying between 0 and 1. Menger elaborated this idea in his formulation of statistical metric spaces where a distribution function rather than a definite number is associated with every pair of elements. The distance between the two points is the number associated with the two points of a metric space. McCulloch (1951) discussed the topology of nervous nets, the concept of heterarchy, and introduced the concept of non-transitivity. If one recalls the self-referential aspects of brain function, it is necessary to also include the concept of non-transitivity in our present considerations.

Let us now consider the validity of the contravariant vector associated with the motor activity in

stochastic space. This is none other than considering the contravariant vector as a physically observable quantity with a certain probability. Frederick Carlton (1976) considered the contravariant observable theorem to be a basic tenet in stochastic space-time.

Our contention is that measurements of dynamical variables are contravariant components of tensors. That is, whenever any measurement can be reduced to a displacement in a coordinate system; it can, de facto, be related to contravariant components of the coordinate system. If the metric is well known, both the covariant and contravariant quantities can be calculated. Carlton gave a simple example to elaborate this point as follows.

If we consider that the astronomical distance to an object is $\bar{r} = \bar{\xi}^1$. Let ξ_1 be the covariant equivalent of the radial coordinate r which can be written as

$$\xi_1 = g_{1i}\xi^i = \frac{r}{1 - 2\,(GM/r)}$$

the contravariant distance becomes

$$\text{Distance} = \int \mathrm{d}r = \bar{r}$$

and covariant distance is given by

$$\bar{\xi} = \int \mathrm{d}\left(\frac{r}{1 - 2\ \ (GM/r)}\right) = \infty$$

This indicates that the contravariant distance is observable (a measurable and physically meaningful quantity).

Concerning physical measurements, when making observations of a dynamical variable, e.g. position or momentum, the measurements are made with reference to a standard such as distance in meters, or current in E/R ratios. Through calculations, one can reduce the given datum to a position in a coordinate system (both in orthogonal or non-orthogonal co-ordinate systems). Margenau (1959) discussed the conditions under which this reduction may be considered as a measurement. He has shown that such a reduction must satisfy two conditions: it must be repeatable with the same results and it must be a physically useful quantity. The measurement does not necessarily mean an interaction with the system to be observed and the recording of a number. This interaction and the recording of a number does not represent simply a length, a mass, a frequency, or energy. The identification requires certain rules or correspondences with preformed theoretical concepts and hence gives rise to a physical dimension.

Frederick has shown that this physically useful quantity is contravariant if we confine ourselves to Minkowski space. This is true even if the metric is non-stochastic. Now, if we consider the stochastic metric where the stochasticity is due to uncertainty in the angle as mentioned above, the measurement x^1 (i.e., contravariant quantity) is still well defined whereas the covariant one x_1 becomes indeterminate since it is a function of the angle.

The functional or internal space will be considered here to be stochastic in nature. By contrast, the external world is largely non-stochastic and so the issue to be considered is that of the conversion from the non-stochastic space of the external world to the stochastic space of CNS function. This can be done by considering the equations transforming the non-stochastic coordinates to stochastic coordinates as stochastic.

Here, if one assumes the existence of Lagrangian (i.e., difference of kinetic and potential energy of a system) and defines a pair of conjugate variables, one can define an observable quantity as a contravariant variable such as

$$p^j = g^{ji}p_i$$

where p_i a covariant vector and hence not observable. Since g^{ji} is stochastic, p^j is a distribution. One way that this distribution of momentum can be tested is by determining the dynamic parameters of physiological tremor (Llinás, 1991). Asanov et al. (1988) considered space-time not merely as a set of points, x^i, but also as other vectors defined over the tangent to the base manifold containing x^i. Physically, these vectors may be velocity or acceleration. If the metric tensor becomes a function of the position variables x^i, this reduces to the metric known as a

Riemannian metric. Now, if the basic geometric objects posses the property of zero-degree homogeneity with respect to each vector of a sequence (e.g., velocity, acceleration vectors) the internal space is considered to be a homogeneous probabilistic space. Note that the above type of metric, i.e., function of both x and ζ, can be shown to be a special case of a more generalized metric known as a Finsler metric.

The dependence of the metric tensor on extra variables such as velocity and acceleration has been shown to be associated with the fluctuating nature of space-time (Asanov et al., 1988). Again this type of dependence can be interpreted as a manifestation of the existence of some internal curvature. This kind of theory involves dependence only on the directions of vectors but not on their lengths. Many authors have discussed the physiological significance of the direction of a given line or curve element. In fact, what we see is only the direction of the curve elements and the deviation of the direction of one curve-element from that of another. For example, let us consider the physiological significance of the direction of a straight line or curve element. The first derivative, $\mathrm{d}y/\mathrm{d}x$, of the curve element can be determined by its steepness and the qualitative information of the second derivative from the curvature. However, the third order or fourth order derivative cannot be estimated physiologically. Indeed, only the direction of the curve elements, or the direction of one curve element from the other, can be determined. All other higher order derivatives are intellectual affairs. Again, in case of left–right rotation, one needs to compare the direction of one curve element with another. The above-mentioned Finsler metric involves only the directions and not the comparison of the direction of one curve element with the other. So the directionality in the sense of left–right rotation cannot be specified as a property definable on Riemannian or non-Riemannian terms alone. In functional geometry, the metric tensor will be a function of an extra variable in the probabilistic sense as defined above, i.e.,

$$\begin{aligned}\mathrm{d}s^2 &= h_{ij}(x,\xi)\mathrm{d}x^i\mathrm{d}x^j \\ &= \mathrm{d}x^2 - (\xi_2 - \xi_1)^2\end{aligned}$$

where $\mathrm{d}x^2$ is the four-dimensional line element in Minkowski space and $(\xi_2 - \xi_1)^2$ the contribution from the fluctuating variables.

This is a five-dimensional fluctuation space. Now, we can also think of a four-dimensional fluctuating space where the fluctuation of space-time arises due to the chaotic three-dimensional oscillation of neurons. This is similar to the idea put forward by Takano (1967) where he considered the fluctuation of four-dimensional space-time arises due to the three dimensional harmonic oscillations. In this framework,

$$(\xi_2 - \xi_1)^2 = \rho_1^2 + \rho_2^2 - \rho_1\rho_2\cos\theta$$

where ρ_1 and ρ_2 are the radii and θ is the angle between them.

Using the above metric tensor we may associate with each arbitrary contravariant vector ξ^i of tangent space $T_n(P)$ at a point P a covariant vector η_j defined by the relation

$$\xi^i = g^{ij}(x,\xi)\eta_j$$

provided

$$\det\left|\frac{\partial}{\partial\xi^j}(g_{ik}(x,\xi))\xi^k\right| \neq 0$$

It is worth mentioning that the directional argument in g^{ij} must coincide with each contravariant vector, ξ^i of $T_n(P)$. This contravariant vector would then be associated with the motor activity. Again, since the directional argument coincides with this contravariant vector, the degrees of freedom in the internal or functional space are linked to the motor activity. Note that in other forms of contravariant vectorial representations, such as in the dream state, the degree of freedom increases, but the sensory activities do not operate. We shall consider this situation next.

The issue of non-orthogonality between the vectors has been analyzed in the context of probabilistic geometry. One can use the concept of probability in defining the inner product between the vectors.

The dream state and probabilistic geometry

In spite of considerable effort, physiological and behavioral characterization of the wakeful and sleep states, as well as their functional meaning, remain elusive. A systematic study by Llinás and Pare (1991) concluded that

(i) The main difference between wakefulness and paradoxical sleep (characterized by the repeated occurrence of periods of rapid eye movements — from which the alternative designation "REM" sleep was derived) lies in the weight given to sensory afferents in cognitive images.
(ii) Otherwise, wakefulness and paradoxical sleep are fundamentally equivalent brain states probably subserved by an intrinsic thalamo-cortical loop with the fundamental difference that intentionality (frontal lobe activity) is absent. As such, dreams are perceptual streams of intrinsic origin that can be sensed but not modulated.

The above framework of dynamic geometry makes it possible to distinguish one individual from another. The uniqueness of individuals must operate in the detail not in principle, i.e. general solutions must operate on the basis of a generalized solution field.

Considering the stochastic nature of specific brain function amongst individuals, they will have both anatomical and functional variances, similar to the differences in their facial characteristics. The variances must be such that they modulate, rather than destroy, the general solution. However, the variances may be so extreme that they negate function or may even be lethal. Within this framework, social variance enriches the system, as variances augment some properties and diminishes the others and so individuality is important.

In the dream state the degrees of freedom are increased because internal perception is not limited by the senses as during the waking state (Llinás, 1987). The stochastic metric tensor, as considered above, is a function of several variables. During the dream state this tensor will be a function of more variables because the internal degrees of freedom are increased. In addition to the reduction in sensory limitations, the frontal cortex that controls motor execution intentions is also not functional (Hobson, 1988). During the waking state sensory input constraints the internal degrees of freedom of brain activity and the frontal cortex augments this constraint by providing perceptions and intentions. It is also worth noting that dreams obey the rules in the sense of causality. However, the rules will be determined by the internal or dynamic geometry and not by the external reality of the waking state.

In the waking state, the sensory systems are operational in the sense that they can communicate with the internal functional state. Sensory systems do not sleep, but the brain does. So in the sleep state the senses cannot modulate the functional geometry of the brain. This operation can be visualized as the instruments making averages over those extra degrees of freedom in the metric structure in the following sense.

The stochastic metric is, $g_{ij}(x,\xi) = \kappa_{ij}(x) + \chi_{ij}(x,\xi)$, where the second term, $\chi_{ij}(x,\xi)$, is the fluctuating part of the metric, g_{ij}. By adopting suitable averaging procedure one can write $\langle g_{ij}(x,\xi)\rangle = \langle \kappa_{ij(x)}\rangle$.

In the dream state, the contravariant vector arising from sensory inputs vanishes. However, the metric tensor does not vanish since there is an increased of degree of freedom during dreaming where the contravariant vectors arise from spontaneous background activity. Here, the metric tensor is a function of all the variables which are treated as degrees of freedom. In this case

$$x^i = g^{ij}x_j = 0$$

and since

$$x_j = 0$$

then

$$g^{ij} = 0$$

Or non-zero [equation says zero].

In REM sleep

The internal degrees of freedom increase in the dream state. Moreover, in case of functional geometry as $g^{ij} = e^i \cdot e^j$, the metric tensor cannot

be diagonalized since no orthogonal coordinate system exists for the real neuronal situation. So, for REM sleep

$$g^{ij} \neq 0$$

In deep sleep

In case of non-REM (NREM) sleep, the degrees of freedom seem to be frozen. Then we can think of the other solution, i.e.,

$$g^{ij} = 0$$

This happens if we consider the orthogonality of functional states of neurons.

Now, if at a certain level in the hierarchy, when the orthogonality property is satisfied, one can diagonalise the matrix (for the metric tensor) then the trace of the matrix associated with the metric (i.e., sum of the diagonal elements) will be zero. If the orthogonality property exists in the NREM sleep state, then it happens in a hierarchically different functional state than that in the REM sleep state. MEG and EEG recordings indicate that there is a long distance coherence at low frequencies (Llinas and Steriade, 2006) where the rich geometrical granularity encountered in REM or in wakefulness is momentarily gone. Thus, the functional states in REM and deep sleep states are basically different. It appears that the functional states will be orthogonally oriented in deep sleep and in REM sleep and the awake state there will be a deviation from orthogonality. In a sense, the deep sleep state is being modulated in REM sleep as well as in the awake state. 40 Hz oscillations are present in REM sleep as well as in deep sleep states. This rhythmic activity might modulate the functional states of neurons from orthogonal to non-orthogonal states.

Implications

The above analysis clearly indicates that dynamic geometry plays a pivotal role in understanding the external world through the CNS. This internal geometry is sense dependent in contrast to deductive geometry used in modern physics. The weak chaotic nature of the oscillations of single neurons makes the metric of the functional geometry a probabilistic one.

The probabilistic nature of the geometry makes it possible to construct a well-defined mathematical transformation between the outside world and the internal world using tensor network theory. The dream state and non-REM sleep state might have consistent explanations within the framework of dynamic geometry. However, we shall consider this issue in a more elaborate manner in a separate paper.

The concept of dynamic geometry will shed new light on the issue of consciousness and its neuronal correlates.

Acknowledgments

We are indebted to the Marine Biology Laboratory, Woods Hole and to New York University School of Medicine for support to complete this work. One of the authors (SR) is also indebted to college of Science, George Mason University.

References

Amari, S. (1994) Information geometry and manifolds of neural networks. In: Grassberger P. and Nadal J.-P. (Eds.), Statistical Physics to Statistical Inference and Back. Kluwer Academic Publishers, The Netherlands, pp. 113–138.

Aristotle. (2004) On the Soul. Kissinger Publishing, MT, USA.

Asanov, G.S., Ponomarenko, S.P. and Roy, S. (1988) Finslerian multi-dimensionality, associated gauge fields and stochastic space-time. Fortschr. Phys., 36: 679–706.

Bekoff, A. (1981) Studies in developmental neurobiology. In: Maxwell Cowan W. (Ed.), Essays in Honor of Viktor Hamburger. Oxford University Press, New York.

Caianiello, E.R. (1961) Outline of a theory of thought-processes and thinking machines. J. Theor. Biol., 2: 204–235.

Carlton, F. (1976) Stochastic space-time and quantum theory. Phys. Rev. D, 13: 3183–3191.

Donald, P.H. (1990) In: Anderson J.A., et al. (Eds.), Neurocomputing, 2. MIT Press, MA, USA, p. 406.

Graf, W., de Waele, C. and Vidal, P.P. (1992) Skeletal geometry in vertebrates and its relation to the vestibular endorgans. In: Berthoz A., Graf W. and Vidal P.P. (Eds.), The Head-Neck Sensory Motor System. Oxford University Press, New York, pp. 129–134.

Hobson, A.J. (1988) The Dreaming Brain. Basic Books Inc., NY, USA.

Ingraham, R.L. (1962) Stochastic lorentz observers and the divergence in quantum field theory. II Nuvo Cimento, 24: p. 1117.

Joliot, M., Ribary, U. and Llinás, R. (1994) Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. Proc. Natl .Acad. Sci. U.S.A., 91: 11748–11751.

Kazantsev, V.B., Nekorkin, V.I., Makarenko, V.I. and Llinás, R. (2003) Olivo-cerebellar cluster-based universal control system. Proc. Natl. Acad. Sci. U.S.A., 100: 13064–13068.

Kazantsev, V.B., et al. (2004) Self-referential phase reset based on inferior olive oscillator dynamics. Proc. Natl. Acad. Sci. U.S.A., 101: 18183–18188.

Kohonen, T. (2001). Self-Organizing Maps. Springer Series in Information Science, Vol. 30 (3rd extended ed.). Springer, Berlin.

Leznik, E., Makarenko, V. and Llinás, R. (2002) Electrotonically mediated oscillatory patterns in neuronal ensembles: an in vitro voltage-dependent dye imaging study in the inferior olive. J. Neurosci., 22(7): 2804–2815.

Llinás, R. (1987) 'Mindness' as a functional state of the brain. In: Blakemore C. and Greenfield S.A. (Eds.), Mind Waves. Basil Blackwell, Oxford, pp. 339–358.

Llinás, R. (1988) The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. Science, 242: 1654–1664.

Llinás, R. (1991) The noncontinuous nature of movement execution. In: Humphrey D.R. and Freund H.J. (Eds.), Motor Control, Concepts and Issues. Wiley, New York, pp. 273–274.

Llinás, R. (2002) I of the Vortex (Bradford Book). The MIT Press.

Llinás, R. and Pare, D. (1991) Of dreaming and wakefulness. Neuroscience, 44: 521–532.

Llinás, R. and Steriade, M. (2006) Bursting of thalamic neurons and state of vigilance. J. Neurophysiol., 95: 3297–3308.

Mach, E. (1959) The Analysis of Sensations. Dover Publications Inc., NY.

Makarenko, V. and Llinás, R. (1998) Experimentally determined chaotic phase synchronization in a neuronal system. Proc. Natl. Acad. Sci. U.S.A., 95: 15747–15752.

Margenau, H. (1959) In: Churchman C.W. and Rahtoosh P. (Eds.), Measurement, Definition and Theories. Wiley, New York, p. 75.

Maturana, U. (2000) Steps to an Ecology of Mind. University of Chicago Press, Chicago.

McCulloch, W.S. (1951). Why the mind is in the brain. In: Jeffress L.A. (Ed.), Cerebral Mechanisms in Behavior, the Hixon Symposium, Wiley, New York, pp. 42–111.

McCulloch, W.S. and Pitts, W.H. (1943) A logical calculus of the ideas immanent in nervous system. Bull. Math. Biophys., 15: 115–133.

Menger, K. (1945) Theory of relativity and geometry. In: Schilpp P.A. (Ed.), Albert Einstein: Philosopher Scientist. Open Court Publishing Co., p. 459 (3 Rev Ed.).

Menger, K. (1951) Probabilistic geometry. Proc. Natl. Acad. Sci. U.S.A., 37: p. 226.

Pellionisz, A. and Graf, W. (1987) Tensor network model of the three-neuron vestibulo-ocular reflex arc. J. Theor. Neurobiol., 5: 127–151.

Pellionisz, A. and Llinás, R. (1982) Space-time representation in the brain: the cerebellum as a predictive space-time metric tensor. Neuroscience, 7: 2949–2970.

Pellionisz, A. and Llinás, R. (1985) Tensor network theory of the metaorganization of functional geometries in the CNS. Neuroscience, 16: 245–273.

Plato. (1991). The Republic, translated by Allan Bloom, Basic books (2nd ed.).

Prigogine, I. and Kondepudi, D. (1998) Modern Thermodynamics: From Heat Engines to Dissipative Structure. Wiley, Chichester.

Roy, S. (1998) Statistical Geometry and Applications to Microphysics and Cosmology. Kluwer Academic Publishers, The Netherlands.

Takano, Y. (1967) Fluctuation of space-time and elementary particles I & II. Prog. Theor. Phys., 38: 1185–1187.

von Bertalnffy, L. (1967) Robots, Men and Minds. George Braziller, New York.

von Bertalanffy, L. (1968) General System Theory. George Braziller, New York.

Von der Malsburg, C. (1998). Dynamic link architecture. In: The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA.

Von Neumann, J. (1951) The General and Logical Theory of Automata. Remarks on the Foundation of Mathematics. Oxford University Press, England, UK.

Welsh, J.P. and Llinás, R. (1997) Some organizing principles for the control of movement based on olivocerebellar physiology. Prog. Brain Res., 114: 449–461.

Wiener, N. (1948) Cybernetics. Wiley, New York.

Wittgenstein, L. (1997). Philosophische Untersuetingen translated by G.E.M. Anscombe. Blackwell, Malden, MA.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 12

Understanding the mind of a worm: hierarchical network structure underlying nervous system function in *C. elegans*

Nivedita Chatterjee[1,*] and Sitabhra Sinha[2]

[1]*AU-KBC Research Centre, Anna University, Chromepet, Chennai 600044, India*
[2]*Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600113, India*

Abstract: The nervous system of the nematode *C. elegans* provides a unique opportunity to understand how behavior ('mind') emerges from activity in the nervous system ('brain') of an organism. The hermaphrodite worm has only 302 neurons, all of whose connections (synaptic and gap junctional) are known. Recently, many of the functional circuits that make up its behavioral repertoire have begun to be identified. In this paper, we investigate the hierarchical structure of the nervous system through k-core decomposition and find it to be intimately related to the set of all known functional circuits. Our analysis also suggests a vital role for the lateral ganglion in processing information, providing an essential connection between the sensory and motor components of the *C. elegans* nervous system.

Keywords: *C. elegans* neural network; nematode behavior; k-core decomposition; brain-mind; hierarchical network; degree correlation; assortativity

Introduction

> Behavior is the result of a complex and ill-understood set of computations performed by nervous systems and it seems essential to decompose the problem into two: one concerned with the question of the genetic specification of the nervous system, and, the other with *the way nervous systems work to produce behavior*.
>
> Brenner (1974)

As suggested by the above quotation, one of the fundamental problems in brain-mind studies is to understand the process by which electrophysiological activity at the level of the neuronal network gives rise to the complete set of stimulus–response behavior characteristic of a particular organism. Ideally, we would like to understand mental phenomena as a direct consequence of activity in the neurons that make up the brain. However, for the human brain having over 10^9 neurons and 10^{12} synaptic connections amongst them, such an undertaking seems impossible with current technology. Even if it had been possible by some means to record the activity of every neuron, it would be extremely hard to make sense of this enormous quantity of data and thereby understand their relation to various mental states.

*Corresponding author. Tel.: +91-44-22543301; Fax: +91-44-22541586; E-mail: charu5176@yahoo.com

DOI: 10.1016/S0079-6123(07)68012-1

For this reason, it seems more fruitful to focus on a much simpler organism having relatively very few neurons, and yet, which has a complex behavioral repertoire capable of surviving successfully in the complex natural environment. The nematode *Caenorhabditis elegans* (*C. elegans*), so far the only organism whose nervous system has been completely mapped, is the perfect system which satisfies these requirements. The hermaphrodite animal, which is ~1 mm in length, has a nervous system comprising 302 neurons, a third of all the somatic cells in its body. The morphology, location and connectivity of each neuron has been completely described and is almost invariant across different individuals (Ward et al., 1975; Hall and Russell, 1991). Approximately 5000 chemical synapses, 600 gap junctions and 2000 neuromuscular junctions have been identified. Moreover, the nematode displays a rich variety of behavioral patterns, including several forms of non-associative learning that persist over several hours, and there is also indication that it is capable of associative learning (Hobert, 2003). In fact, it has already been used extensively as a model system to study the relationship between behavior and genetics (Brenner, 1974). The easy accessibility of the *C. elegans* nervous system to manipulation, has allowed identification of several reflexes which function as the basis of many aspects of organismal behavior (du Lac et al., 1995).

The fact that only ~300 neurons seem to be enough for an organism to survive in the wild has been a particular challenge to scientists involved in modeling the brain/mind, who have struggled to simulate individual aspects of mental activity, e.g., memory, using many thousands of model neurons (Hertz et al., 1991). It seems there is little hope of understanding how the much more complicated human brain works, until we can explain the behavior of *C. elegans* in terms of its neural network dynamics. This is especially so because the complexity of behavior of an organism appears to be related to the complexity of its nervous system. Here *behavior* refers to the set of actions or reactions in relation to the environment, allowing adaptation to various external stimuli. While behavior can be conscious or unconscious, overt or covert, voluntary or involuntary, it requires decision-making on the part of the neural circuits involved.

In *C. elegans*, neuronal circuits have been delineated based on patterns of synaptic connectivity derived from ultrastructural analysis. Individual cellular components of these anatomically defined circuits have previously been characterized on the sensory, motor and interneuron levels (Tsalik and Hobert, 2003). In the present work we have chosen eight functional circuits, namely, (a) touch sensitivity, (b) egg laying, (c) thermotaxis, (d) chemosensory, (e) defecation, and, three types of locomotion: when (f) satiated (feeding), (g) hungry (exploration) and (h) during escape behavior (tap withdrawal). Over the last few decades, in order to gain insight into the neuronal mechanisms regulating these reflexes or behaviors, individual neurons have been selectively and systematically ablated by laser microbeam. For example, laser ablation showed that nine classes of sensory neurons and four classes of interneurons are involved in the basic four steps involved in locomotion: forward and backward movements, omega-shaped turns and resting stages (Wakabayashi et al., 2004). The thermotaxis functional circuit was observed to contain relatively few neurons (Mori and Ohshima, 1995; Mori, 1999). On the contrary, the chemosensory functional circuit involves no less than nine pairs. The interneurons in *C. elegans* receive inputs from many modalities and are often multifunctional. But every functional circuit possesses a few dedicated sensory neurons. For example, the chemosensory circuit not only has chemosensory neurons (Bargmann and Horvitz, 1991; Troemel et al., 1995, 1997; Sambongi et al., 1999; Pierce-Shimomura et al., 2001), but also neurons specific for the olfactory component of chemosensation (Bargmann. et al., 1993; L'Etoile and Bargmann, 2000). Motor neurons in the functional circuits may also be very specific. The egg-laying circuit occurring only in the hermaphrodite animal has specialized motor neurons (Horvitz et al., 1982; White et al., 1986; Desai et al., 1988) some of which direct their synaptic output exclusively to vulval muscles and other motor neurons (Waggoner et al., 1998). With connectivities and composition (Chalfie et al., 1985; White et al., 1986) of several functional circuits identified, the next step is to integrate them to analyze for any patterns that might be emerging.

In this paper we employ a core decomposition method to reveal the fundamental structure underlying the connectivity profile of the *C. elegans* neural network. By using a process that peels away successive layers leaving behind the core of the network, we investigate whether there is correlation between a neuron (a) having a critical functional role and (b) occupying a central position in terms of structure. Our results indicate that there is indeed a structural basis behind the roles played by neurons in the functional circuits.

Materials and methods

Connectivity data

The *C. elegans* nervous system is naturally divided into two parts: the pharyngeal system composed of 20 cells (Albertson and Thomson, 1976) controlling the rhythmic contraction of the pharynx during feeding, which is almost completely isolated from the somatic system consisting of the remaining 282 neurons. In this paper, we focus on the 280 connected neurons of the latter system (2 of the neurons being not connected to any other neurons). The connections among these neurons, both synaptic and gap junctional, have been obtained through reconstructions of electron micrographs of serial sections (White et al., 1986). Further, this data has been collated together and made available in an electronic format (Achacoso and Yamamoto, 1992). In the data set, the neurons have been arranged into 10 ganglia, a classification based on physical proximity of the cell bodies to each other. The actual locations of these ganglia are shown in Fig. 1. In addition, the neurons are specified according to their function type, i.e., sensory, motor, interneuron or combinations thereof.

k-core decomposition

Core decomposition, introduced by Seidman (1983), is a technique to obtain the fundamental structural organization of a complex network through a process of successive pruning. The technique has been used to show core–periphery organization in a large number of biological networks (see Holme, 2005; Wuchty and Almaas, 2005). The *k-core* of a network is defined as the

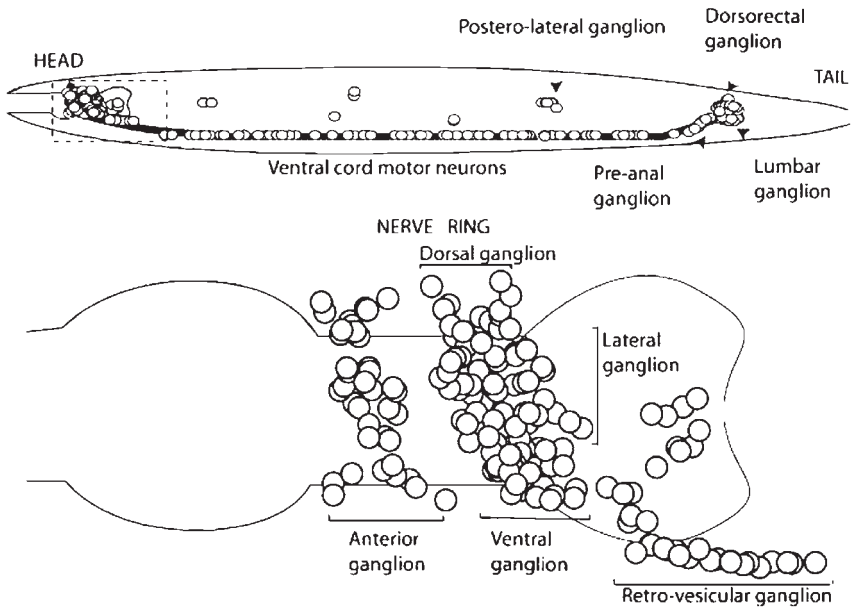


Fig. 1. Diagram indicating the locations of various neural ganglia in *C. elegans*. The bottom figure is a magnified view of the region in the head enclosed with the broken lines in the top figure. (Adapted from Ahn et al., 2006.)

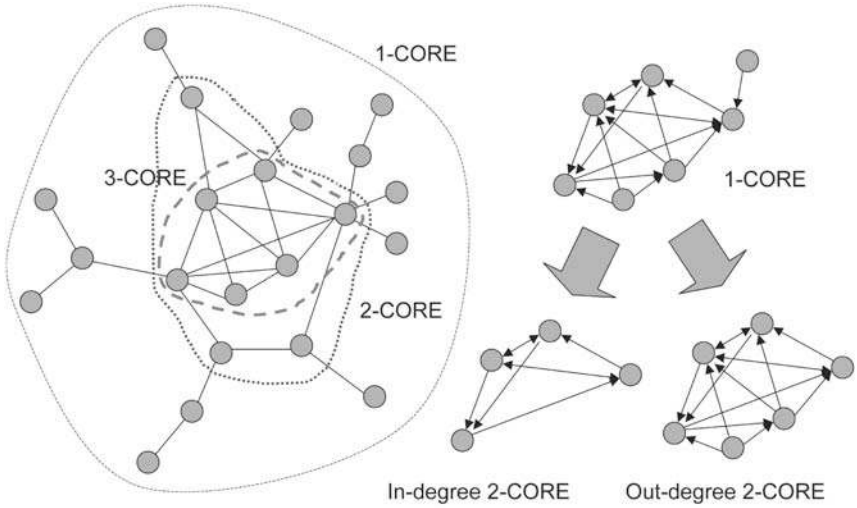


Fig. 2. Schematic diagram showing the k-core decomposition of (left) a undirected network and (right) a directed network, with arrows indicating the direction of connections. For the undirected network, the 1-core is made up of all nodes within the thin broken curve, while the 2-core and 3-core nodes are bounded by the dotted curve and thick broken curve, respectively. For the directed network, the k-core obtained depends on whether one is looking at the in-degree (inward links to a node) or out-degree (outward links from a node).

subnetwork containing all nodes that have degree[1] *at least* equal to *k*. An iterative procedure for determining the k-core is (i) to remove all nodes having degree less than k, (ii) check the resulting network to see if any of the remaining nodes now have degree less than k as a result of (i), and if so (iii) repeat steps (i)–(ii) until all remaining nodes have degree at least equal to k (Fig. 2, left). This resulting network is the k-core of the original network. In particular, the 2-core of a network is obtained by eliminating all nodes that do not form part of a loop (a closed path through a subset of the connected nodes). In fact, there exist at least k paths between any pair of nodes belonging to a k-core.

The procedure indicated above is for an undirected network, i.e., where the links do not have any directionality. However, a synaptic link between two neurons has an inherent direction, and therefore we need to define k-cores for directed networks. By focusing exclusively on either the *in-degree* (number of connections arriving at a neuron from other neurons) or the *out-degree* (number of connections from a neuron to other neurons), one can define k-cores for a directed network. Not surprisingly, one can arrive at different cores for the same network, depending on whether one used the in-degree or out-degree for recursive pruning of the network (Fig. 2, right). It is worth noting here that for a general network, it is possible that the inner k-cores may consist of disconnected parts. However, for *C. elegans*, k-cores for the networks defined in terms of both directed and undirected synaptic connections, as well as for the network defined in terms of gap junctions, remain connected at all orders of k for which the core exists.

Pair-wise degree correlation

While degree is a property associated with a single neuron, one can also look at the relation between degrees of a connected pair of neurons. In particular, one can ask if high-degree neurons connect preferentially to other high-degree neurons, or whether instead, they prefer connecting to low-degree neurons. These two possibilities result in two rather different kinds of network structure, assortative and disassortative, respectively, with

[1]The *degree* of a node (neuron) is the total number of its links or connections.

most biological networks seeming to be of the latter kind (Newman, 2002). In this paper, we use Pearson's correlation coefficient between the degrees of all pairs of connected nodes as a measure of the pair-wise degree correlation. For directed synaptic networks, the correlation coefficient can be measured in four different ways, as one can focus on either the in-degree or the out-degree of the pre- and post-synaptic neurons. Therefore, one can define correlations among (i) pre-synaptic in-degree and post-synaptic out-degree, (ii) pre-synaptic out-degree and post-synaptic in-degree, (iii) pre-synaptic in-degree and post-synaptic in-degree and (iv) pre-synaptic out-degree and post-synaptic out-degree. Each of these measures has distinct functional implications for the neural group concerned. For example, high positive values for (ii) will mark neuronal groups that are primarily involved in carrying information from sensory to motor neurons in the nervous system.

Results

Our approach to understand the functional role of the specific patterns in connectivity among the *C. elegans* neurons is to extract statistically significant structural features among them, i.e., properties that would not be expected to arise in a randomly assembled network. Let us focus on the property of degree, the number of links of a neuron. The degree distribution, i.e., the relative frequency of neurons having various degrees q, sharply decays with q, in a manner that is indistinguishable from a random network. However, by looking at the actual values of in-degree and out-degree of the different neurons, we notice that while sensory neurons have low in-degree and motor neurons have low out-degree, interneurons with high in-degree also tend to have high out-degree (Fig. 3, left). This would not have been expected had the connections among them been made at random. Moreover, the neurons with high degree also tend to be strongly interconnected, another feature not expected in a random network. In fact, the latter feature suggests the existence of a "core group" of neurons, a notion that we shall explore in detail below.

Examining the matrix of synaptic connectivities, arranged according to ganglia, reveals that most connections occur between neurons belonging to the same ganglion, implying a modular structure. Moreover, when the connections between ganglia

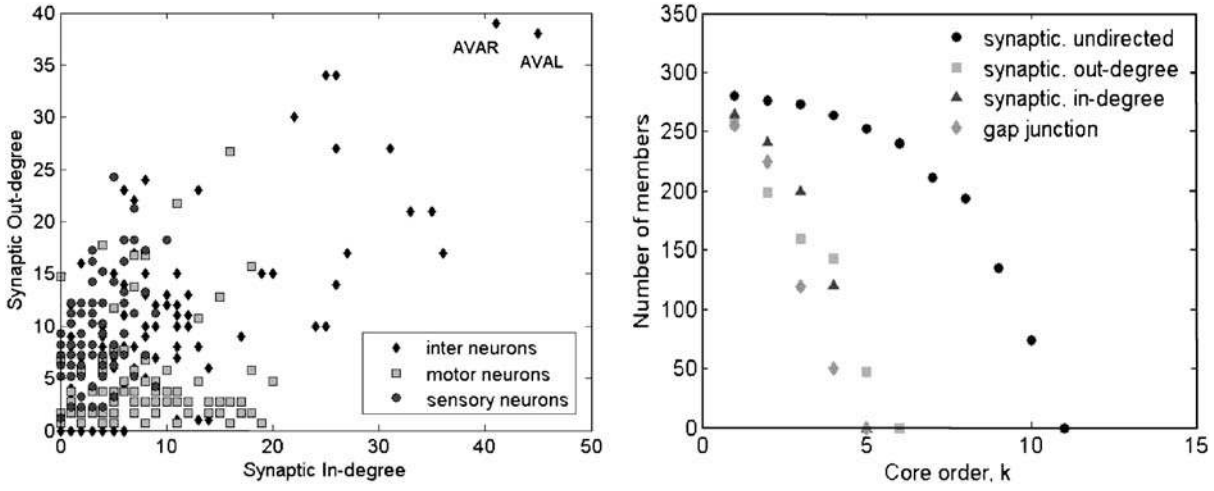


Fig. 3. Left: Relation between synaptic in-degree (number of post-synaptic connections to a neuron) and the synaptic out-degree (number of pre-synaptic connections) for sensory, motor and interneurons. Right: The number of neurons belonging to the k-cores for networks defined according to the type of connection and/or their direction. Except for the undirected synaptic network, the innermost cores are not of very high order.

are examined in detail, we observe that a few receive a significantly higher proportion of the interganglionic connections compared to others. In particular, the lateral ganglion is observed to receive many connections from other ganglia and, in turn, sends out many connections to the ventral cord motor neuron group. This is especially significant as the lateral ganglion hosts the "command" interneurons (White et al., 1986), so-called because they have a prominent role in a large number of functional circuits.

This brings us to the question of whether the *C. elegans* nervous system has a core–periphery structure (i.e., all neurons can be classified as belonging to either a densely connected central core or a sparsely connected periphery), and if so, then what is its functional significance. We shall attempt to answer this question by identifying the neurons belonging to the core group and ascertaining their functional properties, specifically by noting their membership in the different behavioral circuits. Note that, for neural networks there is intuitively a natural division into core and periphery in terms of the function of the neurons, where interneurons that take part in information processing should form the bulk of the core, while the majority of sensory and motor neurons should belong to the periphery.

We use the k-core decomposition technique (explained earlier) to identify the neurons that belong to the inner layers of structural organization for both the synaptic as well as gap-junctional networks. Figure 3 (right) shows how the number of neurons belonging to a k-core decreases with the order of the core, k. It is of interest that for both the directed synaptic networks as well as the undirected gap junctional networks the order of the innermost core is not large, never exceeding five. This is non-trivial, as we show that for the control case of the *undirected* synaptic network (which has no relevance for network function, synapses being essentially directional) the core order can increase to 10.

Investigating the functional types of the neurons making up the inner cores of the synaptic out-degree and in-degree networks, we find that in the former, sensory neurons significantly increase their presence with increasing k, while motor neurons dominate the latter (Fig. 4). This is perhaps not surprising when one recalls that most sensory neurons have out-degree significantly larger than the in-degree, the opposite being true for motor neurons. However, on inspecting the ganglionic membership of the core neurons, we observe over-representation of the lumbar and lateral ganglia neurons in the synaptic out-degree core, whereas in

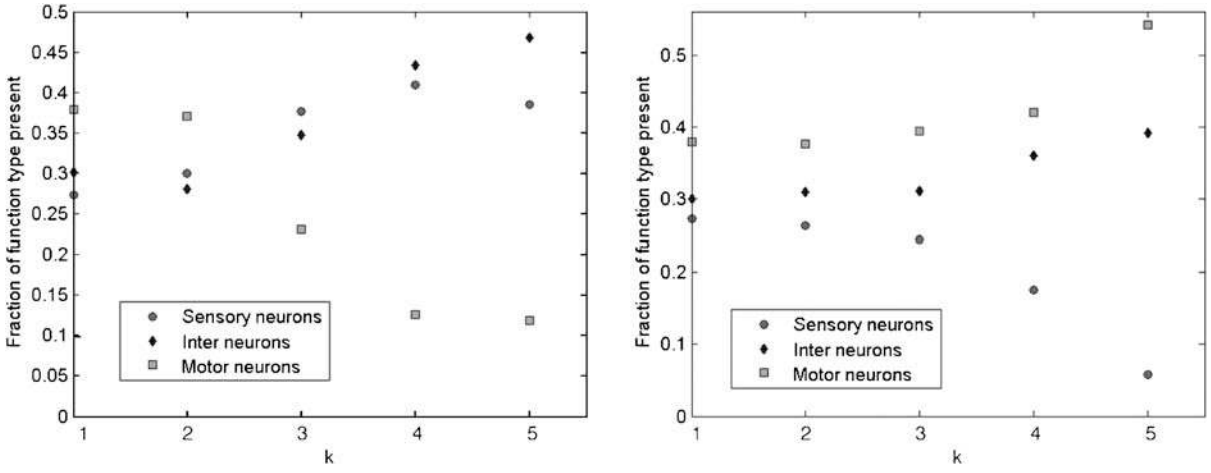


Fig. 4. The fraction of neuron types (sensory, motor or interneuron) in the k-core of the (left) synaptic out-degree network and the (right) synaptic in-degree network. The fraction of motor neurons drops and that of sensory neurons rises as we approach the innermost cores of the out-degree network, while the opposite is true for the in-degree network.

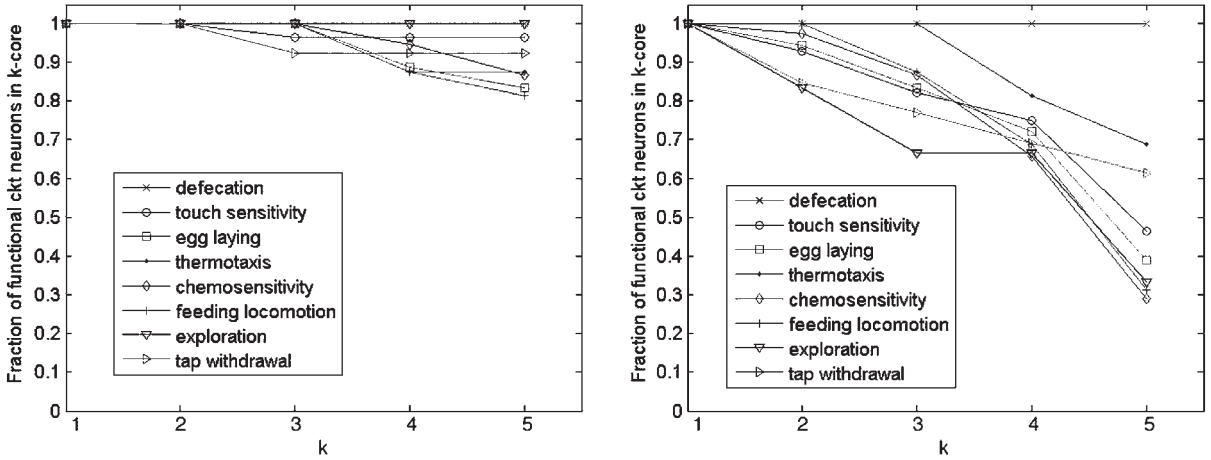


Fig. 5. Fraction of neurons belonging to specific functional circuits in the k-core of the (left) synaptic out-degree and the (right) synaptic in-degree network. Most of the neurons critically involved in various functions are in the innermost cores of the out-degree network.

the in-degree core there is an over-representation of neurons belonging to the lateral ganglion and the ventral cord neural group. This indicates that our earlier observation of the significant position of the lateral ganglion in the synaptic connectivity matrix is not an accident. Indeed, it suggests that the lateral ganglion acts as the ''information processing hub'' of the *C. elegans* nervous system, the principal bridge between its sensory and motor components.

We now turn to examine the role that core neurons play in the organism's behavior. Figure 5 indicates that a large fraction of the neurons involved in the various functional circuits are present in the inner cores of the directed synaptic networks. Moreover, most neurons belonging to the inner core are involved in the different functional circuits. This is perhaps natural because functional circuit membership is determined by eliminating a neuron and observing its behavioral consequence. It stands to reason that eliminating a neuron that belongs to the inner core will have a larger effect than eliminating one in the periphery, and therefore, more likely to result in behavioral anomaly. The identification of the position of a neuron in the k-core hierarchy with its importance in the functional circuits, shows that one can indeed relate structural features of the neural network with behavioral function.

Finally, we explore the pair-wise degree correlation for the gap junctional and directed synaptic network cores. We find that at the lowest order, the gap junction network shows a preference for connections between high-degree and low-degree neurons (i.e., disassortative) but as one goes higher up the hierarchy to the inner cores, they become *less* disassortative. This can possibly be because the gap-junctional network is star-like, with its hub composed of a densely connected group of high-degree neurons, to each of which several low-degree neurons are linked. For the synaptic network, (Fig. 6) we find that the network, even at the lowest order, is assortative, i.e., high-degree neurons show a significant preference for connecting to other high-degree neurons. Moreover, as one proceeds to higher orders of k, this assortative tendency increases for the synaptic in-degree network cores (comprising mostly motor and interneurons), while it decreases for the synaptic out-degree network cores (consisting of mostly sensory and interneurons). This suggests that the group of sensory and interneurons may have a different core structure (more star-like) than the group of motor and *command* interneurons (more clustered).

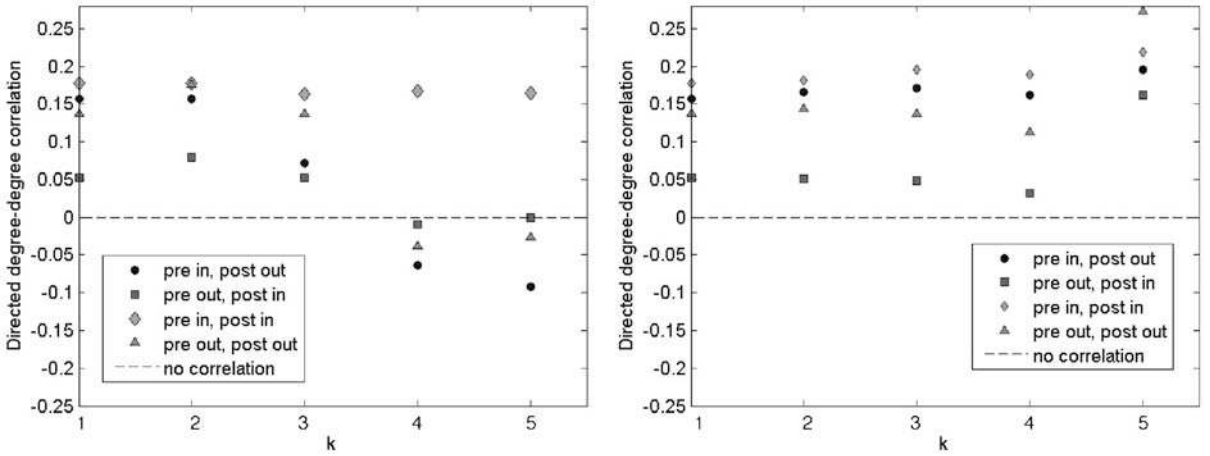


Fig. 6. Pair-wise degree correlation for the synaptic (left) out-degree and the (right) in-degree network cores, defined according to the direction of connections in the pre-synaptic and post-synaptic neurons.

Conclusions

In this paper we have used the k-core decomposition technique to analyze the hierarchical structure of the *C. elegans* nervous system. Our results point towards a key role played by neurons belonging to the lateral ganglion in processing information traveling through the stimulus–response path between the sensory and motor neurons. Comparison of the neurons belonging to the inner cores as defined by network structure with the neurons belonging to the different functional circuits as indicated by their crucial role in behavior, suggests a strong correlation between the two. Almost all neurons identified as belonging to any functional circuit are present in the inner cores. This suggests an intriguing relation between the structural centrality and functional importance of a neuron. In addition, we obtain a glimpse of the possibly different structural principles used in connecting the sensory–interneuron and the motor–interneuron components of the nervous system by investigating the pair-wise degree correlation along the core order hierarchy. The occurrence of assortativity in a biological neural network, in contrast to most other biological networks which are disassortative, is especially intriguing. It may indicate that the nervous system had to face significantly different constraints in its evolutionary path compared to other biological circuits. This may shed light on one of the central questions in evolutionary biology that resonates strongly with the theme of this volume, namely, why did brains or central nervous systems evolve? An alternative could have been a nervous system composed of a set of semi-independent reflex arcs. Structurally, this would have been manifested as a series of parallel pathways that process information independently of each other, rather than the densely connected networks that we are familiar with. This question becomes even more significant in light of the argument that the larger complexity inherent in densely connected networks has led to the emergence of a conscious mind from the simple stimulus–response processing capability of primitive organisms.

References

Achacoso, T.B. and Yamamoto, W.S. (1992) AY's Neuroanatomy of *C. elegans* for Computation. CRC Press, Boca Raton, FL.

Ahn, Y.Y., Jeong, H. and Kim, B.J. (2006) Wiring cost in the organization of a biological neuronal network. Phys. A, 367: 531–537.

Albertson, D.G. and Thomson, J.N. (1976) The pharynx of *Caenorhabditis elegans*. Philos. Trans. R. Soc. Lond. B, 275: 299–325.

Bargmann, C.I., Hartwieg, E. and Horvitz, H.R. (1993) Odorant-selective genes and neurons mediate olfaction in *C. elegans*. Cell, 74: 515–527.

Bargmann, C.I. and Horvitz, H.R. (1991) Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in *C. elegans*. Neuron, 7: 729–742.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. Genetics, 77: 71–94.

Chalfie, M., Sulston, J.E., White, J.G., Southgate, E., Thomson, J.N. and Brenner, S. (1985) The neural circuit for touch sensitivity in *Caenorhabditis elegans*. J. Neurosci., 5: 956–964.

Desai, C., Garriga, G., McIntire, S.L. and Horvitz, H.R. (1988) A genetic pathway for the development of the *Caenorhabditis elegans* HSN motor neurons. Nature, 336: 638–646.

L'Etoile, N.D. and Bargmann, C.I. (2000) Olfaction and odor discrimination are mediated by the *C. elegans* guanylyl cyclase ODR-1. Neuron, 25: 575–586.

Hall, D.H. and Russell, R.L. (1991) The posterior nervous system of the nematode *Caenorhabditis elegans*: serial reconstruction of identified neurons and complete pattern of synaptic interactions. J. Neurosci., 11: 1–22.

Hertz, J., Krogh, A. and Palmer, R.G. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA.

Hobert, O. (2003) Behavioral plasticity in *C. elegans*: paradigms, circuits, genes. J. Neurobiol., 54: 203–223.

Holme, P. (2005) Core–periphery organization of complex networks. Phys. Rev. E, 72: p. 046111.

Horvitz, H.R., Chalfie, M., Trent, C., Sulston, J.E. and Evans, P.D. (1982) Serotonin and octopamine in the nematode *Caenorhabditis elegans*. Science, 216: 1012–1014.

du Lac, S., Raymond, J.L., Sejnowski, T.J. and Lisberger, S.G. (1995) Learning and memory in the vestibulo-ocular reflex. Ann. Rev. Neurosci., 18: 409–441.

Mori, I. (1999) Genetics of chemotaxis and thermotaxis in the nematode *Caenorhabditis elegans*. Ann. Rev. Genet., 33: 399–422.

Mori, I. and Ohshima, Y. (1995) Neural regulation of thermotaxis in *Caenorhabditis elegans*. Nature, 376: 344–348.

Newman, M.E.J. (2002) Assortative mixing in networks. Phys. Rev. Lett., 89: p. 208701.

Pierce-Shimomura, J.T., Faumont, S., Gaston, M.R., Pearson, B.J. and Lockery, S.R. (2001) The homeobox gene lim-6 is required for distinct chemosensory representations in *C. elegans*. Nature, 410: 694–698.

Sambongi, Y., Nagae, T., Liu, Y., Yoshimizu, T., Takeda, K., Wada, Y. and Futai, M. (1999) Sensing of cadmium and copper ions by externally exposed ADL, ASE, and ASH neurons elicits avoidance response in *Caenorhabditis elegans*. Neuroreport, 10: 753–757.

Seidman, S.B. (1983) Network structure and minimum degree. Soc. Netw., 5: 269–287.

Troemel, E.R., Chou, J.H., Dwyer, N.D., Colbert, H.A. and Bargmann, C.I. (1995) Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. Cell, 83: 207–218.

Troemel, E.R., Kimmel, B.E. and Bargmann, C.I. (1997) Reprogramming chemotaxis responses: sensory neurons define olfactory preferences in *C. elegans*. Cell, 91: 161–169.

Tsalik, E.L. and Hobert, O. (2003) Functional mapping of neurons that control locomotory behavior in *Caenorhabditis elegans*. J. Neurobiol., 56: 178–197.

Waggoner, L.E., Zhou, G.T., Schafer, R.W. and Schafer, W.R. (1998) Control of alternative behavioral states by serotonin in *Caenorhabditis elegans*. Neuron, 21: 203–214.

Wakabayashi, T., Kitagawa, I. and Shingai, R. (2004) Neurons regulating the duration of forward locomotion in *Caenorhabditis elegans*. Neurosci. Res., 50: 103–111.

Ward, S., Thomson, N., White, J.G. and Brenner, S. (1975) Electron microscopical reconstruction of the anterior sensory anatomy of the nematode *Caenorhabditis elegans*? J. Comp. Neurol., 160: 313–337.

White, J.G., Southgate, E., Thomson, J.N. and Brenner, S. (1986) The structure of the nervous system of the nematode *C. elegans*. Philos. Trans. R. Soc. Lond. B, 314: 1–340.

Wuchty, S. and Almaas, E. (2005) Peeling the yeast protein network. Proteomics, 5: 444–449.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 13

Neural network modeling

Bikas K. Chakrabarti[1,2,*] and Abhik Basu[2]

[1]*Centre for Applied Mathematics and Computational Science, Saha Institute of Nuclear Physics, Calcutta 700064, India*
[2]*Theoretical Condensed Matter Physics Division, Saha Institute of Nuclear Physics, Calcutta 700064, India*

Abstract: Some of the (comparatively older) numerical results on neural network models obtained by our group are reviewed. These models incorporate synaptic connections constructed by using the Hebb's rule. The dynamics is determined by the internal field which has a weighted contribution from the time delayed signals. Studies on relaxation and the growth of correlations in the Hopfield model are discussed here. The memory capacity of such networks have been investigated also for some asymmetric synaptic interactions. In some cases both the asynchronous (or Glauber; Hopfield) and synchronous (Little) dynamics are used. At the end, in the appendix, we discuss the effects of asymmetric interactions on the statistical properties in a related model of spin glass (new results).

Keywords: neural network models; spin glasses; associative memory; Hopfield model; Little model; retrieval dynamics; relaxation

Introduction

The human brain is formed out of an interconnected network of roughly 10^{10}–10^{12} relatively simple computing elements called neurons. Each neuron, although a complex electro-chemical device, performs simple computational tasks of comparing and summing incoming electrical signals from other neurons through the synaptic connections. Yet the emerging features of this interconnected network are surprising and understanding the cognition and computing ability of the human brain is certainly an intriguing problem (see Amit, 1989; Hertz et al., 1991). Although the real details of the working of a neuron can be quite complicated, for a physics model, following McCullough and Pitts (see Amit, 1989), neurons can be taken as two-state devices (with firing state and nonfiring states). Such two-state neurons are interconnected via synaptic junctions where, as pointed out first by Hebb (see Amit, 1989), learning takes place as the pulse travels through the synaptic junctions (its phase and/or strength may get changed).

Present day super-computers employ about 10^8 transistors, each connected to about 2–4 other transistors. The human brain, as mentioned before, is made up of $\sim 10^{10}$ neurons and each neuron is connected to $\sim 10^4$ other neurons. The brain is thus a much more densely interconnected network, and although it uses much slower operating elements (the typical time scale is of the order of milliseconds for neurons) compared to silicon devices (nanosecond order), many of the computations performed by the brain, such as in perceiving an object, are remarkably faster and suggest an altogether different architecture and approach.

*Corresponding author. Tel.: +91 33 2321 4869;
Fax: +91 33 2337 4637;
E-mail: bikask.Chakrabarti@saha.ac.in

DOI: 10.1016/S0079-6123(07)68013-3

Two very important features of neural computations by the brain which largely outperforms present day computers (though they perform arithmetic and logical operations with ever increasing speeds) are:

- Associative memory or retrieval from partial information.
- Pattern or rhythm recognition and inductive inference capability.

By associative memory, we mean that the brain can recall a pattern or sequence from partial or incomplete information (partially erased pattern). In terms of the language of dynamics of any network it means that by learning various patterns the neural network forms corresponding (distinct) attractors in the configuration space, and if the partial information is within the basin of attraction of the pattern then the dynamics of the network takes it to the (local) fixed point or attractor corresponding to that pattern. The network is then said to have recognized the pattern. A look at neural networks in this way suggests that the learning of a large (macroscopic) number of patterns in such a network means creation of a large number of attractors or fixed points of dynamics corresponding to the various patterns (uncorrelated by any symmetry operation) to be learned. The pattern or rhythm recognition capability of the brain is responsible for inductive knowledge. By looking at a part of a sequence, say a periodic function in time, the brain can extrapolate or predict (by adaptively changing the synaptic connections) the next few steps or the rest of the pattern. By improved training, expert systems can comprehend more complicated (quasi-periodic or quasi-chaotic) patterns or time sequences (e.g., in medical diagnosis). In the following sections we give a brief (biased) review of our studies on associative memory, time series prediction and optimization problems. In this manuscript we discuss asynchronous (Hopfield) and in some cases synchronous (Little) dynamics also.

There have been a fair number of studies indicating that the recall performance of a neural network model increases if the synaptic strength has asymmetry in inter-neuron connectivity. In view of this, we briefly discuss an associated spin glass model with asymmetric interactions in the appendix. We show that such a model lacks the Fluctuation–Dissipation Theorem (FDT) and consequently it is very difficult to analyze the properties of the model.

Hopfield model of associative memory

In the Hopfield model, a neuron i is represented by a two-state Ising spin at that site (i). The synaptic connections are represented by spin–spin interaction and they are taken to be symmetric. This symmetry of synaptic connections allows one to define an energy function. Synaptic connections are constructed following Hebb's rule of learning, which says that for p patterns the synaptic strength for the pair (i,j) is

$$J_{ij} = \frac{1}{N}\sum_{\mu}^{p} \zeta_i^{\mu}\zeta_j^{\mu} \tag{1}$$

where $\zeta_i^{\mu}, i = 1, 2, \ldots, N$, denotes the μ-th pattern learned ($\mu = 1, 2, \ldots, p$). Each can take values ± 1. The parameter N is the total number of neurons each connected in a one to all manner and p is the number of patterns to be learned. For a system of N neurons, each with two states, the energy function is

$$H = -\sum_{i>j}^{N} J_{ij}\sigma_i\sigma_j \tag{2}$$

The expectation is that, with J_{ij}'s constructed as in (1), the Hamiltonian or the energy function (2) will ensure that any arbitrary pattern will have higher energy than those for the patterns to be learned; they will correspond to the (local) minima in the (free) energy landscape. Any pattern then evolves following the dynamics

$$\sigma_i(t+1) = \mathrm{sgn}(h_i(t)) \tag{3}$$

where $h_i(t)$ is the internal field on the neuron i, given by

$$h_i = \sum_{j} J_{ij}\sigma_j(t) \tag{4}$$

Here, a fixed point of dynamics or attractor is guaranteed; i.e., after a certain (finite) number of iterations t^*, the network stabilizes and $\sigma_i(t^*+1) = \sigma_i(t^*)$. Detailed analytical as well as numerical studies (see Amit, 1989; Hertz et al., 1991) shows that the local minima for H in (2) indeed correspond to the patterns fed to be learned in the limit when memory loading factor $\alpha\ (=p/N)$ tends to zero; and they are less than $\sim 3\%$ of the patterns fed to be learned when $\alpha < \alpha_c \sim 0.142$. Above this loading, the network goes to a confused state where the local minima in the energy landscape do not have any significant overlap with the patterns fed, to be learned by the network.

Relaxation studies in the Hopfield model

As already mentioned, after learning (i.e., after the construction of J_{ij}'s as per Hebb's rule in (1) is over), any arbitrary (corrupted or distorted) pattern fed to the network evolves according to the dynamics given in (3). The average of the iteration numbers required to reach the fixed points may be defined as relaxation time τ (i.e., averaged over all the patterns) for the "distorted" patterns considered. If one starts with a pattern obtained by distortion of any of the learned patterns by a fixed amount of corruption and compares their relaxation time (to reach the corresponding fixed point), one expects τ to increase as α increases (from zero) to $\alpha = \alpha_c$ (for fixed N); the so called critical slowing down. In a recent study (Ghosh et al., 1990; see also Chakrabarti and Dasgupta, 1992), systems ranging in size from $N = 1000$ to $N = 16{,}000$ neurons were investigated numerically, in the absence of noise (zero temperature). For a fixed but small initial corruption, the variation of the average convergence or relaxation time (τ) with storage capacity was found as $\tau \sim \exp[-A(\alpha - \alpha_c)]$ where A is a constant dependent on system size N (Ghosh et al., 1990). This is illustrated in Fig. 1.

It may be noted that this critical slowing down of the recall process near α_c is somewhat unusual

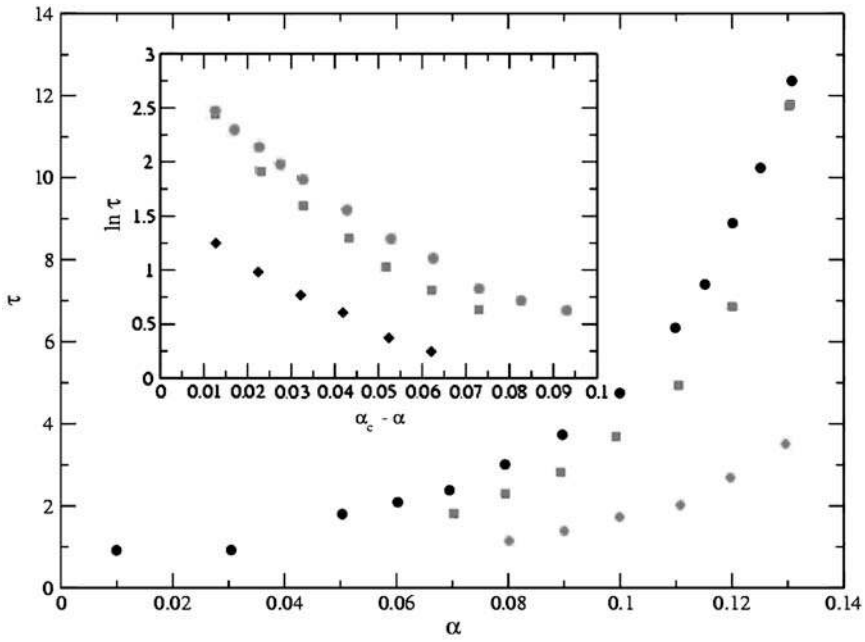


Fig. 1. Plot of average convergence time τ as a function of the initial loading factor α for initial overlap $m(0) = 0.80$ (circles) and $m(0) = 0.95$ (squares) at $N = 16{,}000$ and $m(0) = 0.95$ (diamonds) at $N = 1000$. The inset shows how τ variations fit with the form $\tau \sim \exp[-A(\alpha - \alpha_c)]$ (adapted from Ghosh et al., 1990). (See Color Plate 13.1 in color plate section.)

compared to such phenomena near the phase transition points. Below α_c, the local minima or the metastable states act as the overlap or memory states (not those actual minima, which are spin-glass states) and these metastable states (with overlap) disappear at $\alpha > \alpha_c$. This transition at α_c is therefore not comparable to a phase transition, where the states corresponding to (free) energy minima change. Here, the minimum energy states remain as the spin-glass states for any $\alpha > 0.05$ (see Hertz et al., 1991); only the overlap metastable states disappear at $\alpha \geqslant 0.14$.

A logical extension is to study the variation of τ with synaptic noise. After the learning stage, the patterns were again distorted (by changing the spin states at a fixed small fraction of the sites or neurons) and retrieval is governed by the Glauber type dynamics:

$$\sigma_i(t+1) = \mathrm{sgn}(h_i) \text{ with probability } \frac{1}{[1+\exp(-2h_i/T)]}$$
$$= -\mathrm{sgn}(h_i) \text{ otherwise} \qquad (5)$$

Here T is the synaptic noise and h_i is given by (4). Here, of course, instead of looking for the attractors (which exist only for $T=0$), one measures the average overlap (with the learned patterns) that develops in time and looks for its relaxation to some equilibrium value. Again, the relaxation time increases abruptly as the metastable overlap states disappear across the phase boundary in the $\alpha - T$ plane. Indeed, the (dynamical) phase boundary obtained here (Sen and Chakrabarti, 1989, 1992), studying the divergence of relaxation time, compares favorably with that obtained from the static (free energy) study. It was found that the average relaxation time τ grows approximately as $\exp[(T - T_m)^2]$ near the phase boundary at $T_m(\alpha)$, when $\alpha < \alpha_c$.

Growth of correlations in the Hopfield model

The Hopfield model is solvable in the thermodynamic limit. In the exact solution of the model (see Hertz et al., 1991), it is necessary to assume that the learned patterns are completely random and uncorrelated. Models with correlated patterns have been investigated, where average correlation between all the patterns have been kept finite. A different problem in which only two of the patterns are correlated (i.e., have finite overlap), so that average correlation is still irrelevant, has recently been studied (Chakrabarti and Dasgupta, 1992). The aim was to investigate the ability (or lack of it) of the Hopfield model to store the correlation between two (arbitrary) patterns. The restriction of the correlation between two patterns only ensures that the critical memory loading capacity remains unaffected at $\sim 14\%$; i.e., α_c nearly 0.14. The variation of final correlation between learned patterns with the initial correlation between a randomly chosen pair of patterns (denoted by 1 and 2) to be learned was measured. The initial correlations between the patterns are given by $\rho^{\mu\nu} = \Sigma_i \varsigma_i^\mu \varsigma_i^\nu / N$, where $\rho^{\mu\nu} = \rho_0$ for $\mu = 1$ and $\nu = 2$, and $\rho^{\mu\nu} = 0$ for other values of μ and ν. The system is then allowed to evolve from the initial states $\sigma_i^{(1)}(0) = \varsigma_i^1$ and $\sigma_i^{(2)}(0) = \varsigma_i^2$ following the Hopfield dynamics (3) and (4). The correlation between the states $\sigma^{(1)}$ and $\sigma^{(2)}$ is $\rho_f = \Sigma_i \sigma_i^{(1)}(t^*)\sigma_i^{(2)}(t^*)/N$, where $\sigma_i^{(1)}(t^*)$ and $\sigma_i^{(2)}(t^*)$ denote two states after reaching their respective fixed points. In simulation, the system size is taken as $N = 500$; the patterns between which the correlation is introduced are selected randomly and the averaging is done over 25 configurations for each value of ρ_0 and α. When the ratio ρ_f/ρ_0 is plotted against ρ_0 for different loading α it is found that the ratio $\rho_f/\rho_0 \geq 1$; i.e., the Hopfield model retains correlation of a particular pair of patterns; either it keeps the same correlation or enhances it (see Figs. 3 and 4 of Chakrabarti and Dasgupta, 1992). For studying the $\rho_0 \to 0$ limit more accurately, a larger system of $N = 1000$ was taken. The critical value ρ_c of average ρ_0 upto which the ratio remains unity (and above which it starts deviating towards higher values), is measured against different loading capacity α. As α increases, ρ_c decreases continually and as α exceeds a value of the order of 0.05, ρ_c reaches zero. Hence the Hopfield model generates some correlations, however small, for completely uncorrelated patterns, above a certain value of α ($= 0.05$).

Extended memory and faster retrieval with delay dynamics

In the literature, there have been a number of indications that dynamically defined networks, with asymmetric synaptic connections, may have better recall performance because of the suppression of spin-glass-like states. See the appendix for a discussion on a spin glass model with asymmetric interactions. The exactly solvable models (see Crisanti and Sompolinsky, 1987) indeed uses the same Hebb's rule for connection strength but with extreme dilution (inducing asymmetry). This gives better recall properties. There were also indications that addition of some local memory of each neuron, in the sense that the internal field in (3) is determined by the cumulative effect of all the previous states of the neighboring (as indeed is biologically plausible), gives better recall dynamics in some analog network models. In all these cases, no effective energy function can be defined (because of the asymmetry of or of the local memory effect) and the use of statistical physics of spin-glass-like systems is not possible. All the networks are thus defined dynamically. In a recent simulation study it has been shown that a dynamics with a single time step delay with some tunable weight λ (Sen and Chakrabarti, 1992)

$$\sigma_i(t+1) = \mathrm{sgn}\left[\sum_j J_{ij}(\sigma_j(t) + \lambda\sigma_j(t-1))\right] \quad (6)$$

instead of (3) along with (2), improves the performance of the Hopfield-like model (with sequential updating) enormously. Here the same synaptic connections J_{ij} in Eq. (1), obtained by using Hebb's rule, are employed. In particular, for a network with $N = 500$, the simulations indicated that the discontinuous transition for the overlap function at $\alpha_c \sim 0.14$ disappears and the overlap $m_f^\mu = \sum_i \sigma_i(t^*)\zeta_i^\mu$, for μ-th pattern becomes continuous. Accepting recall with final overlap $m \geq 0.9$, the effective threshold loading capacity α_c increases from 0.14 for $\lambda = 0$ to ~ 0.25 for $\lambda \sim 1$. It was also found that average recall time or relaxation time τ (defined before) becomes minimum around $\lambda \sim 1$ at any particular loading capacity α.

The above interesting and somewhat surprising results for this single step time delayed dynamics are from simulations of a rather small size ($N = 500$) of the network. Here we present the results of a detailed numerical study of this dynamics, both with asynchronous (sequential as well as random; Hopfield-like) and synchronous (Little-like) updating. We check our numerical results for a bigger network (for $250 \leqslant N \leqslant 4000$) and study if there is any significant finite size effect. We find that the fixed point memory capacity improves considerably with λ, both with Hopfield and Little dynamics. We present here the results of a detailed numerical study of this dynamics, both with asynchronous (sequential as well as random; Hopfield-like) and synchronous (Little-like) updating. With 90% final overlap ($m_f \geqslant 0.90$) with the learned patterns, the loading capacity increases to almost 0.25 for $\lambda \sim 1$. The relaxation or average recall timer, for any fixed loading capacity α ($\leqslant 0.25$), is also minimum at $\lambda \sim 1$. We could not detect any significant finite size effect (for $250 \leqslant N \leqslant 4000$) and we believe that this result is true in the thermodynamic limit. We also observe that for $\lambda = 1$, both the Little dynamics and Hopfield dynamics give identical results (all the limit cycles of the Little model disappear to give fixed points). The performance of the network for memory storage in limit cycles for negative λ has also been studied for Hopfield and Little dynamics. Here, the negative λ term provides a damping and limit cycles become frequent but the overlap properties deteriorate significantly. We have also studied the performance of such dynamics for randomly diluted networks. In the next section we discuss the significance of these results.

Simulations and numerical results: Hopfield and Little models

We consider a network of N neurons (spins). The synaptic interactions between neurons i and j are found using the Hebb's rule (1) for a set of p-random Monte Carlo generated patterns. After this 'learning' stage, each pattern is corrupted randomly by a fixed amount (typically 10% of the

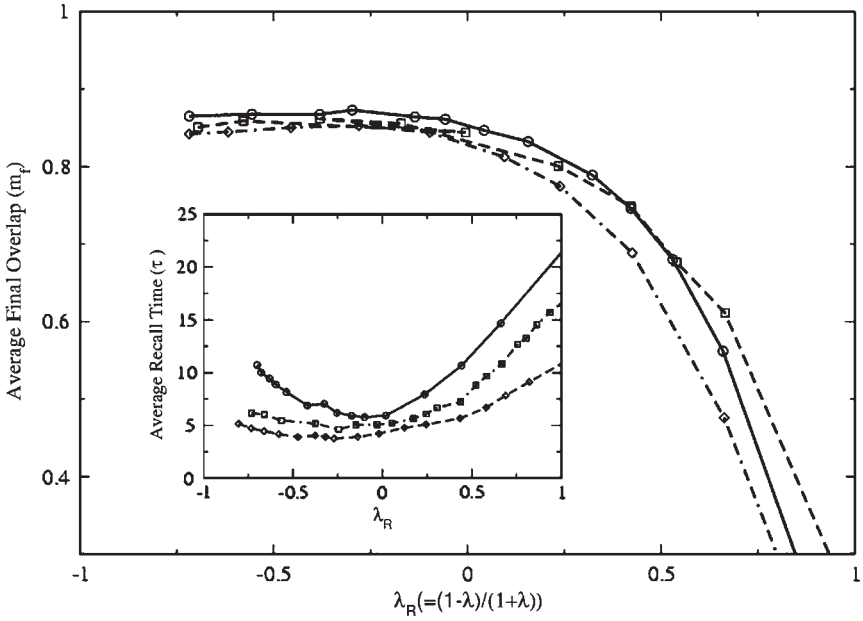


Fig. 2. The variation of the average final overlap m_f (for the fixed point obtained) with λ_R $(=(1-\lambda)(1+\lambda))$, positive λ at a fixed loading capacity $\alpha=0.25$, with sequential dynamics for networks with $N=250$ (solid line), 500 (dashed line) and 1000 (dash-dot line). The inset shows the variation of average recall or relaxation time τ against λ_R for different N (adapted from Maiti et al., 1995).

neuron states are changed randomly; with initial overlap $m_i^\mu = 0.9$ for any pattern μ) and the dynamics (updating process) following (3) starts with these corrupted patterns. Both sequential (Hopfield) and parallel (Little) updating are employed. Since our updating process requires two initial patterns (at $(t-1)$ and $(t-2)$), we start with two randomly distorted patterns with the same value of distortion (for each μ). We study the overlap of the network state with the pattern μ during the updating process and the final overlap m_f^μ of the fixed points (checked for two successive time steps) or limit cycle patterns are noted. The averages over all p patterns (and over two to five sets of random number seed for the entire run) are taken (average of over all p patterns). The number of iterations (t^*) required to reach the fixed points (or limit cycles) are also noted and their average (over the patterns) give the average recall or relaxation time (τ). Our typical results for the Hopfield model (dynamics) are for $N=1000$ and those for the Little model are for $N=500$. We consider mostly positive λ values. Negative λ values do not improve the performance and have been studied just for some typical investigations only.

In Fig. 2, we give the variation of the final average overlap μ (of any pattern) for the fixed point (checked over two successive time steps) with the redefined λ $[\lambda_R=(1-\lambda)(1+\lambda)]$ at fixed loading capacity $\alpha \equiv p/N=0.25$ for sequential as well as random (Glauber) updating. It may be noted that for the Hopfield dynamics ($\lambda=0$) the final overlap $m_f<0.30$ for $\alpha=0.25$. Strictly speaking, this value should be zero in the thermodynamic limit. The results shown in Fig. 2 are for $N=250$, 500 and 1000, and the nonvanishing value of m_f for $\lambda=0$ $(\lambda_R=1)$ at $\alpha=0.25$ is precisely due to this finite size effect as can be easily checked (decreases considerably as N becomes large). It is seen that the final overlap gradually improves with increasing λ and it saturates beyond $\lambda=1$ $(\lambda_R=0)$ to $m_j \approx 0.90$ (for $\alpha=0.25$). The inset figure shows that the average recall time

(or relaxation time) τ and the fraction of limit cycles also decreases as λ increases from $\lambda = 0$ and finally attains an optimal value at $\lambda = 1$. Recall that time τ again shows a minimum at $\lambda = 1$ as in the Glauber (Hopfield) case. An interesting observation is that at $\lambda = 1$ the limit cycles of the Little model (parallel dynamics) disappear and all reach to fixed points with identical overlap properties as in the case of the Hopfield model.

For sequential updating, the variation of the average final overlap m_f with the loading capacity α, for some typical positive values of λ are shown in Fig. 3. The results are for $N = 500$. The inset here shows variation of average recall time (τ) with α for different λ. For negative values of λ even for sequential updating, one gets limit cycles (e.g., the fraction of limit cycles is ~ 0.48, for $\lambda = -0.8$ at $\alpha = 0.25$) and the fixed point fraction decreases further for parallel (Little) dynamics. However, the overlap with the learned patterns decreases considerably and we do not see any significant memory capability (for storage in limit cycles) for negative values of λ (typically $m_f \cong 0.42$ for $\lambda = 0.8$ at $\alpha = 0.25$, using sequential updating) (Fig. 4).

In order to investigate the finite size effect in the improved performance of the network for positive λ values (for sequential dynamics), we studied the variation of average final overlap with $1/N$ at fixed loading capacity α ($= 0.20$ and 0.25) and fixed λ ($= 1$). These results, for $250 \leqslant N \leqslant 4000$, are shown in Fig. 5. The inset there shows the same variation of with $1/N$ at fixed α ($= 0.25$) for $\lambda = 0$. One can clearly compare the finite size effects observed: for $\lambda = 1$, no significant variation of with $1/N$ is observed for both $\alpha = 0.20$ and 0.25 ($m_f \cong 0.4$ for $\alpha = 0.20$, and $m_f \cong 0.8$ for $\alpha = 0.25$ at $\lambda = 1$ as $N \to \infty$). For $\lambda = 0$, however there is significant finite size effect. From the extrapolation we find the extrapolated result for m_f (for large N) to be around 0.29 here at this α ($= 0.25$); see the inset. See Fig. 5 for the variations of m_f with the loading α for various values of λ.

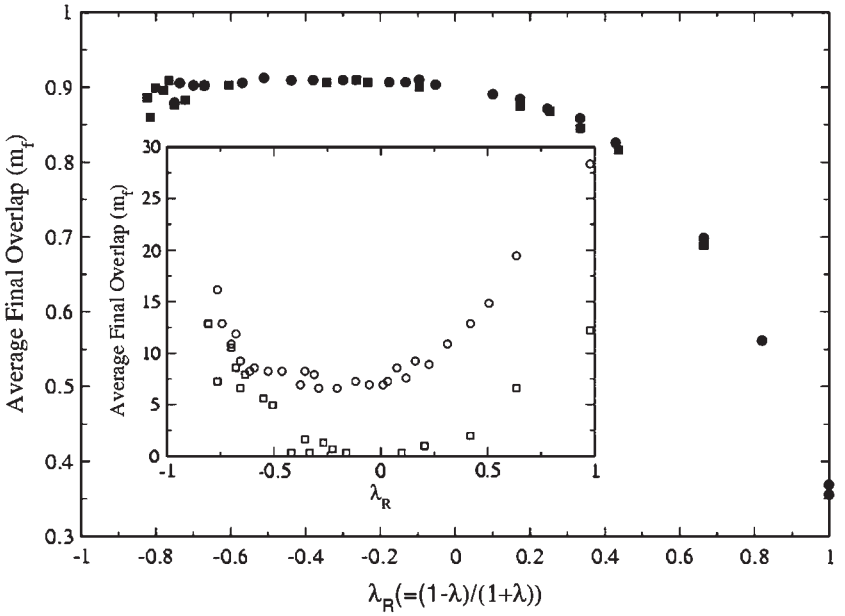


Fig. 3. Average final overlap m_f (for the fixed point obtained) at a fixed loading capacity $\alpha = 0.25$, for a network with $N = 500$ against λ_R ($= (1-\lambda)(1+\lambda)$); positive λ). The results are for parallel updating (Little model), shown by square (filled or unfilled); for the Glauber dynamics (Hopfield model) by circle (filled or unfilled). The inset shows the variation of average recall time τ and the fraction of limit cycles against λ_R (adapted from Maiti et al., 1995).

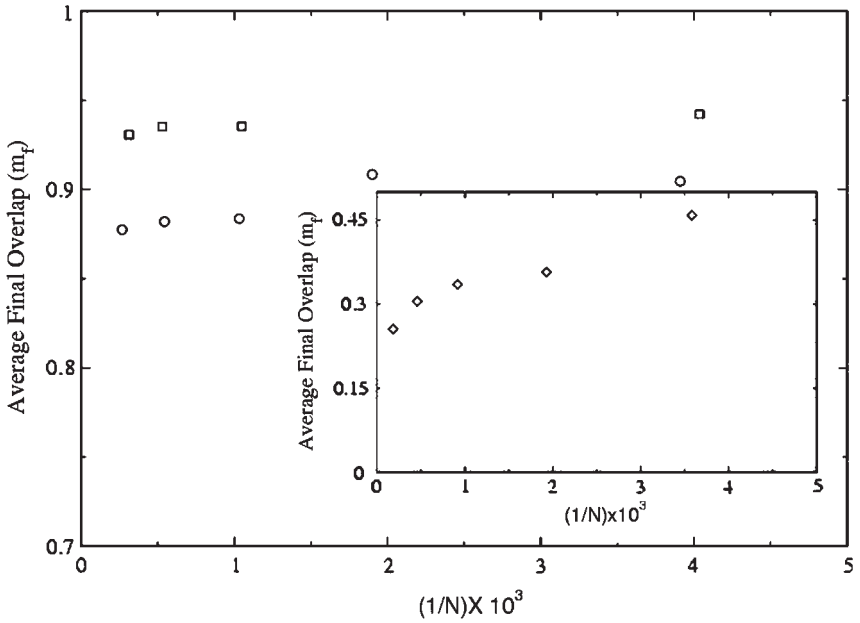


Fig. 4. Variation of average final overlap m_f against $1/N$ for fixed leading capacity $\alpha = 0.2$ and $\alpha = 0.25$ for $\lambda = 1$. The inset shows the same for $\alpha = 0.25$ at $\lambda = 0$. The results are for $250 \leqslant N \leqslant 4000$ (adapted from Maiti et al., 1995).

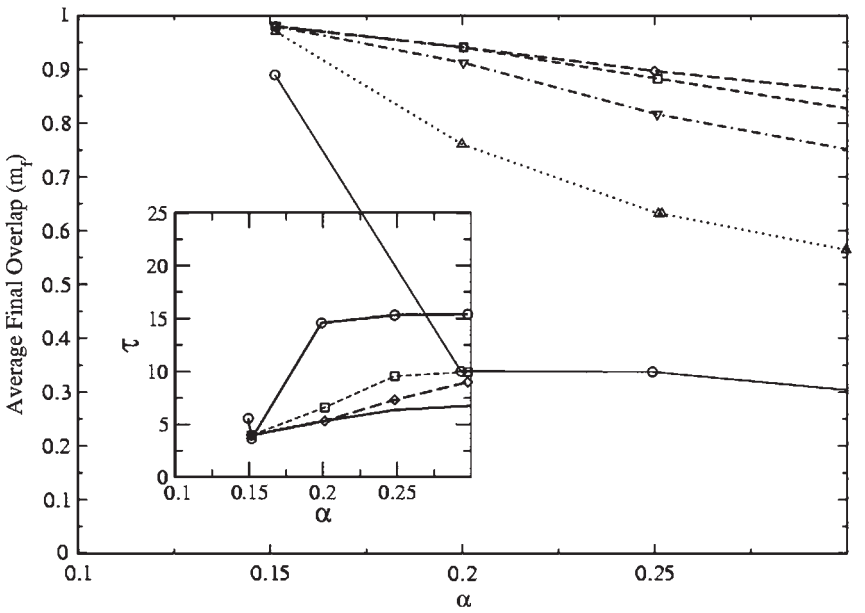


Fig. 5. Variation of average final overlap m_f against loading capacity for some typical values of λ ($\lambda = 0$ for the bottom curve and $\lambda = 0.2$, 0.5, 1.0, 1.5, 2 successively upwards, with sequential dynamics for $N = 1000$). The inset shows the variation of the recall time τ against α for the same values of λ.

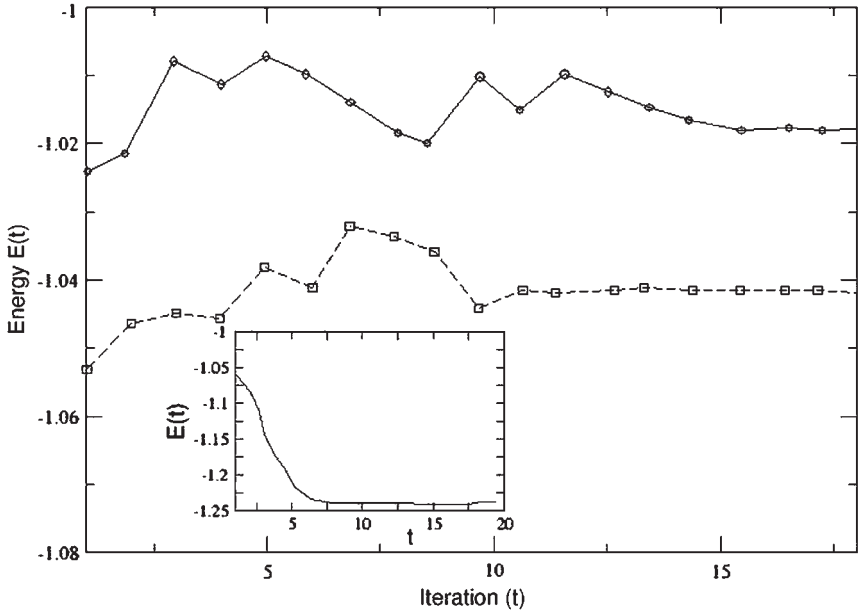


Fig. 6. The variation of the configurational energy $E(t)$ with time (updating iteration) t, for the sequential updating with $\lambda = 4$ for two typical cases (arbitrary distorted patterns) before and after reaching the fixed points. The inset shows the same for a distorted pattern with $\lambda = 0$ dynamics (adapted from Maiti et al., 1995).

It is to be noted that for positive λ, the overlap and other properties saturate beyond $\lambda = 1$ ($\lambda_R = 0$) for any α. This is true even for very large values of λ in the updating (6). This is somewhat tricky in the sense that apparently one time step is almost skipped in every updating; practically $\{S_i(t-2)\}$ (together with only a small fraction of $\{S_i(t-1)\}$) determines $\{S_i(t)\}$ through (6) and is kept in a *buffer* for latter use. Similarly $\{S_i(t-1)\}$ from the buffer is used to determine $\{S_i(t=1)\}$. We have checked independently the performance of such an updating ($S_i(t+1) = \mathrm{sgn}[\Sigma_j J_{ij}(\lambda' S_j(t) + \lambda S_j(t-1)]$, instead of (6), with $\lambda' \to 0$). We find for $\lambda' = 0$, the Hopfield model result is recovered. In fact, this can be seen from Fig. 3, where the overlap m_f decreases rapidly for very large values of λ in (6) ($\lambda \geqslant 10$). It may be noted, however, that the effective size of the domain of attraction of the fixed points depends very much on the ratio λ/λ'. This can be seen easily in the limit $\lambda \to 0$ but λ' finite (or $\lambda' \to 0$ but λ finite), where the change in energy of the network per updating step depends on the value of λ' (or λ), which determines the effective shape of the energy landscape seen by the dynamics.

As can be easily seen, for nonzero λ, the updating dynamics in (6) does not minimize the 'energy' function $E(t)$ defined as $E(t) = \sum J_{ij} S_i(t) S_j(t)$ as the delay term contributes in the internal field. Some typical variations in $E(t)$ with iterations (updating) are shown in Fig. 6 for $\alpha = 0.5$ and $\lambda = 4$. (The large values of α and λ are chosen here to ensure larger τ, so that the effect can be displayed over longer time range.) The results clearly show the contribution of the λ term in escaping over the spurious valleys, in the "energy" landscape. The inset in Fig. 6 shows the same variation for $E(t)$ in the $\lambda = 0$ (Hopfield) case. Here, of course, it clearly decreases monotonically with time.

We have also studied the performance of this dynamics for diluted network (with sequential dynamics; $N = 500$), where a random fraction c of the synaptic connections J_{ij} are removed. We show in Fig. 7, some results for overlap m_f against λ at a fixed loading capacity $\lambda = 0.20$ for some typical

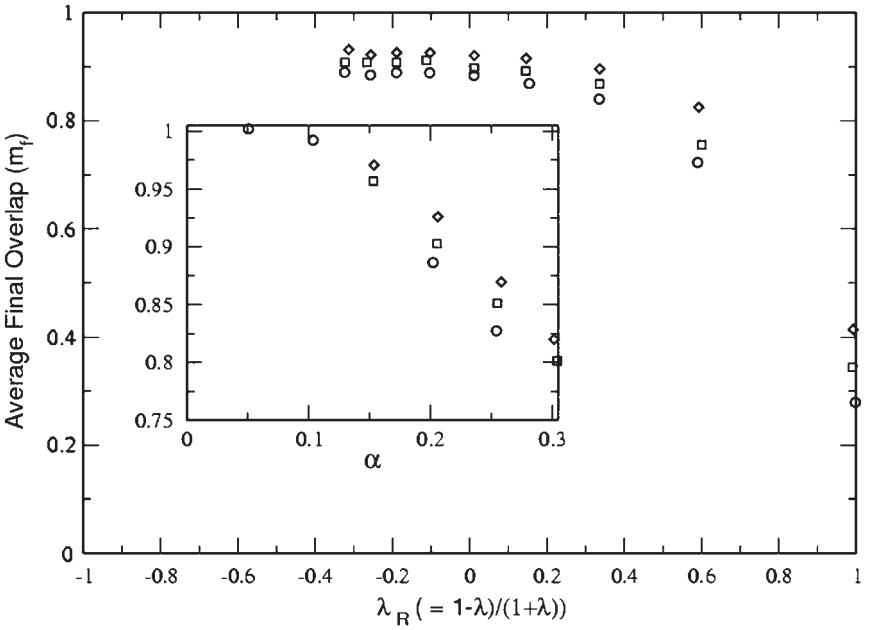


Fig. 7. The variation of average final overlap with, positive λ at a fixed capacity $\alpha = 0.20$ with sequential dynamics for $N = 500$ at some typical values of c (= 0.10, 0.20 and 0.25; the $c = 0.1$ data being at the top). The inset shows the variation of with α at $\lambda = 1$, for the same value of c (adapted from Maiti et al., 1995).

values of dilution concentration c. The inset shows the variation of m_f against α at a fixed λ (= 1) for some typical values of c.

Summary

In this paper, we have reviewed several existing numerical results on neural network models obtained by our group. The models used here have synaptic connections constructed by using the Hebb's rule. The dynamics is governed by the internal field given by (3) and (6). Some (new) analytic results for asymmetric interactions in a related spin glass model is given in the appendix.

Acknowledgment

We are grateful to C. Dasgupta, P. Sengupta, M. Ghosh, G. A. Kohring, P. Maity, A. K. Sen and P. Sen for collaborations at various stages of the above studies and to C. Dasgupta for a critical reading of the manuscript.

Appendix: A spin glass model with asymmetric interactions

A. The dynamical model

We discuss an asymmetric spin-glass model in the mean-field limit. We follow the notations of Crisanti and Sompolinsky (1987). The model consists of N fully connected spins interacting via random asymmetric interactions. The details and more complete studies of the same would be discussed elsewhere (Basu, 2007). We choose the interaction matrix between spins i and j to have the form

$$J_{ij} = J_{ij}^{s} + kJ_{ij}^{a},\ k \geq 0 \tag{7}$$

where J^{s}_{ij} and J^{a}_{ij} are symmetric and antisymmetric couplings respectively, such that

$$J^{s}_{ij} = J^{s}_{ji};\ \ J^{a}_{ij} = -J^{a}_{ji} \tag{8}$$

Each of J^{s}_{ij} and J^{a}_{ij} are zero-mean Gaussian distributed random variables with the variance

$$[(J^{s}_{ij})^2] = \frac{J^2}{N}\frac{1}{1+k^2} \tag{9}$$

As in Crisanti and Sompolinsky (1987) square brackets denote the quench average with respect to the distribution. The parameter k measures the degree of asymmetry in the interactions. Similar to Crisanti and Sompolinsky (1987), Eq. (9) implies

$$N\left[J^2_{ij}\right] = J^2, \quad N\left[J_{ij}J_{ji}\right] = J^2\frac{1-k^2}{1+k^2} \tag{10}$$

Therefore, from the above it is clear that for $k=0$ the model reduces to the ordinary symmetric spin glass model with infinite range interactions (Sompolinsky and Zippelius, 1982), whereas $k=1$ corresponds to a fully asymmetric model (Crisanti and Sompolinsky, 1987).

The dynamics for this model is defined by an appropriate Langevin equation for the i-th spin

$$\Gamma_0^{-1}\frac{\partial}{\partial t}\sigma_i(t) = -r_0\sigma_i(t) - \frac{\delta V(\sigma_i)}{\delta\sigma_i(t)} + \Sigma_j J_{ij}\sigma_j(t) + h_i(t) + \zeta_i(t) \tag{11}$$

In our model we consider a soft spin which varies continuously from $-\infty$ to ∞. The local potential $V(\sigma_i)$ is as chosen in Crisanti and Sompolinsky (1987) and is an even function of σ_i. The function $h_i(t)$ is a local external magnetic field. Further, the stochastic function $\zeta_i(t)$ is a zero-mean Gaussian distributed variable with a variance

$$\langle\zeta_i(t)\zeta_j(t')\rangle = \frac{2T}{\gamma_0}\delta(t-t')\delta_{ij} \tag{12}$$

Note that the above choice of noise correlations would have validated the FDT relating the appropriate correlation functions and the corresponding response functions for the fully symmetric ($k=0$) problem. Furthermore, the static properties of the fully symmetric version of the model can be derived from an appropriate Boltzmann distribution without any reference to the underlying dynamics. However, for the present case with a finite k there is no FDT and, equivalently, no Boltzmann distribution exists for the problem. As discussed in Crisanti and Sompolinsky (1987) and as we show below this makes the finite k problem far more complicated as compared to its fully symmetric version.

B. Dynamic generating functional

The generating functional for the correlation and the response functions for the model Eq. (11) is

$$Z[\phi,\hat{\phi}] = \int D\sigma D\hat{\sigma}\exp\left[\sum_i L_0(\hat{\sigma}_i,\sigma_i) + \frac{1}{2}\Sigma J^{s}_{ij}\{\sigma_j(t)\hat{\sigma}_i(t) + \hat{\sigma}(t')\} + \frac{1}{2}\Sigma J^{a}_{ij}\{\sigma_j(t)\hat{\sigma}_j(t) - \sigma_i(t)\hat{\sigma}_i(t')\}\right] \tag{13}$$

where L_0 is the local part of the action:

$$L_0 = \int dt i\hat{\sigma}(t)[-\Gamma_0^{-1}\partial_t\sigma_i(t) - r_0\sigma_i(t) - 4u\sigma_i^3 + h_i(t) + T\Gamma_0^{-1}i\hat{\sigma}_i(t)]$$

We now average over the distribution of J_{ij} to obtain

$$\langle Z\rangle = \int D\sigma\hat{D}\sigma\exp\left[L_0(\hat{\sigma}_i,\sigma_i) + 2J^2\sigma_j(t)\hat{\sigma}_i(t)\sigma_j(t')\hat{\sigma}_i(t') + 2\sigma_i(t)\sigma_j(t')\sigma_i(t')\hat{\sigma}_j(t)J^2\frac{1-k^2}{1+k^2}\right] \tag{14}$$

Here, the field $\hat{\sigma}$ is the MSR conjugate variable (Martin et al., 1973; Sompolinsky and Zippelius, 1982).

We now take the mean-field limit, which in the present context, is equivalent to assuming infinite range interactions (i.e., all spins are interacting with each other). In that limit, fluctuations, which

are O(1/N), are negligible (Sompolinsky and Zippelius, 1982). We then linearize the remaining quartic terms by Hubbard–Stratonovich transformation (Sompolinsky and Zippelius, 1982). To simplify further, in the limit $N \to \infty$, we substitute the stationary point values. Furthermore, we set due to causality. We thus obtain

$$\langle Z \rangle = \int D\sigma D\hat{\sigma} \exp\left[L_0 + \frac{\beta^2 J^2}{2} \times \int dt dt' \left\{ C(t-t') i\hat{\sigma}_i(t) i\hat{\sigma}_i(t') + 2\frac{1-k^2}{1+k^2} G(t-t') i\hat{\sigma}_i(t)\sigma_i(t') \right\} \right] \quad (15)$$

The effective equation of motion, corresponding to the generating functional (15) above is

$$\Gamma_0^{-1} \frac{\partial}{\partial t}\sigma_i(t) = r_0\sigma_i(t) - 4u\sigma_i^3 + \frac{J^2}{T^2}\frac{1-k^2}{1+k^2} \times \int_{-\infty}^{t} dt' G(t-t')\sigma_i(t') + \psi_i(t) \quad (16)$$

where the effective noise is zero-mean, Gaussian distributed with a variance

$$\langle \psi_i(t)\psi_j(t') \rangle = \frac{2}{\Gamma_0}\delta(t-t')\delta_{ij} + \beta^2 J^2 C(t-t')\delta_{ij} \quad (17)$$

or in the Fourier space

$$\sigma_i(\omega) = G_0(\omega)[\psi(\omega) + h_i(\omega)] - 4uG_0(\omega) \times \int d\Omega_1 d\Omega_2 \sigma_j(\Omega_1)\sigma_j(\Omega_2)\sigma_i(\omega - \Omega_1 - \Omega_2) \quad (18)$$

Further

$$G_0(\omega) = r_0 - i\omega\Gamma^{-1} - \frac{1-k^2J^2}{1+k^2T^2}G(\omega) \quad (19)$$

and the effective noise correlation is given by

$$\langle \psi_i(\omega)\psi_j(-\omega) \rangle = 2\delta_{ij}\left[\frac{2}{\Gamma_0} + \frac{J^2}{T^2} C(\omega) \right] \quad (20)$$

Note that, unlike the case of symmetric couplings, the effective noise no longer does not have any relation with the effective propagator (19). Therefore, the correlation function $C(t)$ and the propagator $G(t)$ are independent of each other in contrast to the case with symmetric couplings where the two are related by the FDT (Sompolinsky and Zippelius, 1982; Crisanti and Sompolinsky, 1987). Further, for the convenience of our subsequent analyses we write the effective noise $\psi_i(t)$ as the sum of three parts

$$\psi_i(t) = \eta_i(t) + x_i(t) + z_i \quad (21)$$

Here

$$\langle \eta_i(\omega)\eta_j(-\omega) \rangle = \delta_{ij}\left[\frac{2}{\Gamma_0} + \frac{J^2}{T^2}\frac{1-k^2}{1-k^2}\hat{C}(\omega) \right] \quad (22)$$

such that $\hat{C}(\omega) = (2T/\omega)\Im G(\omega)$, where $\Im$ refers to the imaginary part. Thus, η_i is the part of the total effective noise which respects the FDT. The function $\hat{C}(t)$ is the part of the correlator $C(t)$ which is related to the propagator $G(t)$ through the FDT. We further define $\hat{C}(t) \equiv C(t) - \hat{q}(t)$ such that $\hat{q}(t) = [\langle \sigma_i(t') \rangle_\eta \langle \sigma_i(t+t') \rangle_\eta]_{x_i z_i}$. Here, $\mathbf{Z}_i$ is the time-persistent part of the Gaussian noise with a variance

$$\langle z_i(\omega) z_j(-\omega) \rangle = 2\pi\delta(\omega)\beta^2 J^2 q \delta_{ij} \quad (23)$$

This then yields

$$\langle x_i(\omega) x_j(-\omega) \rangle = 2\pi\delta_i \left[\frac{2k^2}{1+k^2}\beta^2 J^2 \hat{C}(\omega) + \beta^2\{\hat{q}(\omega) - q\delta(\omega)\} \right] \quad (24)$$

With these effective noise correlators and propagator we now proceed to consider the dynamics at high temperature.

C. Dynamics at high temperature

In the high temperature phase there is no time persistent correlation. We begin by defining a damping function

$$\Gamma^{-1}(\omega) = i\frac{\partial G^{-1}}{\partial \omega} \quad (25)$$

Eqs. (19) and (25) together with the Dyson's equation yield

$$\Gamma^{-1}(\omega) = \frac{\Gamma_0^{-1} + i\frac{\partial\Sigma}{\partial\omega}}{1 - \frac{J^2}{T^2}G^2(\omega)\frac{1-k^2}{1+k^2}} \tag{26}$$

Equation (26), therefore, suggests that the inverse of the effective relaxation time Γ ($\omega = 0$) has a divergence at a critical temperature given by

$$T_c = JG(0)\sqrt{\frac{1-k^2}{1+k^2}} \tag{27}$$

This, however, holds provided $\partial\Sigma/\partial\omega$ has a finite limit for $\omega \to 0$. We now argue below in favour of such a result: we begin by noting that for $k = 0$, i.e., for the symmetric coupling case, $\Gamma(\omega) \sim \sqrt{\psi}$ in the small-ω limit, which in turn implies $G(\omega) \sim \sqrt{\omega}$ (Sompolinsky and Zippelius, 1982). We then note that (ω), a part of the total correlation $C(\omega)$, is related to the propagator $G(\omega)$ and hence has a small-ω dependence of the form $1/\sqrt{w}$. The total correlator $C(\omega)$ is given by

$$C(\omega) = G(\omega)G^*(\omega)\left[\Gamma^{-1}(\omega) + C(\omega)\right] \tag{28}$$

indicating that $C(\omega)$ is *no more* singular than $1/\sqrt{\omega}$. We are then led to the result that $\partial\Sigma/\partial\omega|\omega = 0$ is finite. We, therefore, conclude that for $T > T_c$ spin fluctuations decay as $|1 - T/T_c|^{-1}$, where T_c is given by (27).

D. Statics below

Having argued in favour of the possibility of a critical temperature T_c we now consider the static properties below T_c. We begin by setting up the Fokker–Planck equation (Chaikin and Lubensky, 2000) which governs the time-evolution of the probability distribution of the configurations of σ_i. In particular, we consider the probability distribution $P_1(\sigma, t|\sigma_0, t_0) \equiv \langle\delta(\sigma_i - \sigma_i(t))\rangle\sigma_0, t_0$, which is the probability of the configuration $\{\sigma\}$ at time t.

This implies

$$P_1(\sigma, t + \Delta t|\sigma_0, t_0) = \int D\sigma' P_1(\sigma, t + \Delta t|\sigma', t)P_1(\sigma', t|\sigma_0, t_0) \tag{29}$$

We now calculate the conditional probability $P_1(\sigma, t + \Delta t|\sigma', t)$ from the equation of motion: a Taylor series expansion yields

$$\begin{aligned}\sigma_i(t + \Delta t) = \sigma_i(t) - \Bigg[-r_0\sigma_i(t) - \frac{\delta V}{\delta\sigma_i} + J^2\frac{1-k^2}{1+k^2} \\ \times \int_{-\infty}^{t} dt' G(t - t')\sigma_i(t')\Bigg]\Gamma_0\Delta t \\ + \int_t^{t+\Delta t} \psi_i(t')dt' \\ + \frac{1}{2}\int_t^{t+\Delta t}\int_t^{t+\Delta t} \psi(t_1)\psi(t_2)dt_1dt_2\end{aligned} \tag{30}$$

Here, the noise $\psi = \eta_i + x_i + z_i$. We now perform averages over η_i and x_i which are zero-mean and Gaussian distribute with variances discussed above. After simplifications we finally find

$$\begin{aligned}\langle\delta[\sigma_i - \sigma_i(t + \Delta t)]\rangle_{\sigma_i' t'} \\ = \Bigg[1 + \Delta t\left(-r_0\sigma_i - \frac{\delta V}{\delta\sigma_i}\right) \\ + J^2\frac{1-k^2}{1+k^2}\int_t^{t+\Delta t} dt' G(t - t')\sigma_i(t') - z_i\Delta t\Bigg]\frac{\partial P_1}{\partial\sigma_i} \\ + \Delta t J^2\int_{-\Delta t}^{\Delta t}\left[\frac{2k^2}{1+k^2}J^2\hat{C}(t')J^2[\hat{q}(t') - q]\right]\frac{\partial^2}{\partial\sigma_i^2}P_1\end{aligned} \tag{31}$$

yielding

$$\begin{aligned}\frac{\partial P_1}{\partial t} = T\frac{\partial}{\partial\sigma_i}\Bigg[\frac{1}{T}\left(-r_0\sigma_i - \frac{\delta V}{\delta\sigma_i}\right) - z_i \\ + J^2\frac{1-k^2}{1+k^2}\int_t^{t+\Delta t} G(t - t')\sigma_i(t') \\ + J^2\int_{-\Delta t}^{\Delta t}\hat{C}(t')dt' + J^2\frac{2k^2}{1+k^2}[\hat{q}(t') - q]\Bigg]\end{aligned} \tag{32}$$

The steady-state solution of Eq. (32), P_1 (*steady*) can be obtained by setting $\partial P_1/\partial t$ to zero. From P_1 (*steady*) one would be able to calculate $q \equiv [\langle\sigma\rangle^2_{\eta,x}]_z$. From Eq. (33) one notes that $\hat{q} \equiv [\langle\sigma\rangle^2_{n,x}]_z$ appears in Eq. (32). Therefore, one would need an equation for the distribution $P_2 \equiv [\langle\sigma\rangle_\eta(x,z)]$. The equation is

$$\frac{\partial P_2}{\partial t} = \frac{\partial}{\partial\sigma}\left[-r_0\sigma_i - \frac{\delta V}{\delta\sigma_i} - z_i - x_i(t) + J^2\frac{1-k^2}{1+k^2}\int_t^{t+\Delta t} G(t-t')\sigma_i(t') + J^2\frac{1-k^2}{1+k^2}\int_{-\Delta t}^{\Delta t}\hat{C}(t')dt'\frac{\partial}{\partial\sigma_i}\right]P_2 \quad (33)$$

Clearly there is no steady-state solution since has an explicit time dependence. We formally write the solution of the Eq. (33) as $P_2(t) = \exp[\int_0^t L(t)dt]P_2(t=0)$ where L is the operator

$$\frac{\partial}{\partial\sigma}\left[-r_0\sigma_i - \frac{\delta V}{\delta\sigma_i} - z_i - x_i(t) + J^2\frac{1-k^2}{1+k^2}\int_t^{t+\Delta t} G(t-t')\sigma_i(t') + J^2\frac{1-k^2}{1+k^2}\int_{-\Delta t}^{\Delta t}\hat{C}(t')dt'\frac{\partial}{\partial\sigma_i}\right] \quad (34)$$

and $P_2(t=0)$ is the initial condition. The function (t) is then given by

$$\hat{q}(t) = \int D\sigma P_2(t=0)D\sigma DxDz \times \exp\left[\int_t^{t^0} Ldt'\right]\sigma_i(t')\sigma_i(t'+t)P[x]P[z] \quad (35)$$

Note that we have performed averaging over initial conditions also. Functions $P[z]$ and $P[x]$ are distributions of z and x respectively. Assuming that the initial distribution is normalised we have

$$\hat{q}(t) = \int D\sigma DxDz \times \exp\left[\int_0^{t^0} L(t')dt'\right]\sigma_i(t')\sigma_i(t'+t)P[x]P[z] \quad (36)$$

Further, the $m(z) \equiv \langle\sigma\rangle_{\eta,x} = \int D\sigma P_1[z]\sigma$. These formally complete the discussions on the static properties below T_c. In summary we have investigated a spin glass model with asymmetric spin–spin interactions in the mean-field limit. Due to the asymmetry there is no FDT in the model and consequently analysis of the model becomes far more complicated. Our result [Eq. (28)] suggests that the spin glass transition temperature is reduced in the presence of the asymmetry, and therefore, indicates the possibility of higher memory capacity of neurons with such asymmetric synaptic interactions. Thus asymmetric interactions tend to suppress the spin-glass phase. In this context we refer to some recent related studies: Crisanti and Sompolinsky (1987) studied a spherical model and the Ising version of the problem respectively., and found no spin-glass phase at any finite temperature for any strength of the asymmetry. Further extensive studies on this problem are required for a satisfactory resolution of the issues raised here.

References

Amit, D.J. (1989) Modelling Brain Function. Cambridge University Press, Cambridge.

Basu, A. (2007). Unpublished.

Chaikin, P.M. and Lubensky, T.C. (2000) Principles of Condensed Matter Physics. Cambridge University Press, Cambridge.

Chakrabarti, B.K. and Dasgupta, P.K. (1992) Phys. A, 186: 33–48.

Crisanti, A. and Sompolinsky, H. (1987) Phys. Rev. A, 36: 4922–4939.

Ghosh, M., Sen, A.K., Chakrabarti, B.K. and Kohring, G.A. (1990) J. Stat. Phys., 61: p. 501.

Hertz, J., Krough, A. and Palmer, R.G. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley, Reading, MA.

Maiti, P., Dasgupta, P. and Chakrabarti, B.K. (1995) Int. J. Mod. Phys. B, 9: 3025–3037.

Martin, P.C., Siggia, E.D. and Rose, H.A. (1973) Phys. Rev. A, 8: 423–437.

Sen, P. and Chakrabarti, B.K. (1989) Phys. Rev. A, 40: 4700–4703.

Sen, P. and Chakrabarti, B.K. (1992) Phys. Lett. A, 162: 327–330.

Sompolinsky, H. and Zippelius, A. (1982) Phys. Rev. B, 25: 6860–6875.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 14

A simple Hopfield-like cellular network model of plant intelligence

Jun-ichi Inoue*

Complex Systems Engineering, Graduate School of Information Science and Technology, Hokkaido University, N14-W9, Kita-Ku, Sapporo 060-0814, Japan

Abstract: We introduce a simple Hopfield-like cellular-network model to explain a kind of "intelligence" in plants (Trewavas, 2002), especially, the capacity of plants to act as a memory device. Following earlier observations by Indian scientist J.C. Bose (1923), we regard the plant as a network in which each of the elements is connected via negative interactions. We investigate properties of the model by statistical mechanics.

Keywords: plant intelligence; Hopfield model; associative memories; statistical-mechanical analysis; storage capacity; phase transition

Introduction

Since pioneering studies by Indian scientist J.C. Bose (1923), plants have been regarded as a kind of network which are capable of intelligent responses to environmental stimuli. For example, the dodder coil, which is a plastic plant, explores a new host tree within hours subsequent to initial contact (Trewavas, 2002). This sort of behavior might be regarded as *plant intelligence*. If that is the case, does the plant compute, learn or memorize various spatial and temporal patterns in different environments similar to a computer or human brain (Ball, 2004)? Recently, Peak et al. (2004) pointed out that the plants may regulate their uptake and loss of gases by a distributed computation. As is well known, the ability of neural networks, which is a mathematical model of the brain, is also based on parallel and distributed computation. Therefore, similarities between neural network models of brains and the plant networks should be discussed. Although the behavior of the dodder coil we mentioned above is due to the emergence of intelligence as a macroscopic function, it is important for us to investigate its rationale on a microscopic scale. Over 80 years ago, J.C. Bose (1923) detected electrical signaling between plant cells. Since then, following his experiments, many examples of cross-talk between biochemical signaling pathways in plants have been discovered. Especially, a Boolean representation of networks of signaling pathways is possible in terms of logical gates like AND, OR and XOR etc. These Boolean descriptions make it possible to draw analogies between plant networks and neural network models. Recently, Brueggemann et al. (1998) found the plant vacuolar membrane current–voltage characteristic to be equivalent to that of a Zenner diode. Inspired by their work, Chakrabarti and Dutta

*Corresponding author. Tel.: +81-11-706-7225; Fax: +81-11-706-7391; E-mail: j_inoue@complex.eng.hokudai.ac.jp

DOI: 10.1016/S0079-6123(07)68014-5

(2003) utilized the threshold behavior of plant cell membranes to develop or model gates for performing simple logical operations. They found that the plant network connections are all positive (excitatory by means of neuronal states) or all negative (inhibitory), compared to the randomly positive–negative distributed synaptic connections in real brains. As a result, the plant network does not involve any frustration in their computational capabilities and might lack the distributed parallel computational ability as in the case of associative memories. Even if we observe a single plant cell and provide a mathematical model for the single cell, we will not understand the macroscopic behavior of the plant giving rise to emergent properties like learning, memory and recognition (Genoud and Metraux, 1999; Bose and Karmakar, 2003). With this fact in mind, we investigate one of such collective behavior, namely, associative memories of plant networks, in this chapter. In other words, we try to answer the question "Do plant networks act as memory devices?" Especially, following the analysis by Chakrabarti and Dutta (2003), we investigate the equilibrium properties of the Hopfield model in which both ferromagnetic retrieval and anti-ferromagnetic terms co-exist. The strength of the anti-ferromagnetic order is controlled by a single parameter λ. Within the replica symmetric calculation of statistical physics (Hertz et al., 1991; Nishimori, 2001), we obtain phase diagrams of the system. The λ-dependence of the optimal loading rate α_c, which is defined as the ratio of the maximum number of embedded patterns to the number of neurons in the brain, namely, $P_{\max}/N$, at $T=0$ is discussed.

This chapter is organized as follows. In the next section, we introduce several experiments and observations related to the current–voltage characteristics of the plant cell membrane. In section "The Plant Intelligence Model", we model the plant possessing such properties by using a Hopfield-like network model in which both ferromagnetic-retrieval and anti-ferromagnetic terms exist. In section "Replica Symmetric Analysis", we analyze the model with the assistance of the replica method. In this section, we investigate to what extent the ferromagnetic retrieval order remains against the anti-ferromagnetic disturbance. We also investigate the result of the ferromagnetic disturbance. All the results are summarized in the final section.

The I–V characteristics of cell membranes

In this section, we briefly mention several results concerning properties of the plant units (cells), namely, current (I)–voltage (V) characteristics of their cell membrane. In Fig. 1, we show the typical non-linear I–V characteristics of cell membranes acting as logical gates.

From this figure, we find that the I–V characteristics are equivalent to those of the Zenner diode. Namely, some threshold v_T exists crossing which, the direction of the current changes. By assuming that the current has two directions, that is, ± 1, and the voltage is determined by the weighted contributions, Chakrabarti and Dutta (2003) constructed a mathematical plant cell as a non-linear unit. From the viewpoint of input–output logical units like perceptrons for neural networks, the output of the i-th unit O_i is given by

$$O_i = \Theta\left(\sum_{j=1}^{N} w_{ij} I_j\right)$$

where strength of each connection w_{ij} is all positive or negative, while in the Hopfield model it is given as $\pm$ randomly distributed weight matrix in terms

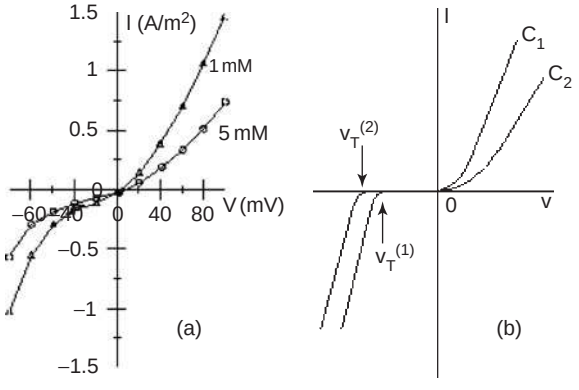


Fig. 1. The non-linear I–V characteristics of cell membranes.

of the Hebbian rule

$$w_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \qquad \xi_i^\mu \in \{-1, 1\}$$

From these experimental results and simple observations, we now ask a natural question, which is, could the plants act as memory devices similar to a real brain? Obviously, in the above definition of a single unit in the plant, there is no frustration as in animal brains. Thus, this paper attempts to elucidate, as to what extent constraints on the sign of the weight matrix influences the ability in plants to retrieve patterns as associative memories.

For this purpose, we introduce a simple plant intelligence model based on a Hopfield-like model in which ferromagnetic retrieval and anti-ferromagnetic ordered phases co-exist. In the next section, we explain its details.

The plant intelligence model

We start from the following Hamiltonian

$$\begin{aligned} H &= \frac{1}{N}\sum\nolimits_{ij}\left(\lambda - \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu\right) S_i S_j \\ &= H_{AF} + H_{FR} \end{aligned}$$

where we defined the following two parts of the total Hamiltonian

$$H_{AF} = \frac{\lambda}{N}\sum_{ij} S_i S_j$$

$$H_F = \frac{1}{N}\sum_{ij}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu S_i S_j$$

In these expressions, the vector $\vec{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$ is μ-th embedded pattern and $\vec{S} = (S_1, ..., S_N)$ stands for neuronal states. A single parameter λ determines the strength of the anti-ferromagnetic order, that is to say, in the limit of $\lambda \to \infty$, the system is completely determined by the energy function H_{AF}. On the other hand, in the limit of $\lambda \to 0$, the system becomes identical to the conventional Hopfield model. The purpose of this chapter is to investigate the λ-dependence of the system, namely, to study the λ-dependence of the optimal loading rate $\alpha_c(\lambda)$ at $T = 0$ by using the technique of statistical mechanics for spin glasses.

Replica symmetric analysis

In order to evaluate macroscopic properties of the system, we first evaluate the averaged free energy

$$[\log Z]_{\vec{\xi}} = [\log tr_{\vec{S}} \exp(-\beta H)]_{\vec{\xi}}$$

where $[\cdots]_{\vec{\xi}}$ denotes the quenched average over the extensive $P = \alpha N$ patterns. To carry out this average and spin trace $tr_{\vec{S}} \exp(-\beta H)$, we use the replica method by using the relation

$$[\log Z]_{\vec{\xi}} = \lim_{n \to 0} \frac{[Z^n]_{\vec{\xi}} - 1}{n}$$

After standard algebra, we obtain the pattern-averaged replicated partition function under the replica symmetric approximation as follows

$$[\log Z]_{\vec{\xi}} = \underset{m,q,M,r}{extr} \exp(nN\Phi(m, q, M, r))$$

with

$$\begin{aligned} \Phi(m, q, M, r) = {} & \frac{\beta}{2}M^2 - \frac{\beta\lambda}{2}m^2 + \frac{\alpha\beta^2 r}{2}(1 - q) \\ & + \frac{a}{2}\left\{\log[1 - \beta(1 - q)] - \frac{\beta q}{1 - \beta(1 - q)}\right\} \\ & - \log \int_{-\infty}^{\infty} Dz \log 2\cosh \beta(\lambda m + \sqrt{\alpha r} z + M) \end{aligned}$$

where we define $Dz = dz \exp(-z^2/2)/\sqrt{2\pi}$. We should keep in mind that physical meanings of m and M are magnetization of the system, overlap between the neuronal state $\vec{S}$ and a specific recalling pattern $\vec{\xi}^1$ among $p = \alpha N$ embedded patterns, respectively. The order parameter q stands for spin glass order parameters. In the next section, we evaluate the saddle point of this free energy density Φ and draw phase diagrams to specify the pattern retrieval properties of the system.

Phase diagrams

In this section, we investigate the phase diagram of the system by solving the saddle point equations. By taking the derivatives of the free energy density Φ with respect to m, q, M and r, we obtain the saddle point equations

$$M = \int_{-\infty}^{\infty} Dz \tanh \beta[(1-\lambda)M + z\sqrt{\alpha r}] = -m$$

$$q = \int_{-\infty}^{\infty} Dz \tanh^2 \beta\left[(1-\lambda)M + z\sqrt{\alpha r}\right]$$

$$r = \frac{q}{[1-\beta(1-q)^2]}$$

We solve the equations numerically to obtain the phase diagram.

We first investigate the noiseless limit $T \to 0$. Obviously, in this limit, $q = 1$ holds. After some algebra, we find that the optimal loading rate α_c is determined by the point at which the solution of the following equation with respect to y vanishes

$$y\left\{\sqrt{\alpha} + \sqrt{\frac{2}{\pi}}(1-\lambda)\exp\left(-\frac{z^2}{2}\right)\right\} = (1-\lambda)\{1-2H(y)\}$$

where we defined the error function by

$$H(x) = \int_x^{\infty} Dz$$

In Fig. 2, we plot the optimal loading rate α_c as a function of λ.

From this figure, we see that the optimal loading rate α_c decreases monotonically. This means that the ferromagnetic retrieval order was destroyed by adding the anti-ferromagnetic term to the Hamiltonian. Thus, we conclude that if the components of the weight matrix of the networks are all positive, the plant intelligence model does not act as a memory device.

Before we solve the saddle point equations for $T \neq 0$, it should be important to determine the phase boundary between the spin glass and paramagnetic phases. The phase transition between these two phases is of first order, by expanding the saddle point equations around $M = q = 0$, we obtain the phase boundary line $T_{SG} = (1-\lambda)(1+\sqrt{\alpha})$.

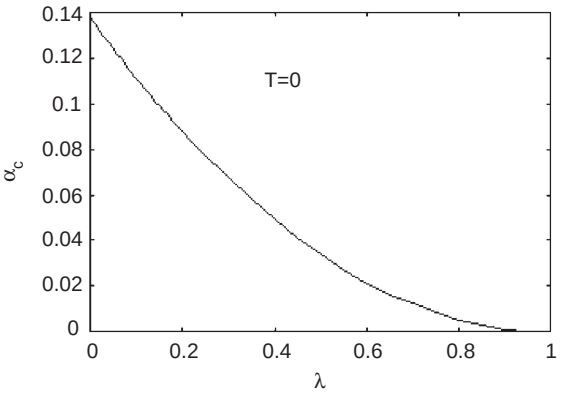


Fig. 2. The optimal loading rate α_c as a function of λ.

Now, we investigate the phase diagram for $T \neq 0$ by solving the saddle point equations numerically. We show the result in Fig. 3.

From this figure, we find that the ferromagnetic retrieval phase shrinks to zero as the anti-ferromagnetic order increases as $\lambda \to 1$. The behavior of the overlap M as a function of α is shown in Fig. 4. From this figure, we find that the overlap M becomes zero discontinuously at $\alpha = \alpha_c$.

We next consider the case of negative λ. From the Hamiltonian, we find

$$H = -\frac{1}{N}\sum_{ij}\sum_{\mu=1}^{P}\xi_i^{\mu}\xi_j^{\mu}S_iS_j - \frac{\lambda'}{N}\sum_{ij}S_iS_j$$
$$\lambda' = -\lambda(>0)$$

As the parameter λ increases, the system changes to the pure ferromagnet.

Let us consider the limit of the parameter $\lambda \to -\infty$ in the saddle point equation: $M = \int_{-\infty}^{\infty} Dz \tanh \beta[(1-\lambda)M + z\sqrt{\alpha r}]$. Then, the term $(1-\lambda)M$ appearing in the argument of $\tanh \beta(\cdots)$ becomes dominant, namely $(1-\lambda)M \gg -z\sqrt{\alpha z}$ even if the loading rate α is large. Consequently, equation $M = \int_{-\infty}^{\infty} Dz \tanh$

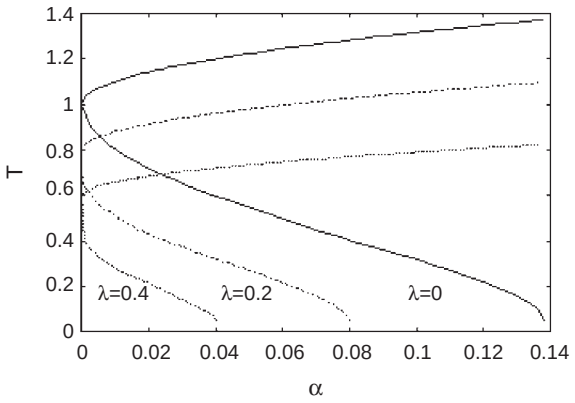


Fig. 3. The phase diagram of the system.

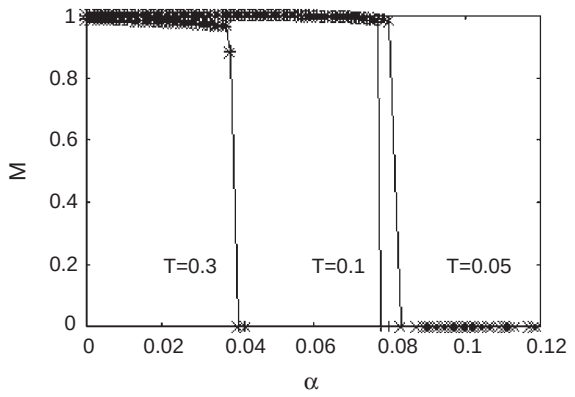


Fig. 4. The overlap M as a function of α for the case of $\lambda = 0.2$ at temperatures $T = 0.05$, 0.1, 0.3.

$\beta[(1-\lambda)M + z\sqrt{\alpha r}]$ leads to

$$M = \tanh \beta[(1-\lambda)M]$$

Apparently, this equation has always a positive solution even if the temperature $T = \beta^{-1}$ is large. In this sense, the factor $(1-\lambda)$ can be interpreted as temperature re-scaling. It is also possible for us to understand this result from a different point of view. In the saddle point equation

$$M = \int_{-\infty}^{\infty} Dz \tanh \beta[(1-\lambda)M + z\sqrt{\alpha r}]$$

the second term appearing in the argument of tanh, $z\sqrt{\alpha r}$ means *cross-talk noise* from the other

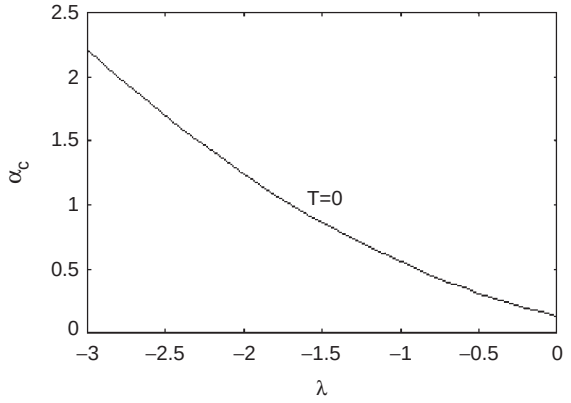


Fig. 5. The optimal loading late α_c as a function of $\lambda(<0)$ at $T=0$.

patterns $\vec{\xi}^{\mu} = (\xi_1^{\mu}, \ldots, \xi_N^{\mu})$, $\mu = 2, .., P$, and obeys Gaussian distribution with zero mean and unit variance. On the other hand, the first term $(1-\lambda)M$ represents the signal of the retrieval pattern $\vec{\xi}^{1}$. Therefore, if the second term $z\sqrt{\alpha r}$ is dominant, the system cannot retrieve the embedded pattern $\vec{\xi}^{1}$. Usually, r in the second term grows rapidly as T increases. Obviously, if α increases, the noise term $z\sqrt{\alpha r}$ also increases. As a result, the signal part $(1-\lambda)M$ becomes relatively small and the system moves from the retrieval phase to the spin glass phase. However, if λ is negative and large, the signal part is dominant, and as a result, the noise part becomes vanishingly small. This is an intuitive reason why the optimal loading rate increases for negative λ.

In Fig. 5, we plot the optimal loading late α_c as a function of $\lambda(<0)$ at $T=0$. As we mention, the optimal loading rate α_c monotonically increases as λ goes to $-\infty$.

Concluding remarks

In this chapter, we introduced a simple model based on a Hopfield-like network to explain the intelligence of plants. Inspired by the studies of Chakrabarti and Dutta (2003), we regarded a plant cell as a non-linear unit and investigated its corrective behavior as a network. In order to investigate the possibility of the plant network as a memory device, we constructed a mathematical

associative memory based on the Hopfield model in which ferromagnetic retrieval and anti-ferromagnetic terms co-exist. The strength of disturbance of pattern retrieval by the anti-ferromagnetic order is controlled by a single parameter. We found that the anti-ferromagnetic order prevents the system from recalling a pattern. This result means that the ability of the plant as a memory device is very weak if we set all weight connections to positive values. On the other hand, if we set all connections of the network to be negative, the capacity for memory increases. Our analysis in this paper was done for fully connected networks, and of course, for real plants, the cell membrane should be located in a finite dimensional lattice (Koyama, 2002) or in a scale-free network (Stauffer et al., 2003). So, extensive studies might be needed to investigate such "structure dependence" in the ability of the plant networks.

Acknowledgments

The author thanks Professor Bikas K. Chakrabarti for collaboration. This chapter was based on our publication (Inoue and Chakrabarti, 2005). He also acknowledges Professor Rahul Banerjee for good organizing the international workshop on Models of Brain and Mind: Physical, Computational and Psychological Approaches in Kolkata 2006.

References

Ball, P. (2004). Do plants act like computers? Nature Science Update Weekly Highlights, 26 January 2004.

Bose, I. and Karmakar, R. (2003) Simple models of plant learning and memory. Phys. Scr., T106: 9–12.

Bose, J.C. (1923) The Nervous Mechanism of Plants. Longmans, London.

Brueggemann, L.I., Pottosin, I.I. and Schoenkecht, G. (1998) Cytoplasmic polyamines block the fast-activating vacuolar cation channel. Plant J., 16(1): 101–105.

Chakrabarti, B.K. and Dutta, O. (2003) An electrical network model of plant intelligence. Ind. J. Phys., A77: 551–556.

Genoud, T. and Metraux, J.-P. (1999) Crosstalk in plant cell signaling: structure and function of the genetic network. Trends Plant Sci., 4(12): 503–507.

Hertz, J., Krough, A. and Palmer, R.G. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley Publishing, Cambridge, MA.

Inoue, J. and Chakrabarti, B.K. (2005) Competition between ferro-retrival and anti-ferro orders in a Hopfield-like network model for plant intelligence. Phys. A, 346: 58–67.

Koyama, S. (2002) Storage capacity of two-dimensional networks. Phys. Rev. E, 65: 016124-1–016124-6.

Nishimori, H. (2001) Statistical Physics of Spin Glasses and Information Processing. Oxford University Press, New York.

Peak, D.A., West, J.D., Messinger, S.M. and Mott, K.A. (2004) Evidence for complex, collective dynamics and emergent, distributed computation in plants. Proc. Acad. Sci. U.S.A., 101: 918–922.

Stauffer, D., Aharony, A., da Fontoura Costa, L. and Adler, J. (2003) Efficient Hopfield pattern recognition on a scale-free neural network. Eur. Phys. J. B, 32: 395–399.

Trewavas, A. (2002) Plant intelligence: mindless mastery. Nature, 415(6874): p. 841.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 15

Retinomorphic image processing

Kuntal Ghosh[1], Kamales Bhaumik[2,*] and Sandip Sarkar[3]

[1]*Centre for Soft Computing Research, Indian Statistical Institute, 203, B.T. Road, Calcutta 700108, India*
[2]*West Bengal University of Technology, BF-142 Salt Lake, Calcutta 700064, India*
[3]*Microelectronics Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Calcutta 700064, India*

Abstract: The present work is aimed at understanding and explaining some of the aspects of visual signal processing at the retinal level while exploiting the same towards the development of some simple techniques in the domain of digital image processing. Classical studies on retinal physiology revealed the nature of contrast sensitivity of the receptive field of bipolar or ganglion cells, which lie in the outer and inner plexiform layers of the retina. To explain these observations, a difference of Gaussian (DOG) filter was suggested, which was subsequently modified to a Laplacian of Gaussian (LOG) filter for computational ease in handling two-dimensional retinal inputs. Till date almost all image processing algorithms, used in various branches of science and engineering had followed LOG or one of its variants. Recent observations in retinal physiology however, indicate that the retinal ganglion cells receive input from a larger area than the classical receptive fields. We have proposed an isotropic model for the non-classical receptive field of the retinal ganglion cells, corroborated from these recent observations, by introducing higher order derivatives of Gaussian expressed as linear combination of Gaussians only. In digital image processing, this provides a new mechanism of edge detection on one hand and image half-toning on the other. It has also been found that living systems may sometimes prefer to "perceive" the external scenario by adding noise to the received signals in the pre-processing level for arriving at better information on light and shade in the edge map. The proposed model also provides explanation to many brightness–contrast illusions hitherto unexplained not only by the classical isotropic model but also by some other Gestalt and Constructivist models or by non-isotropic multi-scale models. The proposed model is easy to implement both in the analog and digital domain. A scheme for implementation in the analog domain generates a new silicon retina model implemented on a hardware development platform.

Keywords: ganglion cells; non-classical receptive field; Gaussian derivatives; zero-crossings; low-level illusions; image half-toning; silicon retina

Introduction

One of the most effective means for survival in the case of *Homo sapiens*, are erect postures and binocular vision, which leads to the perception of the external world by their eyes and the brain. It is, therefore, not surprising that since its inception, many computer vision algorithms are designed to be as anthropomorphous as possible by mimicking the functioning of the eye and the brain. For example, the most popular codes for edge detection are designed from cues obtained

*Corresponding author. Tel.: +91-33-23210731;
Fax: +91-33-23341032; E-mail: kamales.bhaumik@wbut.ac.in

DOI: 10.1016/S0079-6123(07)68015-7

through the study of the physiology of the early visual system. The process of receiving an image by a nearly two-dimensional retina is quite similar to the planar recording of the images by any digital device. However, there are three major differences. Firstly, for a retinal image the pixel sizes are not exactly identical; secondly, the pixels do not span the entire retinal surfaces (there are minute gaps) and lastly, the density of the photoreceptors also varies between the fovea and other regions of the retina. In spite of these differences the retinal image may be well approximated as a digital image. Images, recorded by the retina or any digital device, are often contaminated with noise. (In some special situations, noise may add more information, as will be shown in this chapter). The first step in the processing is, therefore, filtering out the noise as far as practicable. It has been observed that the Gaussian filters are best suited for smoothening an image, because it optimizes both the spatial variance and also the bandwidth. For a two-dimensional image, the Gaussian filtering has the added advantages, namely, it is rotationally symmetric, it preserves the neighborhood characteristics and it is separable into two one-dimensional Gaussians. Image smoothening is, therefore, generally achieved by convolving it with a Gaussian function.

Let us now look into the mechanism of image processing in the human eye. Since the retina is an extension of the brain, a good deal of image processing occurs in the retina itself. Classical investigations by Barlow (1953), Kuffler (1953) and Wiesel (1960) had shown that information about the image is extracted in the successive layers of the retina by a center-surround effect. For example, a bipolar cell in the outer plexiform layer (the layer immediately beneath the layer of photoreceptor cells) receives information from a large number of photoreceptors distributed over a circular zone mainly through a network of horizontal cells. While the receptors in the central region of this zone send information to the bipolar in a positive fashion, the information from the receptors in the periphery of this zone arrives with a reversal of signature. As a result a central bright spot with dark background is the best stimulus for exciting a bipolar cell. (These bipolar cells are known as on-centre cells. There are also off-centre bipolar cells for which a dark spot with bright background is the most appropriate stimulus.) Such an antagonistic effect is further enhanced at a ganglion cell in the inner plexiform layer that collects the information from a large number of bipolar cells through a network of amacrine cells. Ganglion cells may also be of two types: on-centre or off-centre and each of these types are highly contrast sensitive. It is clear from these observations that the retina sends the contrast information to the visual cortex for further processing. Probably the simple cells of the primary visual cortex detect the contrast edges of an image (Hochstein and Spitzer, 1984) from this information through an appropriate mechanism. Any particular ganglion cell in the inner (deeper) layer of the retina may receive information from the photoreceptors located only in a particular area of the retinal surface. That particular area (assumed to be circular or elliptical in shape) is called the receptive field of that particular ganglion cell. Strength of the output from a photoreceptor to the ganglion cell should be a maximum when they are in closest proximity. It is also natural to assume that the contributions received by a ganglion cell from other receptors will smoothly fall off with the distance. Such a distribution in which the peak value is at the center and the value falls of with distance can be easily fitted with a Gaussian function. This would be true for both positive (center) and negative (surround) inputs. Consequently the net input to a ganglion cell is obtained from a difference of two Gaussian inputs (Fig. 1), the central one (positive) having a smaller variance than the surround (negative). This prompted physiologists like Rodieck (1965) or Enroth-Cugell and Robson (1966) to develop a model of difference of Gaussian (DOG) model for which the resultant looks like a Mexican hat in two-dimension. The DOG model is very effective in explaining a large number of experimental findings in retinal responses and hence it is the universally accepted model for the centre-surround antagonistic effects observed at a retinal ganglion cell.

Apart from the DOG model, as suggested by the physiologists, one can look for an alternative

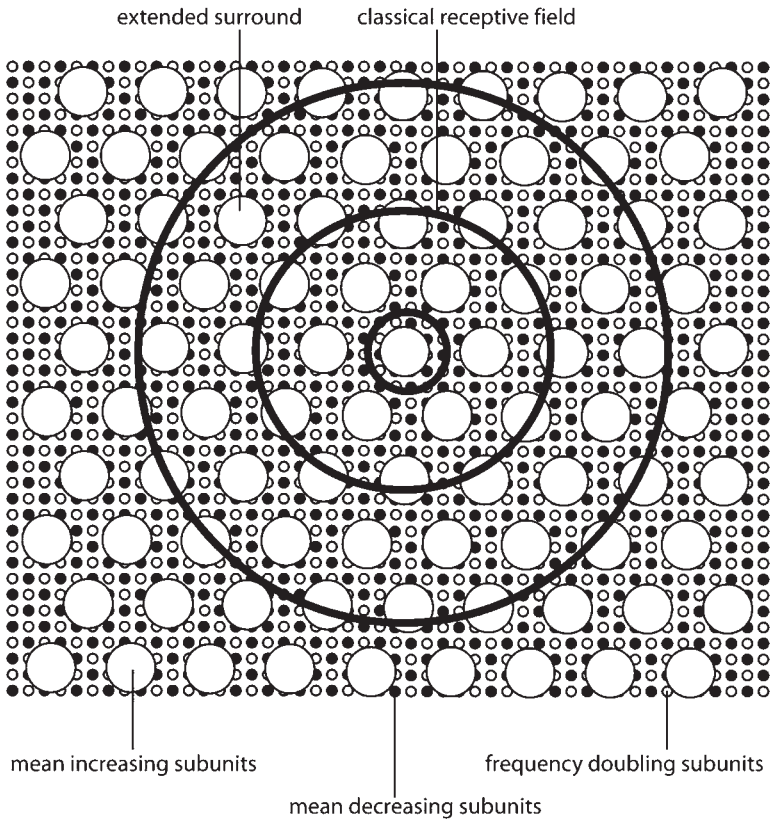


Fig. 1. The non-classical receptive field structure.

model from the computational viewpoint. Since the primary job of the retina is to generate the necessary information for detection of contrast edges in the image, a retinomorphic image-processing model should base itself on this task. Detection of edge means detection of discontinuity, which may be best achieved with the help of a derivative function. The model should, therefore, have two operations, namely smoothening the image (preferably with a Gaussian) and then to take the derivative. A good choice for a combined filter would be, therefore, to take a derivative of the Gaussian. Marr and Hildreth (1980) pointed out that the Laplacian, which is a rotationally invariant and linear differential operator, may be a good choice. Laplacian operated on a two-dimensional Gaussian, known as Laplacian of Gaussian (LOG) function, is a good alternative for DOG, because of its computational ease. Marr and Hildreth further argued that for a certain ratio of the two scale parameters in DOG, LOG could be considered as a good approximation to DOG. For over the last half a century, almost all the popular algorithms for edge detection use one or other variants of DOG or LOG model.

Modified model

In the previous section, we have discussed the genesis of the DOG model from the viewpoint of retinal physiology and also the argument for proposing an almost equivalent but computationally advantageous alternative model called LOG. Till date, nearly all the popularly accepted models for edge detection are merely variants of one of

these models. Difference of Gaussians or the DOG model looks like

$$\mathrm{DOG}(\sigma_1,\sigma_2)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{x^2}{2\sigma_1^2}}-\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{x^2}{2\sigma_2^2}} \quad (1)$$

Here σ_1^2 and σ_2^2 are the variances (or scale parameters) of the two Gaussian functions representing the center and the surround respectively. LOG model is obtained by operating a Laplacian operator on a two-dimensional Gaussian function. In two-dimension it looks like

$$\nabla^2 G(r)=-\frac{1}{\pi\sigma^2}\left[1-\frac{r^2}{2\sigma^2}\right]e^{-\frac{r^2}{2\sigma^2}} \quad (2)$$

In view of the recent experimental observations on the retinal physiology, we propose to modify these existent models. Experiments by Hochstein and Shapley (1976) prove that some photoreceptor cells outside the classical receptive field are capable of modulating the behavior of the ganglion cells. It confirms the earlier reports of an extended surround of retinal ganglion cells (McIlwain, 1966; Ikeda and Wright, 1972). Existence of such cells was verified through a large number of experiments, though there were inconsistencies regarding the nature of polarity of the output of those cells. While the previous studies by Ikeda and Wright (1972) as well as recent studies by Kaplan and Benardete (2001) and many others indicate a positive output (increasing the mean firing rate), there are also evidences in favor of negative output (decreasing the mean firing rate) from these cells (Enroth-Cugell and Jakiela, 1980; Kruger, 1984). Whatever may be the polarity of the output, there is no doubt that the actual receptive field of a ganglion cell is much more widely spread than that depicted by the classical picture. Such a non-classical receptive field is shown in Fig. 1, following the suggestion of Passaglia et al. (2001), where it is conjectured that the "mean increasing" and "mean decreasing" units would remain either active or inactive depending on the desired task of the retina. Some of the effects of the Extended Classical Receptive Field (ECRF) may be emulated by modeling the corresponding response behavior as a linear combination of three zero-mean Gaussians at three different scales (Ghosh et al., 2006a–c), which perhaps represent the classical center, the classical antagonistic surround and the non-classical extended disinhibitory surround (mostly contributed by the amacrine cells) respectively, in ascending order of their widths. So in one-dimension

$$\begin{aligned}\mathrm{ECRF}(\sigma_1,\sigma_2,\sigma_3)=&A_1\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{x^2}{2\sigma_1^2}}\\&-A_2\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{x^2}{2\sigma_2^2}}\\&+A_3\frac{1}{\sqrt{2\pi}\sigma_3}e^{-\frac{x^2}{2\sigma_3^2}}\end{aligned} \quad (3)$$

Here ECRF represents the response function for the extended classical receptive field of retinal ganglion cells, σ_1, σ_2 and σ_3 represent the scales of the center, the antagonistic surround and the extended disinhibitory surround respectively and A_1, A_2 and A_3 represent the corresponding amplitudes. The Eq. (3) may be looked upon as a modification to Eq. (1) for classical receptive field, with an extra correction term added for the extended surround. We have observed that this modified model shows considerable improvement over the results obtained from the classical model in various aspects of image processing, like edge detection, image enhancement etc. (Ghosh et al., 2006a, c). In this chapter, we shall specially emphasize on two contributions of this model in some areas, which the image analysts generally do not consider, namely, low-level illusions and half toning of an image.

Low-level brightness–contrast illusions

Visual illusions are generated by a complex mechanism of the brain and are realized at the perceptual level, apparently beyond the structure and function of the retina. It is generally believed, particularly by the Constructivist theorists, that the models of illusions should be studied through a top-down approach (Adelson, 2000). However, for some illusions, known as low-level brightness–contrast illusions, the major information is supposed to be extracted by the contrast sensitive

ganglion cells of the retina. A retinomorphic image-processing algorithm is expected to account for this special class of illusions. It is indeed true that the classical DOG or LOG models are sufficient to explain many of these illusions like Mach band illusion, Simultaneous Brightness–contrast illusion and Grating Induction illusion (Palmer, 1999). However the classical models fail to explain a large number of low-level illusions. These models cannot account for some finer but important points in the Hermann Grid illusion and fails completely in explaining the White Effect illusion (Ghosh et al., 2006a–c). Our modified model, not only provides an adequate explanation of these illusions, it also explains four different types of Todorovic Effect, the Checkerboard stimulus and the Howe Stimulus. These latter illusions are, till date, believed to be beyond the realm of retinal processing. In this section, we shall describe these illusions and show how the modified model provides adequate explanation for these phenomena. Since Mach band illusion, the Simultaneous Brightness–Contrast illusion and the Grating Induction illusion are explainable both by the traditional DOG model and also by our modified ECRF model, we are avoiding the discussion of those illusions. We are giving below the explanation provided by the ECRF model for those illusions in which the DOG model failed either partially or totally. The illusory images have all been convolved with a two-dimensional version of the proposed ECRF filter, followed by a horizontal line scan of the convolved images that reproduce the respective illusory perceptions.

Hermann grid illusion

This refers to the illusory impression of gray blobs found at the intersection of the white lines when one looks at a grid of black squares on a white background (Fig. 2a). The conventional isotropic DOG model can provide an explanation for the generation of these blobs as they are generated due to lateral inhibition between the center and surround of the receptive field. However, with conventional DOG filters the peripheral area appears darker than the illusory dark spots, which is contrary to our perceived experience. This happens because with the isotropic DOG filter, the peripheral area receives inhibition from all directions, whereas the intersections of the grids receive inhibition only from four directions. It is clear (Fig. 2b) that ECRF model can reproduce these features quite faithfully. This is because in the ECRF model an effect of disinhibition is introduced from the extended surround region that neutralizes the effect of inhibition so that the lightness of the peripheral region appears intermediate between the streets and the crossings in Hermann Grid.

White effect illusion

White (1979) discovered an illusory effect that does not depend on the amount of dark or white borders in the vicinity of the test patch. In a square grating of black and white bars, if identical gray segments are used to replace part of the black bars and also part of the white bars, then former gray segments look brighter than the latter (Fig. 3a). Conventional isotropic DOG filters, fail to simulate this illusion and produce results contrary to our perception. White (1981) explained the effect by proposing a model in which inhibition was supposed to be stronger along the bars than across them. Such a supposed anisotropy in lateral inhibition was not observed in White's effect on checkerboard (to be discussed later) or isotropic random dot patterns, thereby disproving the theory. Gestalt theorists believe that White's effect can be understood only in terms of perception at a higher level (Agostini and Proffitt, 1993) and perhaps this effect is beyond the realm of low-level brightness–contrast illusion. Thus to probe whether the explanation of the White effect could have a basis in retinal physiology, we find that the illusion can be faithfully explained with the ECRF model (Fig. 3b).

Todorovic effects

Todorovic (1997) occluded a test patch on a black background by four white squares and vice-versa to create some novel illusions (Fig. 4). DOG filters completely fail to simulate these illusions (Blakeslee and McCourt, 1999; Palmer, 1999).

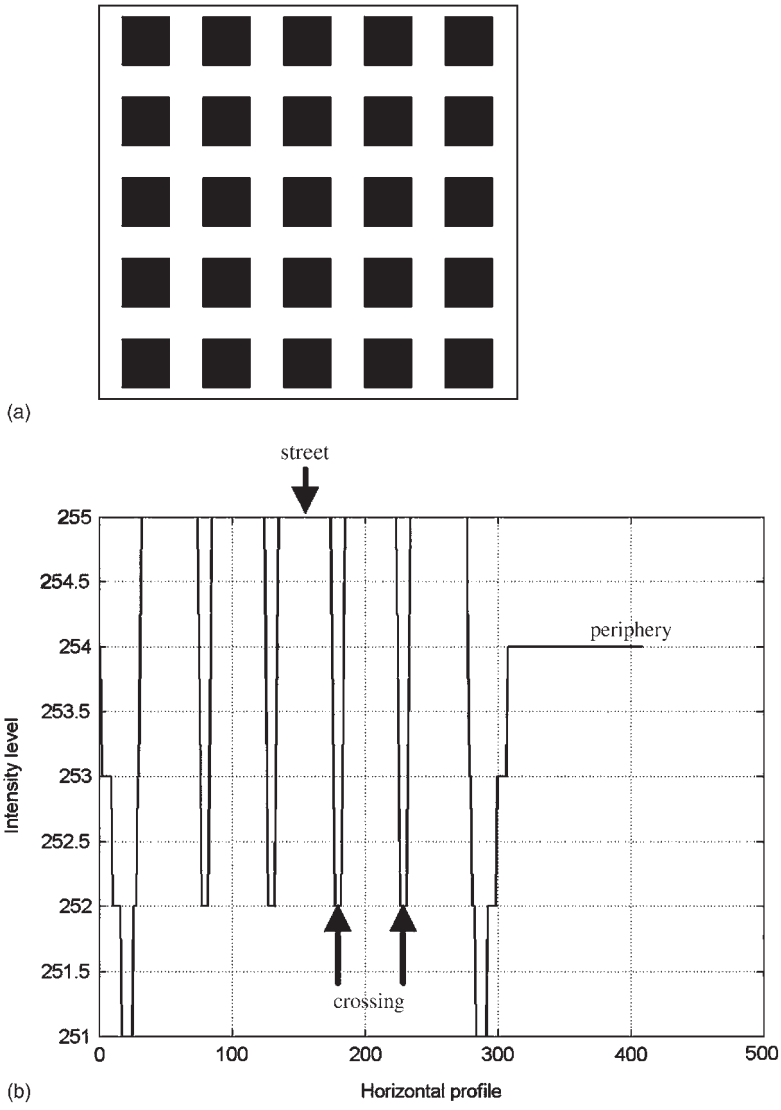


Fig. 2. (a) The Hermann Grid illusion. (b) Explanation of the Hermann Grid illusion with the *ECRF* filter along a horizontal line profile through a street and the crossings.

In fact Blakeslee and McCourt (1999) added considerable modifications to the traditional DOG model to explain these effects. However they could explain only the first two types and eventually proposed that perhaps the last two types (Blakeslee and McCourt , 1999) might represent perception at a higher level. Even without contesting that hypothesis, we may still consider these illusions to be caused by brightness–contrast based low-level visual signal processing. In fact the ECRF model can be used to solve all the Todorovic effects including the last two. We have shown here only the first Todorovic effect. Success of ECRF model in other three Todorovic effects is described in detail in Ghosh et al. (2006b). The scales of the Gaussians corresponding to center, surround and the extended surround for these simulations have been chosen as $\sigma_1 = 0.7, \sigma_2 = 3\sigma_1$ and $\sigma_3 = 9.3\sigma_1$.

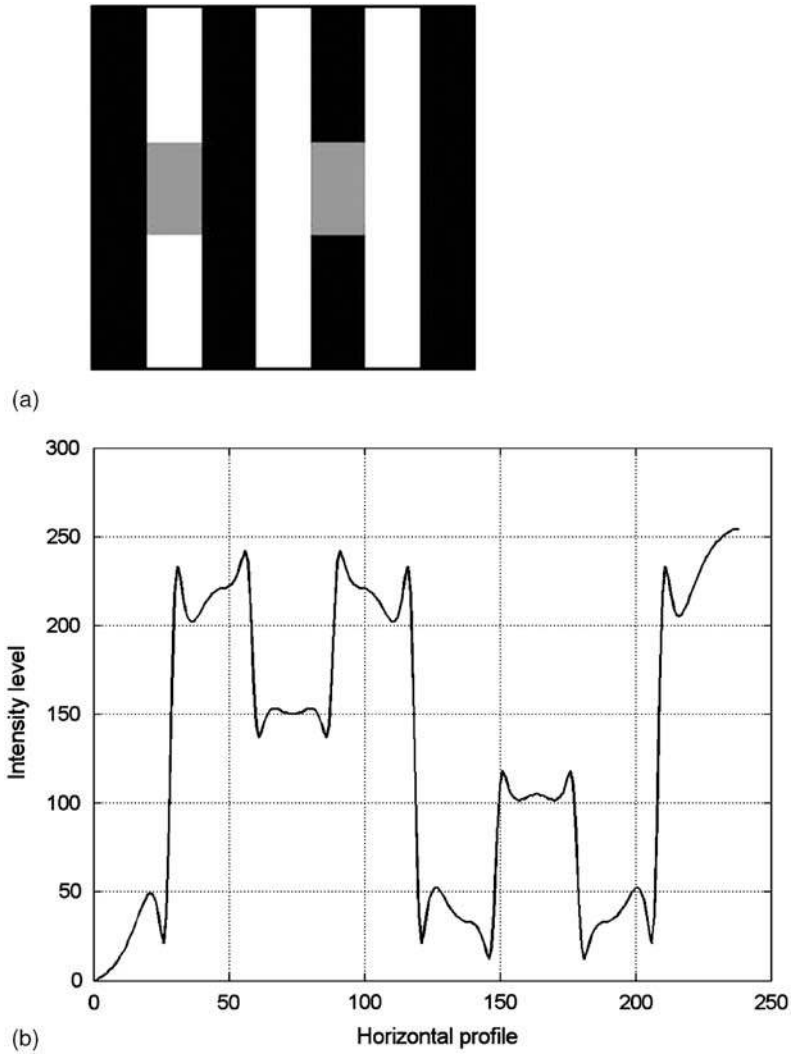


Fig. 3. (a) The White effect illusion. (b) Explanation of the White effect illusion by convolving the image with the *ECRF* filter along a horizontal line profile through the two test patches in the convolved image.

These values may not be unique, but Shou et al. (2000) have shown that the diameter of the extended surround is about ten times that of the classical center.

The hypothetical modification of the DOG model in which the lateral inhibition is supposed to have orientation preferences, may be effective in explaining White effect but fails completely in the Checkerboard illusion. In the Checkerboard stimulus (DeValois and DeValois, 1988), shown in Fig. 5, the test patch with a darker neighborhood on the left appears less bright than the one with brighter neighborhood on the right. There is no orientation anisotropy in the checkerboard stimulus and hence it cannot be explained either by DOG or its suggested modification (White, 1981). We find from Fig. 5b that the *ECRF* model is capable of explaining this illusion. The White effect, Checkerboard illusions and Todorovic illusions have led many investigators from the Gestalt school to abandon analysis of images on the basis of spatial filtering and receptive field.

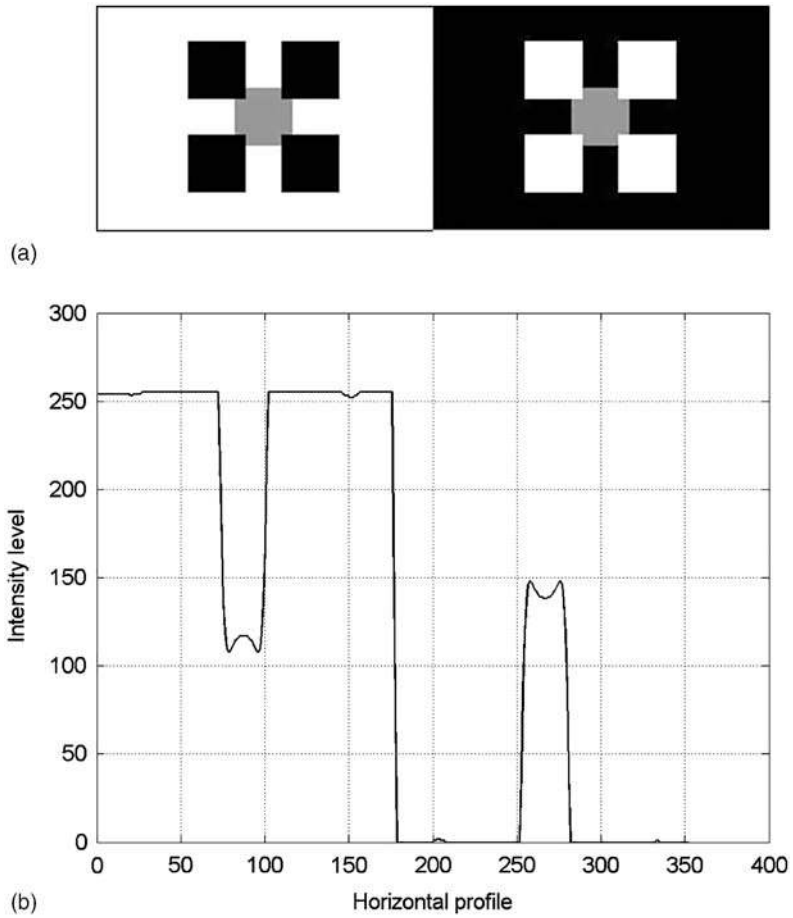


Fig. 4. (a) The Todorovic illusion. (b) Explanation of the illusion by convolving the image with *ECRF* filter along a horizontal line profile through the two test patches in the convolved image.

We would like to emphasize that the conclusion of the Gestalt school may not be true in general.

The present analysis, hints at the possibility that in the processing of low-level illusions, the number of contributing sub-units (Passaglia et al., 2001) from the ECRF and also their weight factors may vary from one illusion to the other. At least at the cortical level, Fiorani et al. (1992) and Gilbert and Wiesel (1992) have demonstrated that the size of the effective surround of a receptive field may vary with the nature of the stimulus. Wennekers and Suder (2004) have fitted such dynamic receptive fields by sum of amplitude-modulated spatial Gaussians. The present model, based on retinal physiology, is able to explain not only the standard low-level brightness–contrast illusions, but also those, which were so far supposedly totally outside the purview of low-level vision.

Extension ECRF model in derivative form

In the beginning, we have discussed in detail how the popular models like DOG or LOG were generated. Marr and Hildreth (1980) proposed an empirical equivalence of the two models to convert a linear combination of Gaussian to a derivative form. In detecting edge or any discontinuity, obviously the derivative functions are quite effective. Particularly interesting is the zero-crossings of

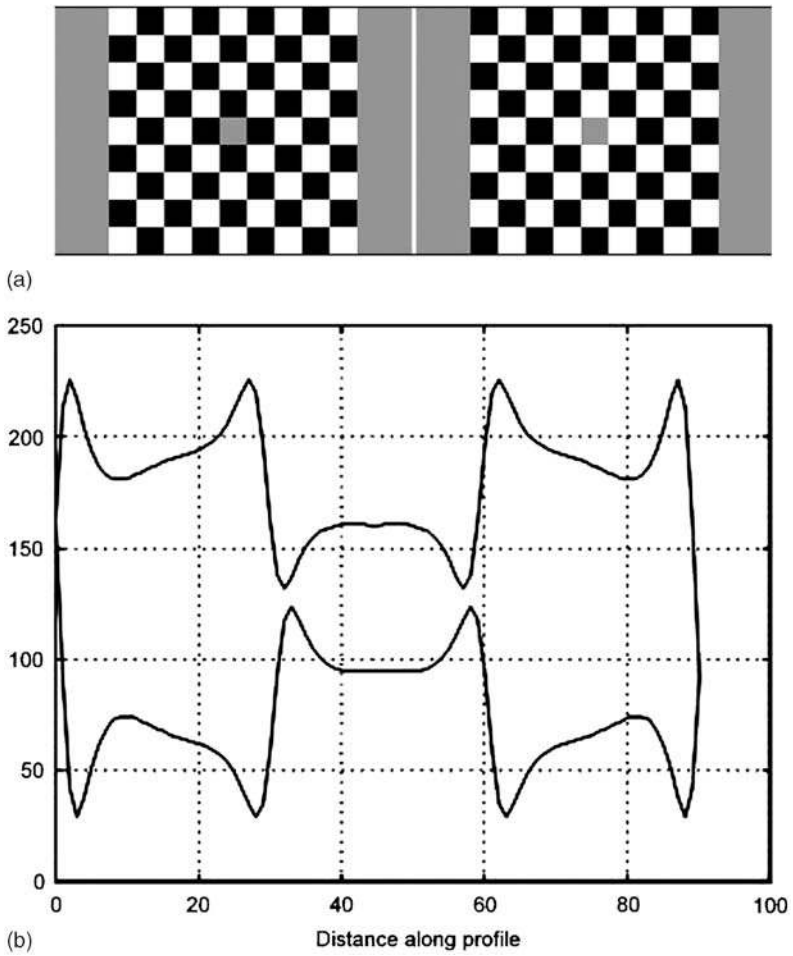


Fig. 5. (a) The DeValois and DeValois checkerboard stimulus. (b) The horizontal line profiles through the two test patches. From the horizontal line profiles it is clear that the test patch on the left in brighter neighborhood appears brighter compared to the one on right, as is also our perceptual experience.

the second derivatives (Fig. 6). However, it is also clear from Fig. 6 that mere detection of these zero-crossings in the cortex (Hochstein and Spitzer, 1984) does not provide any information about the height of the step or in other words the edge strength, a problem which we would like to address in the light of the ECRF model.

Long after the proposal of Marr and Hildreth, a theoretical justification of the equivalence between a linear combination of Gaussians and the higher derivatives of Gaussian has been provided by the following theorem (quoting partially) of Ma and Li (1998).

Theorem: Any *2kth* order derivative filter can be designed simply as the weighted linear combination, $(k+1)$ in number, of an even function, each of the $(k+1)$ functions having the same kernel, differing only in scale.

Corollaries of the theorem

Let us define a function h_{2k} using the primitive Gaussian filter $g(x)$ as

$$h_{2k}(x) = \sum_{j=0}^{k} \frac{\alpha_j}{\sigma_j} g\left(\frac{x}{\sigma_j}\right) \tag{4}$$

$$\text{where} \quad g\left(\frac{x}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{5}$$

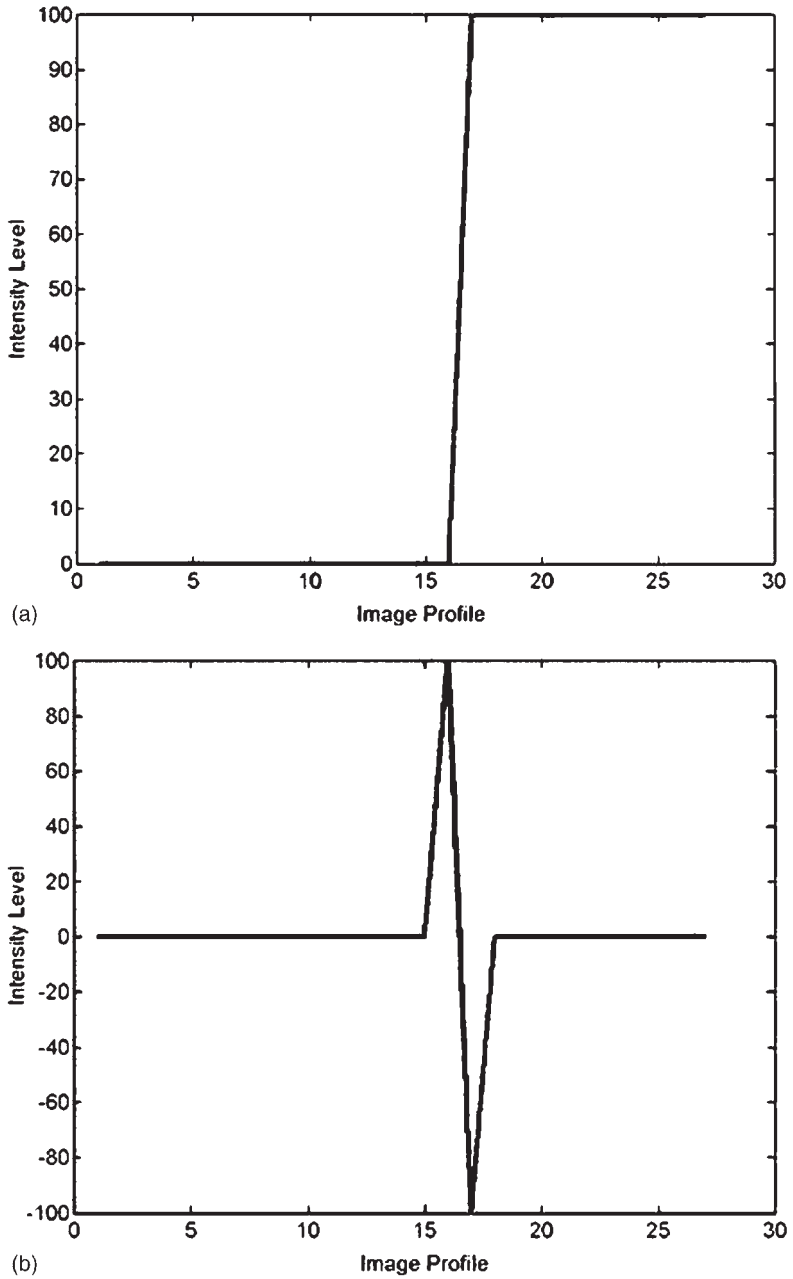


Fig. 6. (a) A function showing a simple one-dimensional step edge. (b) Second order derivative of the step edge shows a zero-crossing. Note that the location of zero-crossing faithfully reproduces the location of the edge. However no information about the step height can be obtained by merely detecting the zero-crossing.

Ma and Li showed that h_{2k} is *2kth* order derivative filter if the weight functions α_j's satisfy certain specified conditions (Ghosh et al., 2005b).

By taking $k=1$, for the second order derivative, in one-dimension

$$h_2(x) = \alpha_0\left(\frac{1}{\sigma_0}g\left(\frac{x}{\sigma_0}\right) - \frac{1}{\sigma_1}g\left(\frac{x}{\sigma_1}\right)\right) \tag{6}$$

For a scale (standard deviation in this case) ratio t of the two Gaussian kernels, i.e., if $\sigma_1 = \sigma$ and $\sigma_0 = t\sigma$

$$h_2(x) = \alpha_0\left(\frac{1}{(t\sigma)\sqrt{2\pi}}e^{-\frac{x^2}{2(t\sigma)^2}} - \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}\right) \tag{7}$$

This clearly provides the equivalence between LOG and DOG. For $t<1$, the first and second Gaussian in the left hand side of the Eq. (7), may be identified as the contribution from the classical center (mostly coming from bipolar cells) and classical surround (mostly coming from horizontal cells) respectively in correspondence to the contrast sensitive classical receptive field model of the retina.

Using the same theorem, a fourth order derivative of Gaussian can be expressed as a combination of three zero-mean Gaussian kernels with three scales σ_0, σ_1 and σ_2 and two scale ratios t and p, such that $\sigma_2 = \sigma, \sigma_1 = t\sigma$ and $\sigma_0 = p\sigma$, to yield

$$h_4(x) = k\sigma^2\begin{pmatrix} \{(1-t^2)/p\sigma\sqrt{2\pi}\}\exp(-x^2/2p^2\sigma^2) \\ -\{(1-p^2)/t\sigma\sqrt{2\pi}\}\exp(-x^2/2t^2\sigma^2) \\ +\{(t^2-p^2)/\sigma\sqrt{2\pi}\}\exp(-x^2/2\sigma^2) \end{pmatrix} \tag{8}$$

Polarities of the three Gaussians, defined in Eq. (8), indicate the possibility of considering the three Gaussians as representatives of the center, surround and disinhibitory extended surround portions of the non-classical receptive field respectively, in ascending order of their widths. Thus $h_4(x)$ essentially covers the basic aspects of ECRF, discussed earlier. We would now prefer to express $h_4(x)$ as a correction term over $h_2(x)$.

Experimental observations (McIlwain, 1966; Ikeda and Wright, 1972; Passaglia et al., 2001) on non-classical receptive fields indicate that the central region is much smaller than the extended surround, or in other words σ_0 is negligible in comparison to σ_2. Assuming the ratio $\sigma_0 : \sigma_2$ to be very small or in other words $p \to 0$ one can easily obtain (Ghosh et al., 2005a)

$$h_4\left(x, \sigma^{///}\right) \to m\delta(x) + h_2\left(x, \sigma^{//}\right) \tag{9}$$

where $\delta(x)$ is the Dirac-delta function in one-dimension. Here $\sigma^{//}$ and $\sigma^{///}$ are functions of σ_0, σ_1 and σ_2. In two-dimension the equivalent expression would be given as

$$\nabla^4 G(r) \to m\delta(x,y) + \nabla^2 G(r) \tag{10}$$

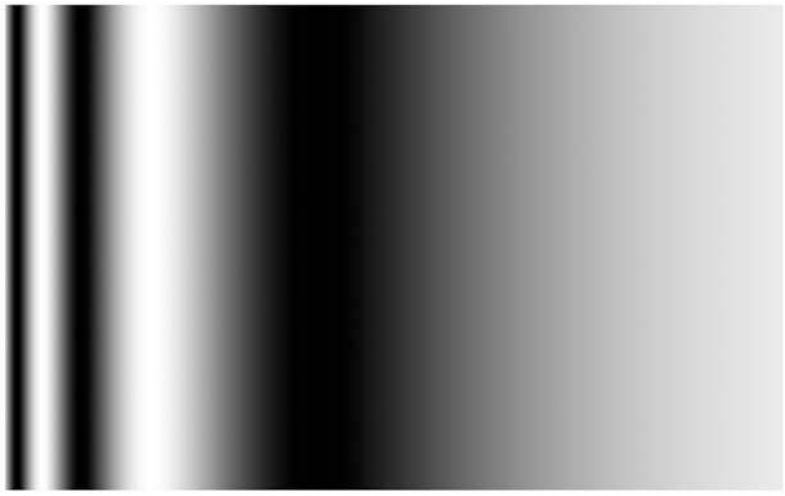
Following identical steps one can show that a linear combination of four Gaussians may be expressed equivalently by a sixth order derivative of Gaussian, which may be further expressed as

$$\nabla^6 G(r) \to m\delta(x,y) + \nabla^4 G(r) \tag{11}$$

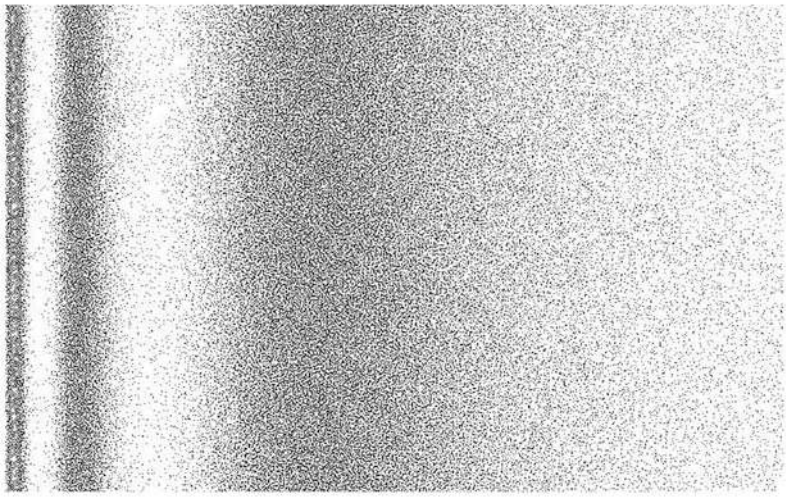
Since the higher order derivatives are expressed here as a summation of a delta function and a lower derivative, additional advantage is likely to be obtained through the detection of zero-crossing by such an operator which may lead to information about edge strength, that is unavailable in the zero-crossing map directly obtained from a simple derivative operation, whatever might be the order. It may be shown (Ghosh et al., 2005a) that $\nabla^4 G$, expressed as $m\delta(x,y) + \nabla^2 G(r)$ gives rise to two zero-crossings, one for $\nabla^2 G$ and one for the direct current (DC), i.e. the mean firing rate shift of the convolved response due to the presence of the Delta function in the filter function. Similarly one can show (Ghosh et al., 2007) that the operator $m\delta(x,y) + \nabla^4 G(r)$ produces four zero-crossings at each gray level transition, three for $\nabla^4 G$ and the fourth for DC shift, with the relative distance of each of the first three from the fourth storing the gray scale information in the image. Thus when higher order derivatives are used, it is possible to obtain significant information about the gray scale of an image from its zero-crossing map. Such a computed zero-crossing map is found to retain

intensity information of the original image in the sense of a half toning, where the intensity variation is mapped to the density variation of the zero-crossing points. We shall illustrate this statement with examples.

Let us first consider the filter h_4. Two kinds of images were considered for these studies. They are synthetic images that are perfectly noise free and natural images that generally contain intrinsic noise. The image in Fig. 7a is constructed using the function $0.5\sin(a/x)+0.5$. Figure 7b is the zero-crossing map after noise has been added to this image. The density of zero-crossings clearly depicts the intensity of the gray value. The effect of



(a)



(b)

Fig. 7. Influence of noise on zero-crossing image for the case of (a) original image created by using the function $0.5 \times \sin(a/x)+0.5$; (b) zero-crossings after adding noise to the original image. Variation of density of zero-crossings in (b) represents the variation of intensity in the original image (a).

noise in a natural image is depicted in Fig. 8. The image shown in Fig. 8a is the original image, image in Fig. 8b is the zero-crossing image of the LOG convolved image of Fig. 8a. As expected, such a LOG-convoluted zero-crossing map does not retain much of the intensity information of the original image. Figure 8c is the zero-crossing map of the original image, contaminated with a Gaussian noise ($\sigma = 0.03$). Contamination by such an extra noise, in addition to the inherent noise present in the natural image, shows (through visual inspection) some improvement in retaining the intensity information of the original image. It may be pointed out over here that there are evidences (Douglass et al., 1993) that living systems sometime prefer to "perceive" the external scenario by adding noise to the received signals in the pre-processing level. This may be due to the fact that though natural images contain noise, the amount of noise becomes optimum, for extracting the relevant information, after the addition of some external noise to the original image. In human vision, pupil noise, for example can play a constructive role in preserving intensity information in the zero-crossing maps assuming that such maps are computed by convolving the image with the class of operators, proposed above. We are tempted to make another remark here. Any zero-crossed map is a compressed version of the original image. But for reasons mentioned previously, such maps, which are binary images, cannot serve the role of being compressed versions of the gray scale distribution in the image. But for our proposed class of filters, for the consequent binary maps which do represent the gray level distribution, we arrive at a new and unconventional methodology for image compression.

It is well known (Weiss, 1994) that depending on geometrical complexity of the image contour, higher order derivative filters may be used to extract the local invariants of shape. However since the problem of de-localization rapidly increases with the increase of scale (Marr, 1982), there is very little freedom in these algorithms in choosing the scale of smoothening. Hence, though important subtle local variations of the edge can be captured by the higher order derivative filters, the overall processing of a noisy image may worsen as one moves from lower to higher derivatives, due to uncontrolled smoothening. It has already been shown with the filter h_4 that the present algorithm may in fact take advantage of the noise present in the image, resulting in a pronounced "fuzzy" effect in the sense of a half tone in the zero-crossing map. The Dirac-delta function with its multiplicative constant imparts into the zero-crossing map, additional information about image intensity.

We are now going to examine this for the higher order derivative filter h_6 (Eq. 11), by externally adding noise to the simple image of a three-dimensional box (Fig. 9). In this image, there are

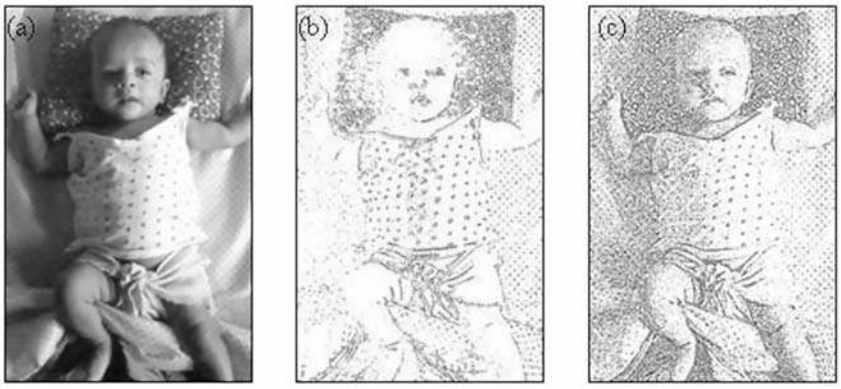


Fig. 8. Influence of noise on zero-crossing image for the case of natural image (a) original image (I); (b) zero-crossing image of $I \otimes LOG$; (c) zero-crossing image computed from Gaussian noise contaminated image convolved with the proposed filter h_4.

three distinct gray levels, one in the background and two more (the top and the sides) in the box. It can be seen, that this intensity variation between the background and the two gray levels in the three-dimensional box has been reflected in the edge map. Apparently the design of this filter as an edge extractor may appear to be strange, because instead of coming up with a clean edge map it is emphasizing shallow shading gradients. However, from the point of view of image reconstruction, like in the human visual system, such shading information is important and should not be lost. Higher order perceptual processes in the cortex require that information. The present model has the possibility to form the basis of such perceptual processes. This effect of density variation of the zero-crossings in tune with the intensity variation in the image, imbued by the Dirac-delta function, can be viewed in all orientations since the delta function is combined with only isotropic higher order derivatives. The phenomenon is similar to a half-toning effect in the zero-crossing map. This concept of multi-scale higher order Gaussian derivatives, which effectively behave as "fuzzy" derivatives in the sense of a half-toning effect, thus yields a new method of edge detection as well as a new and unconventional methodology for image compression taking the help of noise.

Hardware implementation of ECRF filter

In this concluding section we are presenting a very brief account of our progress in hardware implementation of the proposed ECRF filter. The details of the hardware design (Sarkar et al., 2006) are beyond the scope of this chapter. The implementation is mostly in the light of the famous silicon retina model of Mead and Mahowald (1988, 1989) and the consequent developments in edge detecting chips (Bair and Koch, 1991; Park et al., 2003). These chips follow the classical DOG model and implement the effect through a difference of exponential or DOE model, because the latter can be easily incorporated by computing the difference in outputs of two two-dimensional CMOS resistive networks. Effect of extended receptive field, as discussed in the previous sections, can be taken into account by modifying the DOE model with a corresponding ECRF model given by

$$\mathrm{ECRF} = A_0\delta(x) + \frac{A_1}{2\sigma_1}e^{-\frac{|x|}{\sigma_1}} + \frac{A_2}{2\sigma_2}e^{-\frac{|x|}{\sigma_2}} + \frac{A_3}{2\sigma_3}e^{-\frac{|x|}{\sigma_3}} \tag{12}$$

where $\sigma_1 < \sigma_2 < \sigma_3$ are the scale factors and A_1, A_2 and A_3 are the weight factors.

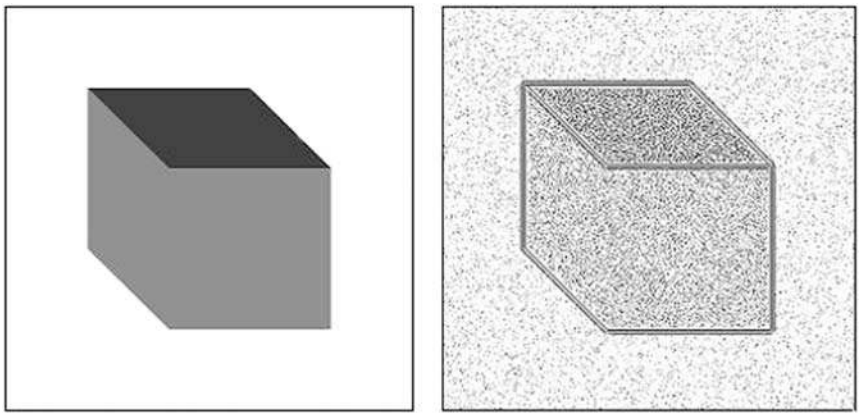


Fig. 9. Zero-crossing map of the three-dimensional box, with externally added noise, using the filter $m\delta(x,y) + \nabla^4 G(r)$ at $\sigma = 1.2$ and $m = 0.1$. The intensity variation in the image (between the background and the top and the sides of the box) is reflected in the map in the sense of a half-toning effect or "fuzzy" derivative computation.

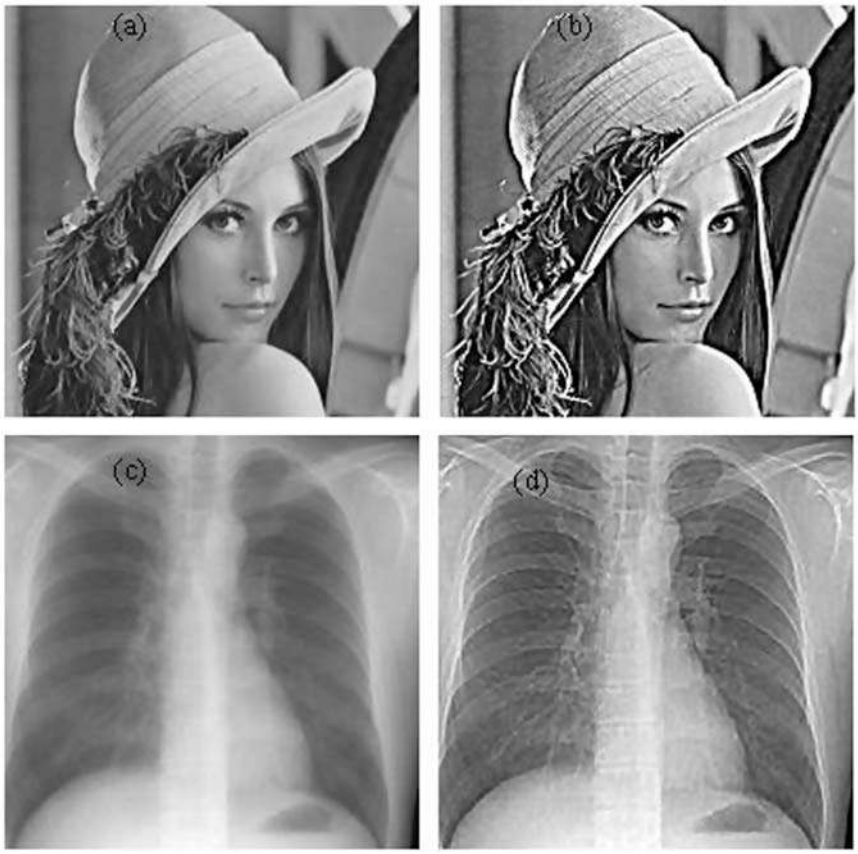


Fig. 10. Image enhancement of (a) a benchmark image and (c) an X-ray image by the proposed silicon retina model. The results are shown in (b) and (d).

In our proposed implementation scheme we have followed the linear resistive network in CMOS technology with transistors, as suggested by Mead (1989). In the proposed multi-layered VLSI the photo detectors occupy the first layer and for each of the exponential terms of Eq. (12) one layer is devoted. We have completed the software simulation of the circuit. By tuning the scaling factor represented by A_0, the amplitude of the delta function, it is possible to produce a light and shade information into the zero-crossing map of a natural image, in the sense of a half tone, apart from detecting simple luminance edges or produce direct image enhancement. Figure 10, for example shows image enhancement with the help of the proposed circuit.

Conclusion

We have presented a modified model of edge detection on the basis of recent physiological data on the retina. The model performs much better than its predecessors in edge detection and image enhancement. Moreover, the model provides explanation for some low-level illusion experiments, which were hitherto beyond the realm of edge detecting algorithms. The model also provides an opportunity to utilize noise in extracting some of the image features. We are, at present, in the process of hardware implementation of the model to facilitate real time image processing.

References

Adelson, E.H. (2000) In: Gazzaniga M. (Ed.), The New Cognitive Neurosciences (2nd ed.). MIT Press, Cambridge, MA, pp. 339–351.

Agostini, T. and Proffitt, D.R. (1993) Perceptual organization evokes simultaneous lightness contrast. Perception, 22: 263–272.

Bair, W. and Koch, C. (1991) An analog VLSI chip for finding edges from zero-crossings. Neural Inf. Process. Syst., 3: 399–405.

Barlow, H.B. (1953) Summation and inhibition in the frog's retina. J. Physiol., 119: 69–88.

Blakeslee, B. and McCourt, M.E. (1999) A multiscale spatial filtering account of the White effect, simultaneous brightness contrast and grating induction. Vis. Res., 39: 4361–4377.

DeValois, R.L. and DeValois, K.K. (1988) Spatial Vision. Oxford University Press, New York.

Douglass, J.K., Wilkens, L., Pantazelou, E. and Moss, F. (1993) Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance. Nature (London), 365: 337–340.

Enroth-Cugell, C. and Jakiela, H.G. (1980) Suppression of cat retinal ganglion cell responses by moving patterns. J. Physiol., 302: 49–72.

Enroth-Cugell, C. and Robson, J.G. (1966) The contrast sensitivity of the retinal ganglion cells of the cat. J. Physiol., 187: 517–552.

Fiorani, Jr., M., Rosa, M.G.P., Gattass, R. and Rocha-Miranda, C.E. (1992) Dynamic surrounds of receptive fields in primate striate cortex: a physiological basis for perceptual completion? Proc. Natl. Acad. Sci. U.S.A., 89: 8547–8551.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2005a) A possible mechanism of zero-crossing detection using the concept of extended classical receptive field of retinal ganglion cells. Biol. Cybern., 93: 1–5.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2005b). A new mechanism of "fuzzy" edge detection by multi-scale Gaussian filters in the light of human visual system. Proceedings of the Second Indian International Conference on Artificial Intelligence, ISBN 0-9727412-1-6, Pune, India, Dec. 20–22, pp. 3234–3251.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2006a) New vision tools from the comparative study of an "old" psychophysical model and a "modern" computational model. In: Ijspeert A.J., Masuzawa T. and Kusumoto S. (Eds.), Biologically Inspired Approaches to Advanced Information Technology, 3853. Springerpp. 236–251. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2006b) A possible explanation of the low-level brightness–contrast illusions in the light of an extended classical receptive field model of retinal ganglion cells. Biol. Cybern., 94: 89–96.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2006c) Proposing new methods in low-level vision from the Mach band illusion in retrospect. Pattern Recognit., 39: 726–730.

Ghosh, K., Sarkar, S. and Bhaumik, K. (2007) Understanding image structure from a new multi-scale representation of higher order derivative filters. Image Vis. Comput., doi:10.1016/j.imavis.2006.07.022. 25: 1228–1238.

Gilbert, C.D. and Wiesel, T.N. (1992) Receptive field dynamics in adult primary visual cortex. Nature, 356: 150–152.

Hochstein, S. and Shapley, R.M. (1976) Linear and nonlinear spatial subunits in Y cat retinal ganglion cells. J. Physiol., 262: 265–284.

Hochstein, S. and Spitzer, H. (1984) Zero-crossing detectors in primary visual cortex. Biol. Cybern., 51: 195–199.

Ikeda, H. and Wright, H.J. (1972) Functional organization of the periphery effect in retinal ganglion cells. Vis. Res., 12: 1857–1879.

Kaplan, E. and Benardete, E. (2001) The dynamics of primate retinal ganglion cells. Prog. Brain Res., 134: 1–18.

Kruger, J. (1984) The shift-effect enhances X- and suppresses Y-type response characteristics of cat retinal ganglion cells. Brain Res., 201: 71–84.

Kuffler, S.W. (1953) Discharge patterns and functional organization of mammalian retina. J. Neurophysiol., 16: 37–68.

Ma, S.D. and Li, B. (1998) Derivative computation by multiscale filters. Image Vis. Comput., 16: 43–53.

Marr, D. (1982) Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman and Company, New York.

Marr, D. and Hildreth, E. (1980) Theory of edge detection. Proc. R. Soc. Lond. B, 207: 187–217.

McIlwain, J.T. (1966) Some evidence concerning the physiological basis of the periphery effect in the cat's retina. Exp. Brain Res., 1: 265–271.

Mead, C.A. (1989) Analog VLSI and Neural Systems. Addison-Wesley, MA.

Mead, C.A. and Mahowald, M.A. (1988) A silicon model of early visual processing. Neural Netw., 1: 91–97.

Palmer, S.E. (1999) Vision Science: Photons to Phenomenology. MIT Press, Cambridge, MA.

Park, D., Kim, J., Kim, H., Park, J., Shin, J. and Lee, M. (2003) A foveated-structure CMOS retina chip for edge detection with local light adaptation. Sens. Actuators A Phys., 108: 75–80.

Passaglia, C.L., Enroth-Cugell, C. and Troy, J.B. (2001) Effects of remote stimulation on the mean firing rate of cat retinal ganglion cell. J. Neurosci., 21: 5794–5803.

Rodieck, R.W. (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vis. Res., 5: 583–601.

Sarkar, S., Ghosh, K. and Bhaumik, K. (2006). Digital image processing in the light of some newer aspects of low-level visual processing. IEEE Proceedings of International Conference on Engineering of Intelligent Systems, Islamabad, Pakistan, April 22–23, pp. 210–215.

Shou, T., Wang, W. and Yu, H. (2000) Orientation biased extended surround of the receptive field of cat retinal ganglion cells. Neuroscience, 98: 207–212.

Todorovic, D. (1997) Lightness and junctions. Perception, 26: 379–395.

Weiss, I. (1994) High-order differentiation filters that work. IEEE Trans. Pattern Anal. Machine Intell., 16: 734–739.

Wennekers, T. and Suder, K. (2004) Fitting of spatio-temporal receptive fields by sums of Gaussian components. Neurocomputing, 58–60: 929–934.
White, M. (1979) A new effect of pattern on perceived lightness. Perception, 8: 413–416.
White, M. (1981) The effect of the nature of the surround on the perceived lightness of gray bars within square-wave test gratings. Perception, 10: 215–230.
Wiesel, T.N. (1960) Receptive fields of ganglion cells in cat's retina. J. Physiol., 153: 583–594.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 16

Modeling the sub-cellular signaling pathways involved in reinforcement learning at the striatum

Shesharao M. Wanjerkhede and Raju S. Bapi*

Department of Computer and Information Sciences, University of Hyderabad, Gachibowli, Hyderabad, India

Abstract: A general discussion of various levels of models in computational neuroscience is presented. A detailed case study of modeling at the sub-cellular level is undertaken. The process of learning actions by reward or punishment is called 'Instrumental Conditioning' or 'Reinforcement Learning' (RL). Temporal difference learning (TDL) is a mathematical framework for RL. Houk et al. (1995) proposed a cellular signaling model for interaction of dopamine (DA) and glutamate activities at the striatum that forms the basis for TDL. In the model, glutamatergic input generates a membrane depolarization through *N*-methyl-D-aspartate (NMDA), α-amino-5-hydroxy-3-methyl-4-isoxazole propionic acid (AMPA), metabotropic glutamate receptors (mGluR), and opens calcium two plus ion (Ca^{2+}) channels resulting in the influx of Ca^{2+} into the dendritic spine. This raises the postsynaptic calcium concentration in the dendritic spine leading to the autophosphorylation of calcium/calmodulin-dependent protein kinase II (CaMKII). The timely arrival of the DA input at the neck of the spine head generates a cascade of reactions which then leads to the prolongation of long-term potentiation (LTP) generated by the autophosphorylation of CaMKII. Since no simulations were done so far to support this proposal, we undertook the task of computational verification of the model. During the simulations it was found that there was enhancement and prolongation of autophosphorylation of CaMKII. This result verifies Houk's proposal for LTP in the striatum. Our simulation results are generally in line with the known biological experimental data and also suggest predictions for future experimental verification.

Keywords: NMDA; AMPA; mGluR; RL; TDL; CaMKII; LTP; GENESIS; Kinetikit; signaling pathways; reinforcement learning; striatum

Computational models of neural systems

Computational modeling is important for understanding what nervous systems do and how they function. Modeling facilitates in understanding the relationship between different brain structures, their functions and provides a framework for linking different levels of biological organization. These models may be based on biophysical, anatomical, and physiological findings, but their primary purpose is to describe the process underlying the phenomenon. In general, models can be categorized into three broad classes — models of generic mechanism, models of specific neuronal systems, and models of generic operation of the network and the system (MacGregor, 1987). Here, generic refers to the fact that the mechanism pertains to or is appropriate to large classes or

*Corresponding author. Tel.: +91-40-23134014;
Fax: +91-40-23010780; E-mail: bapics@uohyd.ernet.in

DOI: 10.1016/S0079-6123(07)68016-9

groups as opposed to specific entities. Models of the first class deal with spike generation, dendritic trees, interneuronal communication, and biophysical/chemical levels of control. The second class deals with models of simpler invertebrate systems, lower level vertebrate systems, limbic systems, and neocortical systems. The last class pertains to the models of neural networks underlying cognitive operations or clinical models of disease (MacGregor, 1987).

Computational modeling of neural systems is also done at various levels such as the network level, the single cell level, and the biochemical level. Here we briefly discuss these models (belonging to different levels) and their utility. Network level is based on the network of neuronal connections between different populations of neurons from different regions. These network models use a simplified representation of neurons without incorporating all the geometrical details and channel properties. The network level models (e.g., the temporal difference (TD) models of dopamine (DA) cell activity during learning, where, DA cells respond to predicted events during classical conditioning) can explain learning based on the 'prediction error signal' (Brown et al., 1999; Suri, 2002; Pan et al., 2005).

Single cell models, on the other hand, can provide specific information about the dynamics of cellular responses to synaptic input. In a computational model, each neuron is divided into several compartments. In a neuron, the dendrite may be compartmentalized into the soma of the neuron and the axon as shown in Fig. 1.

Thus, the whole neuron may be represented as a collection of compartments. These compartments are then modeled with equations describing equivalent electrical circuits (Rall, 1959). The behavior of each compartment and its interaction with other neighboring compartments can be described by differential equations. However, each compartment must be made small enough to satisfy the electrical iso-potential property (Bower and Beeman, 1998). Thus a compartment represents a patch of gated membrane containing synaptic channels. The compartment model can help in the computation of both membrane voltage for any nonlinear input and time-dependent membrane properties (Bower and Beeman, 1998).

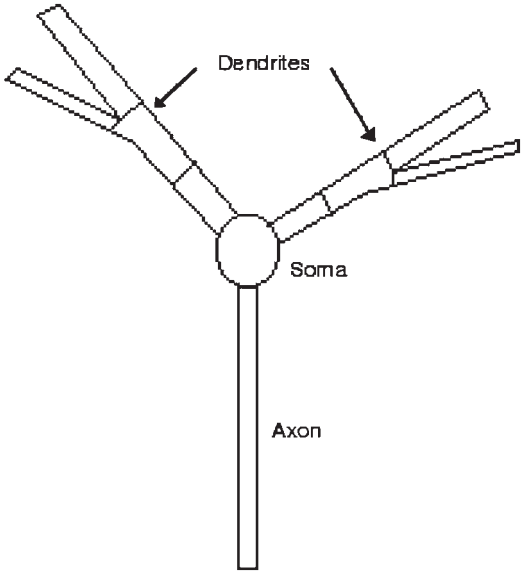


Fig. 1. Compartmental model of a neuron.

The third category, i.e., biochemical models, deals with the molecules involved in cellular signaling pathways. These molecules mediate biochemical reactions. The chemical reactions enable the processing and communication of information along signaling pathways. Models belonging to this category are taken up for detailed exposition in the subsequent sections. In particular, a detailed case study is presented on the signaling pathways and the protein molecules involved in the interaction of glutamate and DA neurotransmitters mediating reinforcement learning (RL). In the next section, an introduction to the methodology used for modeling signaling pathways is given.

Modeling the signaling pathways

Basics of biochemical equations

Second messenger pathways involving DA, the mechanisms of calcium influx into the spine head as well as other signaling pathways are modeled as

a series of bimolecular and enzymatic reactions. These are stoichiometric interactions between substrate molecules that form product molecules. A stoichiometric interaction is that reaction which specifies the number of each type of molecule required in the reaction. In biochemical systems, signal transmission occurs through two mechanisms: (a) protein–protein interactions and (b) enzymatic reactions. All binding reactions, e.g., a binding reaction in which A binds to B to form AB, is represented by the Eq. (4), given below. The reaction order is the number of simultaneously interacting molecules and biochemical reactions are classified as zero order, first order, second order, etc.

Zero order reaction

A reaction is of zero order when the rate of reaction is independent of the concentration of materials. Thus for a zero order reaction, rate of the reaction is a constant. When the limiting reactant is completely consumed, the reaction stops abruptly.

First order reaction

In the first order reaction, the rate is proportional to the concentration of a single reactant raised to the first power. In this reaction, a single substrate becomes the product. The rate constant is the rate (units per second) at which substrate becomes a product.

$$\mathrm{A} \underset{k_b}{\overset{k_f}{\Leftrightarrow}} \mathrm{B} \tag{1}$$

The ratio of backward rate constant to forward rate constant gives the concentration of substrates and products at equilibrium as in Eq. (1), i.e., when there is no net change in concentration over time.

$$\frac{k_b}{k_f} = \frac{[\mathrm{A}]}{[\mathrm{B}]} \tag{2}$$

In Eq. (2), the ratio k_b/k_f is represented as K_d and is termed as the equilibrium constant. These reactions are modeled with differential equations that express the rate of change of the quantity of the molecule with respect to time.

$$\frac{\mathrm{d}[\mathrm{B}]}{\mathrm{d}t} = k_f[\mathrm{A}] - k_b[\mathrm{B}] \tag{3}$$

The rate constants give the frequency of transitions between substrates and products.

Second order reaction

The rate of a second order reaction is proportional to either the concentration of a reactant squared or the product of concentrations of two reactants. In a second order reaction, each molecule of the product requires one molecule of one substrate and one molecule of another substrate (or two molecules of a single substrate).

$$\mathrm{A} + \mathrm{B} \underset{k_b}{\overset{k_f}{\Leftrightarrow}} \mathrm{AB} \tag{4}$$

The above reaction is analytically equivalent to the differential equation of the form shown in Eq. (5).

$$\frac{\mathrm{d}[\mathrm{AB}]}{\mathrm{d}t} = k_f[\mathrm{A}][\mathrm{B}] - k_b[\mathrm{AB}] \tag{5}$$

Another kind of biochemical reactions in which the two-substrate molecules are converted into two product molecules is of the form shown in Eq. (6).

$$\mathrm{A} + \mathrm{B} \underset{k_b}{\overset{k_f}{\Leftrightarrow}} \mathrm{C} + \mathrm{D} \tag{6}$$

where k_f and k_b are the rate constants for the forward and backward reactions.

The rate constants k_f and k_b are determined by the dissociation constant K_d and the time constant *t*. K_d is defined as $K_d = k_b/k_f$ and *t* indicates the reaction speed toward equilibrium. The above reaction is again analytically equivalent and can be

represented by the following differential equation shown in Eq. (7).

$$\frac{d[A]}{dt} = k_b[C][D] - k_f[A][B] \tag{7}$$

At equilibrium again these two rates are equal, i.e.,

$$k_f[A][B] = k_b[C][D] \tag{8}$$

$$\frac{k_f}{k_b} = \frac{[C][D]}{[A][B]} \tag{9}$$

The quantity on the right is the equilibrium constant, K_{eq}, so it follows that

$$K_{eq} = \frac{k_f}{k_b} \tag{10}$$

Third order reaction

Third and higher order reactions can be described in a similar manner. The order of the reaction is the number of substrate molecules required for the product.

$$A + 2B \underset{k_b}{\overset{k_f}{\Leftrightarrow}} AB \tag{11}$$

The differential equation describing the rate of change of product is the forward rate constant times all the substrates minus the backward rate constant times all the products (if more than one).

$$\frac{d[AB]}{dt} = k_f[A][B]^2 - k_b[A][B] \tag{12}$$

Enzymatic equations

The second type of reactions is enzymatic reactions. These are special type of reactions comprising two steps in which the enzyme is regenerated in the second step. The enzyme is not consumed and each enzyme molecule can make multiple product molecules. As shown in Eq. (13), enzyme reactions are generally modeled by the Michaelis–Menten formulation (Bhalla, 1998)

$$E + S \underset{k_b}{\overset{k_f}{\Leftrightarrow}} ES \overset{k_{cat}}{\rightarrow} P + E \tag{13}$$

where S, E, and P are substrate, enzyme, and product, respectively and ES is the complex product of enzyme E and substrate S. The maximum enzyme velocity V_{max} is defined as $V_{max} = k_{cat}[E]_{total, \text{ where } [E]total}$ is the total concentration of the enzyme and for $[S] \gg K_m$. The Michaelis constant K_m is defined as $K_m = (k_b + k_{cat})/k_f$. The above enzyme reaction can also be converted to a differential form for simulation purposes. The differential form of enzymatic reactions are as shown in the following Eq. (14).

$$\frac{d[ES]}{dt} = k_f[E][S] - (V_{max} + k_b)[ES] \tag{14}$$

$$\frac{d[P]}{dt} = V_{max}[ES] \tag{15}$$

Under Michaelis–Menten conditions, the above equations can be simplified (Blackwell and Kotaleski, 2002). These conditions are: (1) the quantity of the enzyme–substrate complex rapidly reaches equilibrium, (2) the backward reaction rate for the second step is zero, and (3) the amount of substrate is in excess (and so the enzyme quantity is rate limiting). Under the steady-state conditions, the equation describing the quantity of enzyme–substrate complex can be solved. Thus, in steady-state conditions

$$\frac{d[ES]}{dt} = 0 \tag{16}$$

$$k_f[E][S] - (V_{max} + k_b)[ES] = 0 \tag{17}$$

$$[E] = E_{tot} - [ES] \tag{18}$$

Substituting Eq. (18) in Eq. (17), we have the following

$$k_f(E_{tot} - [ES])[S] - (V_{max} + k_b)[ES] = 0 \tag{19}$$

$$k_f E_{tot}[S] - k_f[ES] \times [S] - (V_{max} + k_b)[ES] = 0 \quad (20)$$

$$k_f E_{tot}[S] = k_f[ES][S] + (V_{max} + k_b)[ES] \quad (21)$$

$$k_f E_{tot}[S] = (k_f[S] + (V_{max} + k_b))[ES] \quad (22)$$

$$[ES] = \frac{k_f E_{tot}[S]}{k_f[S] + (V_{max} + k_b)} \quad (23)$$

$$[ES] = \frac{E_{tot}[S]}{[S] + (V_{max} + k_b)/k_f} \quad (24)$$

Let

$$K_m = \frac{(V_{max} + k_b)}{k_f} \quad (25)$$

Substituting K_m from Eq. (25), Eq. (24) becomes

$$[ES] = \frac{E_{tot}[S]}{[S] + K_m} \quad (26)$$

where, K_m is the Michaelis–Menten constant. Substituting the value of [ES] from Eq. (26) in Eq. (15), we have

$$\frac{d[P]}{dt} = V_{max} \frac{E_{tot}[S]}{[S] + K_m} \quad (27)$$

Letting the velocity V_1 to be as shown in Eq. (28)

$$V_1 = \frac{V_{max}[S]}{[S] + K_m} \quad (28)$$

$$\frac{d[P]}{dt} = E_{tot} V_1 \quad (29)$$

Rearranging the terms in Eq. (28)

$$K_m = [S]\left[\frac{V_{max}}{V_1} - 1\right] \quad (30)$$

When

$$V_1 = \frac{V_{max}}{2} \quad (31)$$

then

$$K_m = [S] \quad (32)$$

When the observed reaction rate is half of the maximum possible reaction rate, the substrate concentration is numerically equal to the Michaelis–Menten constant. Thus the Michaelis constant K_m is defined as the substrate concentration at half-the-maximum velocity. The unit of K_m is μM. The ratio k_{cat}/K_m determines the relative rate of the reaction at low substrate concentrations and is known as the specificity constant.

The magnitude of K_m characterizes the enzyme kinematics. A small K_m indicates that the enzyme requires only a small amount of substrate to become saturated. Hence, the maximum velocity is reached at relatively low substrate concentrations. On the other hand, a large K_m indicates the need for high substrate concentrations to achieve maximum reaction velocity. The substrate with the lowest K_m upon which the enzyme acts as a catalyst is frequently assumed to be the enzyme's natural substrate, though this is not true for all enzymes (Blackwell and Kotaleski, 2002). The Michaelis–Menten equation has the form as shown in the following plot.

At high substrate concentrations, the rate is represented by the point C (Fig. 2), where the rate of the reaction is almost equal to V_{max}. If the Michaelis–Menten plot is extrapolated to very high substrate concentrations, then rate is equal to V_{max}. As [S] becomes very large, the velocity of the reaction will not increase indefinitely, but, for a fixed amount of $[E]_{total}$, it will reach a limiting value termed V_{max}, the maximal velocity. At lower substrate concentrations, such as at points A and B, the lower reaction velocities indicate that at any moment only a portion of the enzyme molecules are bound to the substrate. At the substrate concentration denoted by point B, exactly half the enzyme molecules are in an ES complex at that instant and the rate is exactly one half of V_{max}. At substrate concentrations near point A the rate is directly proportional to substrate concentration and the reaction rate is said to be of first order. To avoid curvilinear plots of enzyme catalyzed reactions, a better

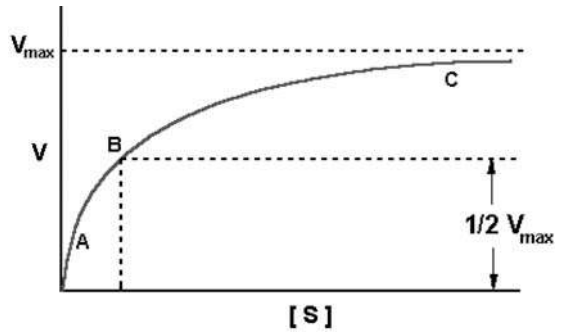


Fig. 2. Rate of reaction (V) versus substrate concentration [S]. A, B, and C represent points of low-, medium-, and high-substrate concentration.

method for determining the values of V_{max} and K_m was formulated by Hans Lineweaver and Dean Burk and is termed the Lineweaver–Burk or double reciprocal plot (Blackwell and Kotaleski, 2002).

$$\frac{1}{v} = \left[\frac{K_m}{V_{max}[S]} + \frac{1}{V_{max}}\right] \tag{33}$$

Plots of $1/v$ versus $1/[S]$ yield straight lines with slope of K_m/V_{max} and an intercept on the ordinate at $1/V_{max}$. Figure 3 shows slope $= K_m/V_{max}$ and intercept $= 1/V_{max}$. Substrate concentration is usually expressed in μM or mM and enzyme velocity in units of concentration of product per time. It is sometimes normalized to enzyme concentration, so the units are in terms of concentration of product per time per concentration of enzyme.

The Lineweaver–Burk transformation of the Michaelis–Menten equation is useful in the analysis of enzyme inhibition. A disadvantage of the Lineweaver–Burk plot is that most experimental measurements involve relatively high [S] and are thus crowded toward the left side of the plot. Since major part of clinical drug therapy is based on inhibiting the activity of enzymes, analysis of enzyme reactions using the tools described above has been fundamental to modern drug design in the pharmaceutical industry.

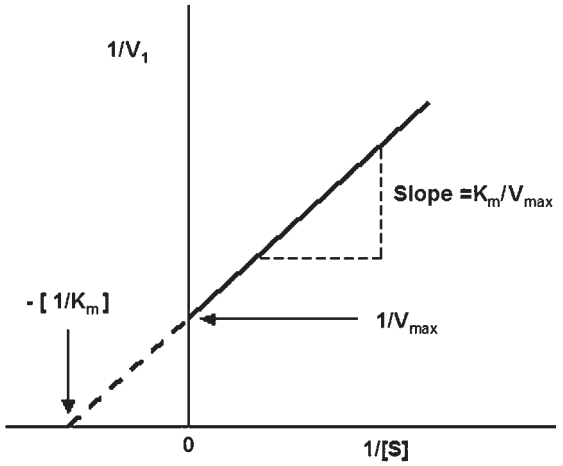


Fig. 3. Lineweaver–Burk plot.

Introduction to GENESIS/Kinetikit

GEneral NEural SImulation System (GENESIS) is a popular neural simulation software package (Bower and Beeman, 1998) for constructing biologically realistic neuronal simulations. Kinetikit is a graphical interface for biochemical computation that works as part of GENESIS (Bhalla, 2004). The sub-cellular level of modeling, such as the interacting biochemical signaling pathways that underlie the processes of synaptic transmission and channel activation, can easily be done with Kinetikit. We take up an example model as shown in Eq. (34) to illustrate the use of the Kinetikit package.

$$\mathrm{DA} + \mathrm{D1} \underset{k_b}{\overset{k_f}{\Leftrightarrow}} \mathrm{DAD1} \tag{34}$$

where DA stands for the dopamine neurotransmitter, D1 for the dopamine receptor and DAD1 is their complex. DA and D1 are the substrates and DAD1 is the product, respectively. A generic model Eq. (34) is implemented in Kinetikit and the results are shown in Figs. 4 and 5.

Figures 4 and 6 are dumpfiles of Kinetikit, and Fig. 4 represents Eq. (34) describing the activity of the D1 receptor. Figures 6 and 7 show the enzymatic reaction and the resultant graphs. With this introduction, we will now take up the detailed case study in the subsequent sections.

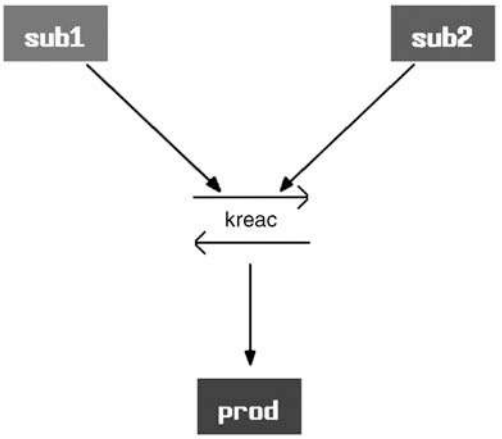


Fig. 4. Kinetikit representation of a typical biochemical reaction of the type shown in Eq. (34).

A case study of modeling at the sub-cellular level

Models of reinforcement learning

Reinforcement learning (RL) has received a lot of attention recently both in Neuroscience and in Machine Learning communities (Houk et al., 1995; Sutton and Barto, 1998; Wanjerkhede and Bapi, 2005). In the neuroscience literature, there has been an accumulation of data pertaining to the biochemical pathways that function when an organism learns actions that repeatedly lead to reward. The process of learning actions by reward or punishment is called 'instrumental conditioning' or 'reinforcement learning'. In the machine learning literature recent mathematical models such as TDL for learning by reinforcement have led to

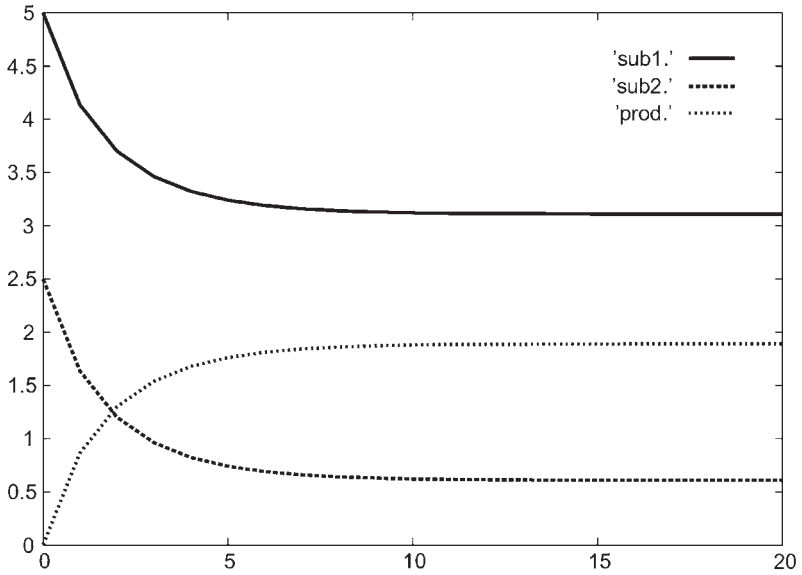


Fig. 5. The concentration–time profile for a typical biochemical reaction of the type shown in Eq. (34). X and Y axes represent time in seconds and concentrations in μM, respectively. The forward and backward rate constants taken are $0.1\,\mu M^{-1}s^{-1}$ and $0.1\,s^{-1}$, respectively.

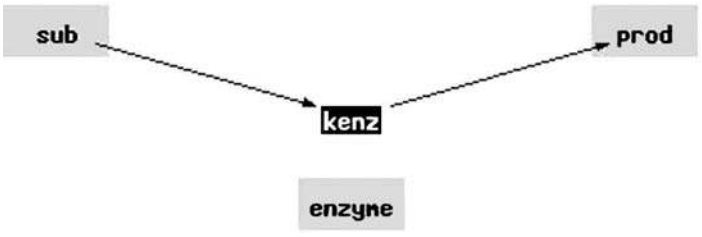


Fig. 6. Kkit dumpfile for the enzymatic reaction shown in Eq. (34).

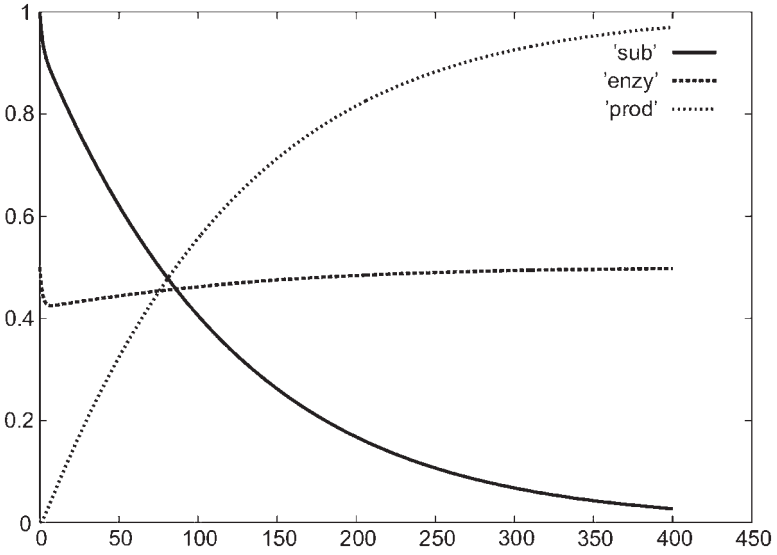


Fig. 7. Variation of the substrate, enzyme and the product with time. X and Y axes represent time in seconds and concentration in μM, respectively.

successful applications in various domains such as robotics, control, etc. (Sutton and Barto, 1998). Further, recently it has been shown that the prediction error signal computed in the TDL model is astonishingly similar to the single unit recordings made in the midbrain DA neurons of monkeys (Schultz, 1998). Thus, it appears that TDL might offer a biologically realistic mathematical framework for RL. In the TDL framework, it is assumed that learning of a chain of stimuli or actions is facilitated through DA modulation.

Several models have attempted to describe the DA cell behavior by a TDL model. The firing pattern of DA neurons in the basal ganglia appears to be a prediction error signal in TDL model. The firing pattern of DA neurons was also found to reflect information regarding the timing of delayed rewards, as could be seen by the precisely timed depression of DA firing when an expected reward was omitted. Houk et al. (1995) presented one of the first proposals of actor–critic models of the basal ganglia.

Sub-cellular signaling pathway model

We undertook the design and simulation of a biophysical model corresponding to Houk's theoretical proposal. This model is presented as a case study of modeling at the sub-cellular level. Figure 8 outlines the interaction between several intracellular signals in the spines of striatal neurons. In the block diagram, the signal transduction pathways are based on the literature (Houk et al., 1995). At the top left of Fig. 8, glutamate is released at the terminals of cortical afferents in response to conditioned stimulus, initiating a cascade of intracellular signaling reactions. The three types of receptors for glutamate namely, *N*-methyl-D-aspartate receptor (NMDAR), α-amino-5-hydroxy-3-methyl-4-isoxazole propionic acid receptor (AMPAR), and mGluR produce a mixture of membrane depolarization and increase the intracellular Ca^{2+}. This effect in turn leads to plasticity. Increase in intracellular Ca^{2+} activates calmodulin (CaM), which activates CaMKII, which then potentiates glutamate receptors via feedback mechanism as shown in Fig. 8. Bound CaM activates its substrates and initiates LTP (Houk et al., 1995). The production of LTP requires an additional reaction namely, the autophosphorylation of CaMKII (CaMKII_a). This keeps the molecule activated for minutes as opposed to a fraction of a second in the dephosphorylated state. The conversion of CaMKII into an autophosphorylated form

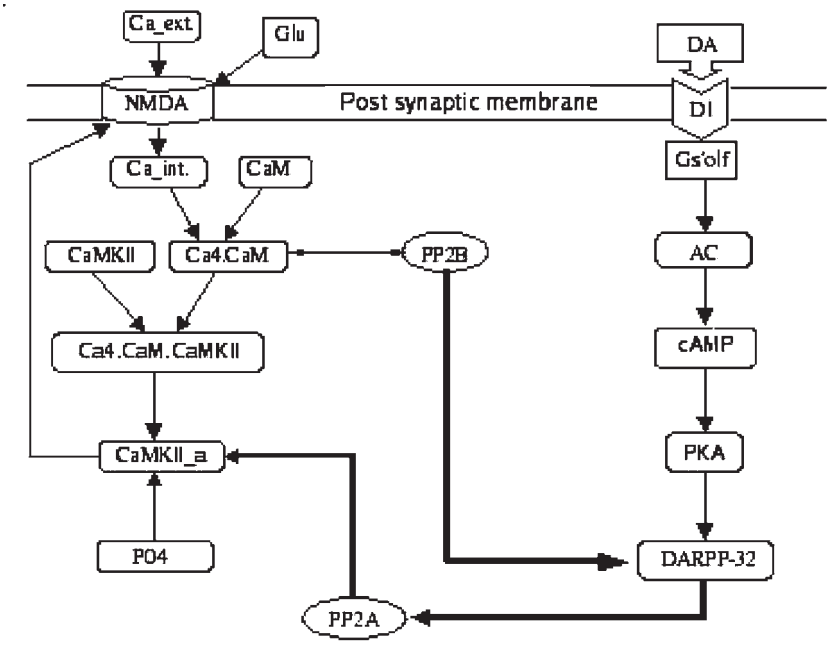


Fig. 8. Simplified diagram for the signaling pathways simulated for the interaction between the glutamate and the dopamine neurotransmitters. Glu, glutamate; NMDA, NMDA receptor; Ca_ext, Ca outside the spine head; Ca_in, Ca inside the spine head; CaM, calmodulin; Ca.CaM, calcium–calmodulin complex; Ca4.CaM, calcium–calmodulin complex with four calcium molecules; Ca4.CaM.CaMKII, a complex of bound form of CaMKII; AC, adenylyl cyclase; Gs, stimulatory G-proteins; Gs/olf, complex; cAMP, cyclic adenosine monophosphate; DARPP-32, Mr = 32,000 dopamine- and cyclic AMP-regulated phosphoprotein; PKA-active, active form of protein kinase A; PP2A, protein phosphatase 2A; PP2B, protein phosphatase 2B; CaMKII a, autonomous form of CaMKII.

depends on the arrival of a properly timed DA reinforcement signal. DA activates D1 receptors located on the spines, which activates cyclic adenosine 3′,5′-monophosphate (cAMP), active form of protein kinase A (PKA), and finally Mr = 32,000 dopamine- and cyclic AMP-regulated phosphoprotein (DARPP-32), as shown in Fig. 8. The activation of DARPP-32 involves other steps that are not shown in the block diagram. Activation of DARPP-32 inhibits phosphatase 2A and thus can disinhibit the autophosphorylation of CaMKII. But this disinhibition requires a properly timed arrival of the DA signal.

Houk et al. (1995) assumed that the proper timing of DA neuron firing and activation of DARPP-32 is appropriate to initiate LTP. Activated DARPP-32 inhibits protein phosphatase 2A (PP2A), thus removing inhibition of CaMKII autophosphorylation. This by itself will not phosphorylate CaMKII. Autophosphorylation occurs only under the condition that CaMKII is already activated by CaM.CaMKII binding. Another condition for DA being effective relates to protein phosphatase 2B (PP2B) activity. The PP2B is rapidly activated and inactivated by transients in CaM activation. When CaM is activated, PP2B blocks the ability of DA to activate DARPP-32. Thus the time course of CaM activity defines a prohibitive factor (Houk et al., 1995), and the time course of DA activity reflects the combination of the permissive and prohibitive factors (see Fig. 8).

The signaling pathways depicted in Fig. 8 are simulated in Kinetikit and the results are discussed below. The primary result from the simulation study is related to the effect of the affinity between the autophosphorylated form of CaMKII and NR2B (subunit of NMDA receptor). Our simulation results show that the amplitude of the autophosphorylated form of CaMKII with affinity

for NR2B was greater than the phosphorylated form without affinity. With affinity, it was above 12 μM (see Fig. 10d), whereas without affinity it was just below 1 μM for 40 molecules of NMDA receptor (graph not shown here).

Our simulations also establish that the phosphorylation of NMDA receptors by the autonomous form of CaMKII produces a larger number of open states of NMDAR (see Fig. 9). This in turn results in a large influx of Ca^{2+} into the spine head of spiny neurons in the striatum. This increase in calcium into the spine head increases the phosphorylation of CaMKII that is involved in learning and memory (Lisman and Goldring, 1988; Lisman, 1989; Lisman and Zhabotinsky, 2001). Nimchinsky et al. (2004) concluded that the number of open states of NMDAR is in the range of 30–60% of the total number of NMDAR molecules. Our simulation results are in agreement with this conclusion when the number of NMDAR molecules is in the range of 10–100. Beyond 100 NMDAR molecules, our results suggest that the concentration of open state NMDAR attains saturation (see Fig. 9).

As shown in Fig. 9, the concentration of the open state NMDAR attains saturation as the number of NMDAR molecules is increased. The total number of molecules of NMDAR that are in the open state is ~1.26 μM for 400 molecules of NMDAR. Interestingly for 40 molecules of NMDAR, the number of molecules in the open state is 22 and this corresponds to around 60% of the total number of molecules of NMDAR. This result is in close agreement with the proposal of Nimchinsky et al. (2004).

However, the bulk of evidence supports receptor saturation where all the glutamate receptors are saturated by glutamate molecules released at the synaptic cleft in the central nervous system. Glutamate reaches a high enough concentration to saturate all the postsynaptic receptors (Frerking and Wilson, 1996). But McAllister and Stevens (2000) have shown that NMDA receptors may not be saturated by a single vesicle release of glutamate. In agreement with this, our simulation results show that open state saturation is less than 60% of the total molecules and does not exceed 50 molecules (data not shown here).

In our simulation results we observed not only the prolongation of autophosphorylation activity as proposed by Houk et al. (1995) but also an enhancement in autophosphorylation of CaMKII (Fig. 10d). These results are in agreement with the proposal of Brown et al. (1999) which proposes that a mGluR-mediated delayed Ca^{2+} spike can be amplified and thus serve to transiently increase rather

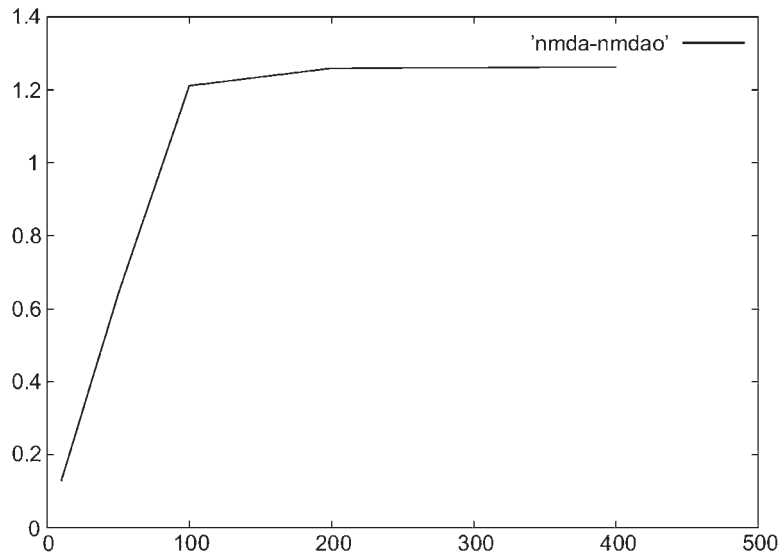


Fig. 9. Concentration of open State of NMDA receptors with the number of NMDA receptors.

than decrease striosomal cell activity. A Ca^{2+} spike combined with a phasic burst of DA acting on striosomal D1 receptor would also allow LTP in striosomal cells and could result in the potentiation of glutamate receptors (LTP) (Houk et al., 1995). These results suggest that the autophosphorylation of CaMKII depends not only on NMDA receptors but also on the affinity between autophosphorylated form of CaMKII to NR2B subtype of NMDAR.

Barria et al. (1997) have shown that the maximum phosphorylation of CaMKII at the T^{286} site occurred at 5 min following an LTP inducing stimulus. However, Holmes (2000) suggested that T^{286} phosphorylation occurs during the first few seconds. But, in our model simulations the phosphorylation of CaMKII occurred around 200 s after the stimulus onset at 0 s (see Fig. 10d). This result is different from both of the above suggestions of Barria et al. (1997) and Holmes (2000). From the simulation results it appears that the time course of phosphorylation of CaMKII depends on the number of NMDA receptor molecules.

For less number of NMDA receptor molecules the time course of phosphorylation is slower, and as the number of molecules increased the temporal profile shifts toward the stimulus onset time (Fig. 10d). The autophosphorylation of CaMKII occurs earlier for a large number of molecules (Fig. 10d). Thus our results suggest that the

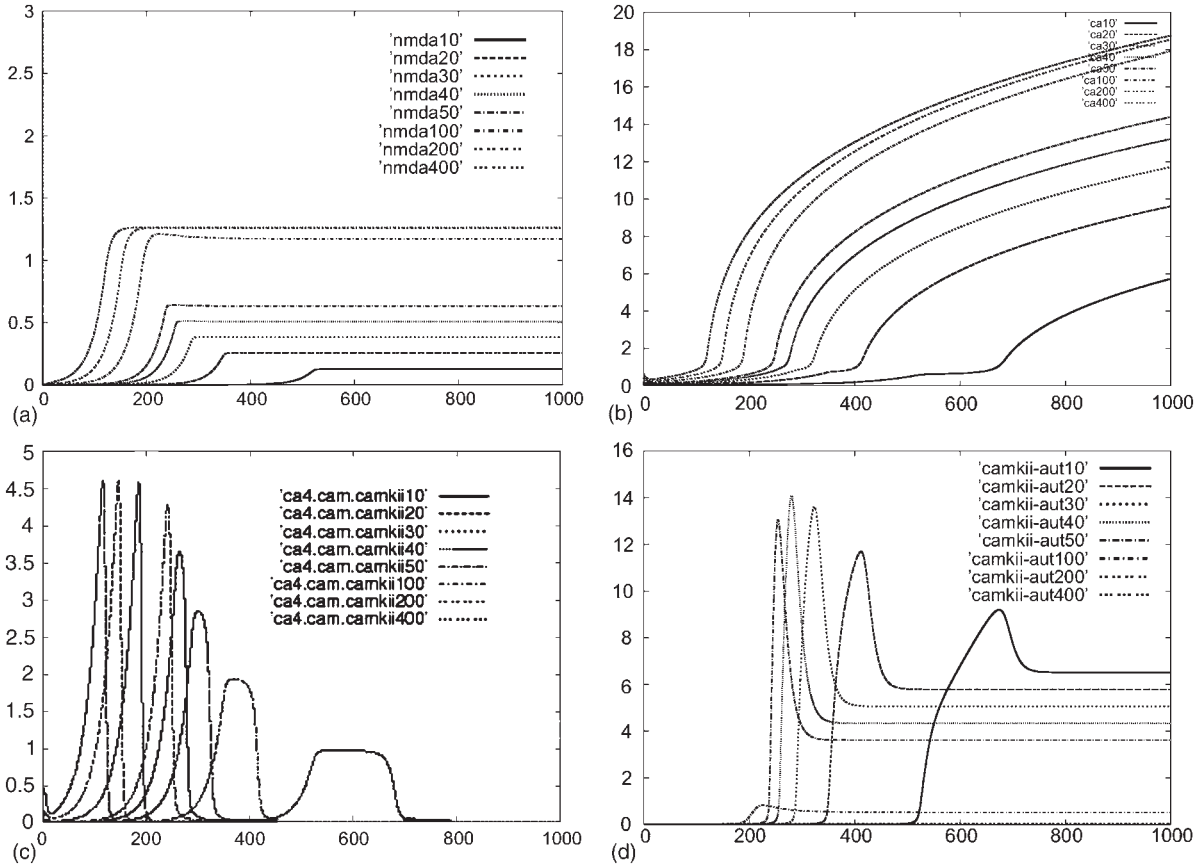


Fig. 10. Time course of autophosphorylation of CaMKII with affinity between autonomous form of CaMKII to NR2B subunit of NMDA receptor. (a) Number of open states of NMDAR at 10, 20, 40, 50,100, 200, and 400 molecules of NMDA receptor. (b) Time course of Ca^{2+} with the number of NMDA receptors in the spine head. (c) Activity of bound form of Ca4.CaM.CaMKII. (d) Activity of autophosphorylated form of CaMKII is maximal at 40 molecules of NMDAR.

autophosphorylation of CaMKII depends not only on NMDA receptors but also on the positive feedback mechanism from autophosphorylated form of CaMKII to NR2B subtype of NMDAR.

The role of autophosphorylation and the generation of autonomous activity have been the subject of multiple models. Earlier models have postulated that this form of enzyme could function as a molecular switch for learning and memory by retaining a memory of Ca^{2+} transients in the form of autonomous activity (Hudmon and Schulman, 2002). The autophosphorylated form of CaMKII, for all concentrations of NMDAR molecules, never goes to zero. As shown in Fig. 10d, it rises to a maximum, then decreases to a certain minimum value and remains constant thereafter for longer times, even for hundreds of minutes (data not shown here). We suggest that this change could function as a 'memory trace' (Otmakhov et al., 2004). This may form the basis for the 'eligibility trace' required for learning and memory necessary for an agent undergoing reinforcement learning (Houk et al., 1995).

The concept of 'eligibility trace condition' is invoked in delayed associative conditioning tasks where there is a need for availability of the memory trace of the antecedent signal at the time of a subsequent reward (Barto, 1995). Pan et al. (2005) suggested that a prolonged eligibility trace results from the circuits regulating DA cell activity in the brain. However, our simulation results point out that prolonged eligibility trace can result simply from intracellular biochemical signaling cascade instead of an extracellular circuit-level phenomenon. Further, our results related to the effect of feedback loop, the saturation phenomenon and the time course of LTP have implications for cellular models of LTP in other brain regions.

Figure 11 shows the autophosphorylation of CaMKII when the dopaminergic stimulus (St2) is applied first at the spine head and then the glutamatergic stimulus (St1) applied at the spine neck. This pairing of stimuli is shown in Fig. 11 for various delays. In the graph, the inter-stimulus interval (ISI) timings are shown as stisi1, stisi10, stisi100, etc. The results show that maximum LTP is more than 45 μM and extends for all delays of glutamatergic and dopaminergic stimuli.

Overall, the results indicate that Houk's theoretical proposal is valid for prolongation of LTP under specific conditions. The simulation results also show that autophosphorylation of CaMKII may form the basis for eligibility trace condition required in learning.

Conclusions

The aim of computational modeling is to elucidate the process underlying the phenomenon. Models of neural systems can be constructed either at the generic mechanism level, to understand a specific neural system or to understand the generic operation of a network/neural subsystem. In general, computational models of neural system can be grouped into three categories — network models, single cell models, and biochemical pathway models. We present here a detailed case study of a model of the sub-cellular signaling pathways participating in RL. In biochemical systems, signal transmission occurs through protein–protein interactions and enzymatic reactions. Through these reactions information is transduced and communicated along the signaling pathways. It is possible to construct first-order differential equation models of such reactions and simulate them using popular software packages such as GENESIS/Kinetikit. We undertook the design and simulation of a pathway model corresponding to Houk's theoretical proposal for the interaction of glutamate and DA to mediate RL in the striatum. We have taken the spine head as a single compartment for the biochemical signaling model in which various types of signaling molecules and signaling pathways are involved. The simulation results show that Houk's proposal is valid for prolongation of LTP and autophosphorylation of CaMKII may form the basis for eligibility trace condition required in learning.

Abbreviations

AMPAR	α-amino-5-hydroxy-3-methyl-4-isoxazole propionic acid receptor
cAMP	cyclic adenosine 3′,5′-monophosphate

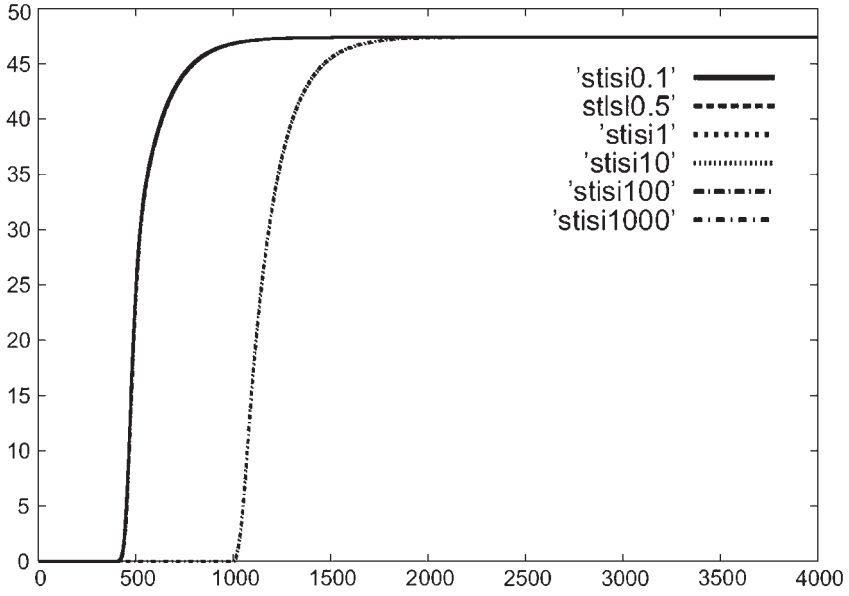


Fig. 11. Formation of LTP for different inter-stimulus intervals indicated as *stisi*, for different arrival times of the dopamine signal.

CaM	calmodulin
CaMKII	calcium/calmodulin-dependent protein kinase II
DA	dopamine, a neurotransmitter
DARPP-32	Mr = 32,000 dopamine- and cyclic AMP-regulated phosphoprotein
D1	dopamine receptor
LTP	long-term potentiation
mGluR	metabotropic glutamate receptor
NMDAR	*N*-methyl-D-aspartate receptor
PKA	active form of protein kinase A
PP2A	protein phosphatase 2A
PP2B	protein phosphatase 2B
RL	reinforcement learning
TDL	temporal difference learning

References

Barria, A., Muller, D., Derkach, V., Griffith, L.C. and Soderling, T.R. (1997) Regulatory phosphorylation of AMPA-type glutamate receptors by CaMKII during long term potentiation. Science, 276: 2042–2045.

Barto, A.G. (1995) Adaptive critics and the basal ganglia. In: Houk J.C., Davis J.L. and Beiser D.G. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 215–232.

Bhalla, U.S. (1998) The network within: signaling pathways. In: Bower J.M. and Beeman D. (Eds.), The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System (2nd ed.). Springer, New York, NY, pp. 169–190.

Bhalla, U.S. (2004) Signaling in small subcellular volume I. Stochastic and diffusion effects on individual pathways. Biophys. J., 87: 733–744.

Blackwell, K.T. and Kotaleski, H.J. (2002) Modeling the dynamics of second messenger pathways. In: Kotter R. (Ed.), Neuroscience Databases: A Practical Guide. Kluwer Academic Publishers, Boston, MA, pp. 63–80.

Bower, J.M. and Beeman, D. (1998) The Book of Genesis: Exploring Realistic Neural Models with the General Neural Simulation System (2nd ed.). Springer, New York, NY.

Brown, J., Bullock, D. and Grossberg, S. (1999) How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected reward cues. J. Neurosci., 19(23): 10502–10511.

Frerking, M. and Wilson, M. (1996) Saturation of postsynaptic receptors at central synapses. Curr. Opin. Neurobiol., 6: 395–403.

Holmes, W.R. (2000) Models of calmodulin trapping and cam kinase II activation in a dendritic spine. J. Comput. Neurosci., 8(1): 65–85.

Houk, J., Adams, J.L. and Barto, A.G. (1995) A model of how the basal ganglia generate and use neural signals that predict

reinforcement. In: Houk J., Davis J. and Beiser D. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 250–268.

Hudmon, A. and Schulman, H. (2002) Structure-function of the multifunctional Ca^{2+}/calmodulin-dependent protein kinase II. Biochem. J., 364: 593–611.

Lisman, J. (1989) A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. Proc. Natl. Acad. Sci. U.S.A., 86: 9574–9578.

Lisman, J.E. and Goldring, M.A. (1988) Feasibility of long term storage of graded information by the Ca^{2+}/calmodulin-dependent protein kinase molecules of the post-synaptic density. Proc. Natl. Acad. Sci. U.S.A., 85: 5320–5324.

Lisman, J.E. and Zhabotinsky, A.M. (2001) A model of synaptic memory: a CaMKII/PP1 switch that potentiates transmission by organizing an AMPA receptor anchoring assembly. Neuron, 31: 191–201.

MacGregor, R.J. (1987) Neural and Brain Modeling. Academic Press, San Diego, CA.

McAllister, A.K. and Stevens, C.F. (2000) Nonsaturation of AMPA and NMDA receptors at hippocampal synapses. Proc. Natl. Acad. Sci. U.S.A., 97: 6173–6178.

Nimchinsky, E.A., Yasuda, R., Oertner, T.G. and Svoboda, K. (2004) The number of glutamate receptors opened by synaptic stimulation in single hippocampal spines. J. Neurosci., 24(8): 2054–2064.

Otmakhov, N., Tao-Cheng, J.H., Carpenter, S., Asrican, B., Dosemeci, A., Reese, T.S. and Lisman, J. (2004) Persistent accumulation of calcium/calmodulin-dependent protein kinase II in dendritic spines after induction of NMDA receptor-dependent chemical long-term potentiation. J. Neurosci., 24: 9324–9331.

Pan, W.-X., Schmidt, R., Wickens, J.R. and Hyland, B.I. (2005) Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci., 25(26): 6235–6242.

Rall, W. (1959) Branching dendritic trees and motoneuron membrane resistivity. Exp. Neurol., 1: 491–527.

Schultz, W. (1998) Predictive reward signal of dopamine neurons. J. Neurophysiol., 80: 1–27.

Suri, R.E. (2002) TD models of reward predictive responses in dopamine neurons. Neural Netw., 15: 523–533.

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

Wanjerkhede, S.M. and Bapi, R.S. (2005) Role of presynaptic reuptake on dopamine modulation of cortico-striatal activity in TD learning. In: Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN '05), Vol. 4, pp. 2145–2149.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 17

Rhythmic structure of Hindi and English: new insights from a computational analysis

Tanusree Das, Latika Singh and Nandini C. Singh*

National Brain Research Centre, NH-8, Nainwal Mode, Manesar 122050, India

Abstract: Much information about speech rhythm is believed to be embedded in low frequency temporal modulations of the speech envelope. Using novel methods of spectral analysis we construct a spectro-temporal modulation spectrum and extract low frequency temporal modulations of spoken utterances to study the rhythmic structure of English and Hindi. The results of our spectral analysis reveal a narrower temporal bandwidth for Hindi as compared to English. We also calculate variability in syllable durations and find that variability in English is greater than Hindi. We relate temporal bandwidth of the modulation spectrum to variability in syllable duration and suggest that narrow bandwidth in the modulation spectrum implies less variability, whereas broad bandwidth implies greater variability in syllable duration. Our results also demonstrate that syllabic information is contained in low frequency temporal modulations of the speech envelope. Our results suggest that the modulation spectrum can be explored as a promising tool to study the temporal structure of language.

Keywords: language rhythms; modulation; spectrum; syllabicity; temporal structure; Hindi; English

Introduction

The pattern of syllable timing also called speech rhythm has been the basis for classification of spoken language. The rhythm class hypothesis proposed by Pike (1945) and Abercrombie (1967) classified spoken languages on the basis of two types of syllabic rhythm patterns: (a) stress-timed rhythm and (b) syllable-timed rhythm. According to this hypothesis, speakers of syllable-timed languages exhibit isochronous syllabic durations wherein syllabic units have equal duration whereas speakers of stress-timed languages exhibit isochronous stress durations wherein the duration from one stressed syllable to the next is believed to be roughly equal. This hypothesis categorizes Hindi as a syllable-timed language and English as stress-timed. Thus in this context, speakers of a syllable-timed language like Hindi would be expected to exhibit a regular pattern of syllabic durations whereas speakers of a stress-timed language, such as English would exhibit a variable pattern of syllable durations.

However this classification has been based purely on empirical evidence and since then there have been numerous attempts to find the acoustic correlates of speech rhythm with only partial success in most cases (see Ramus, 2002; Grabe and Low, 2002, for discussion). The most popular of these approaches has been the one proposed by

*Corresponding author. Tel.: +91-124-233 8920-28, Ext. 222; Fax: +91-124-233 8927/28; E-mail: nandini@nbrc.ac.in

DOI: 10.1016/S0079-6123(07)68017-0

Ramus et al. (1999) wherein the acoustic correlate is defined not in terms of syllabic units but in terms of vocalic and inter-vocalic intervals. This approach based on the variability of vowels relies on the feature that stress-timed languages allow vowel reduction unlike syllable-timed languages, which do not. Thus the major point of difference between the hypothesis proposed by Ramus et al. (1999) and the syllabic hypothesis has been in the choice of the parameter. In the syllabic hypothesis the acoustic correlate of rhythm is captured by the variability in syllabic durations whereas in the vocalic–intervocalic intervals hypothesis, it is the variability in the duration of vowels. However, the hypothesis based on the variability of the vocalic and intervocalic intervals was fairly successful and was able to demonstrate that for a group of eight languages, which consisted of languages from both rhythmic classes, the syllable-timed languages clustered in one group, whereas stress-timed languages clustered in a different group. It has however been unable to account for cross-linguistic control of speech rate, structure of corpus and still needs to be generalized to many more languages.

In this paper we present an approach based on the spectro-temporal features of the speech envelope to study syllabic rhythm. Timing is a fundamental part of sensory and motor processing (Buonomano, 2003). It is well known that speech contains information at different time scales (Rosen, 1992) and one of the primary features critical to the acquisition of language is the ability to categorize information at different time scales. Low-frequency temporal modulations of the order of hundreds of milliseconds are believed to contain information regarding syllabic rhythm and melody recognition, while higher frequency modulations in the range of tens of milliseconds encode information concerned with pitch and roughness (Steeneken and Houtgast, 1980; Rosen, 1992). Using methods of spectral analysis we develop a new computational tool called the modulation spectrum to study the distribution of energy in different amplitude fluctuations of the speech envelope (Singh and Theunissen, 2003). Since the fluctuations in the amplitude envelope can be viewed in both the time and the frequency domain, we obtain a joint distribution of the spectral and temporal modulations in the speech signal.

We study the spectro-temporal energy distribution of spoken utterances in English and Hindi. Since syllabic durations are typically between 100 and 500 ms we focus on the low temporal modulations and find that the temporal bandwidth for English is broader than Hindi. We also measure syllable durations for both English and Hindi and find greater variability in the syllable durations of English as compared to Hindi. We relate temporal bandwidth with variability in syllable duration and suggest that narrow temporal bandwidth indicates less variability in syllable durations as would be seen in a syllable-timed language. On the other hand, broader temporal bandwidth would indicate greater diversity in syllable duration events and would be a feature seen in stress-timed languages. We thereby also demonstrate that syllabic information is contained in low frequency temporal modulations.

The paper is organized as follows: In section ''Materials and Methods'' we describe the participants and database used in the current study. In section ''Methods of Analysis'', the methods of acoustic analysis have been discussed. In section ''Results'' we present the results and in section ''Discussion'' we discuss the conclusions and directions for future research.

Materials and methods

Speech recordings were obtained from five native adult monolingual speakers of Hindi (H) and English (E) respectively. Subjects were recorded directly on a laptop computer with a high quality microphone, at a sampling rate of 22,050 Hz. None of the subjects reported any form of speech or language disorder nor could any be detected during the recording procedure.

A common passage 'The North wind and the Sun' served as the reading material for all the speakers. The English and Hindi translations of this passage are available on the website of the International Phonetics Association. All the subjects were

recorded reading the passage in their native language in what they consider 'normal reading' for this text. The speakers were asked to practice reading the passage before the final narration. Each narrative lasted approximately 2 min.

Methods of analysis

Three kinds of analysis were carried out namely spectrographic analysis, modulation spectrum analysis (MSA) and analysis of syllable durations.

Spectrographic analysis

Broadband spectrograms for bisyllabic words from speakers of Hindi and English were obtained using a 300 Hz filter.

Modulation spectrum analysis

For the MSA (Singh and Theunissen, 2003), speech productions from all the speakers for each group were combined to form a single speech sample that was approximately 10 min long. The calculation of the probability distributions of the amplitude envelope of sounds in the spectrographic representation and their time–frequency correlations gives the modulation spectra. The modulation spectrum therefore provides information about the different spectro-temporal amplitude modulations in speech and the corresponding power associated with them. Based on the specific time scales, the temporal modulations provide both segmental and suprasegmental information, whereas the spectral modulations provide information about harmonic and formant structure. Because of the time–frequency trade off one cannot generate spectrographic representations that exhibit both high spectral and temporal modulations at the same time (Singh and Theunissen, 2003). Since, for the present study the focus was on studying temporal structure, a broadband filter namely 300 Hz was chosen for both languages. For the current study, the focus was to extract syllabic information at 100–300 ms, i.e. between 3 and 10 Hz.

Mean syllabic durations

Mean syllabic durations were also estimated for 8 bisyllabic words and 2 trisyllabic words in Hindi; and 4 bisyllabic words and 4 trisyllabic words in English for each speaker. Thus mean syllable durations for total of 42 words in for Hindi and 40 words in English was estimated. Words from both languages were manually segmented into syllables from spectrograms using sound-editing software, using both auditory and visual cues. Two professional linguists confirmed the segmentations. This was followed by analysis of syllabic durations using a program written in MATLAB.

Results

Spectrographic analysis

Representative wideband spectrograms from speakers of Hindi and English plotted in Fig. 1 show differences in segmentation patterns for both languages. The presence of energy fluctuations across a frequency spectrum at particular times in the spectrogram are called spectral modulations (ω_x), and temporal modulations (ω_t), are energy fluctuations at a particular frequency over time. Temporal modulations, which are of the order of 2–10 Hz are believed to encode syllabic rhythm.

Modulation spectrum analysis (MSA)

The spectral and temporal modulations defined above are obtained using the procedure defined in Singh and Theunissen (2003). Modulation spectra for English and Hindi were computed and have been plotted in Fig. 2a. The figure clearly exhibits differences in the distribution of spectro-temporal energy between English and Hindi. Since the current basis for rhythmic classification of language is based on timing, we focus only on the temporal modulations and plot the energy distribution at various temporal modulations in Fig. 2b. We define the temporal bandwidth of the modulation spectrum for Hindi (Bw(H)) and English (Bw(E)) as that value of the temporal modulation frequency at which the maximum

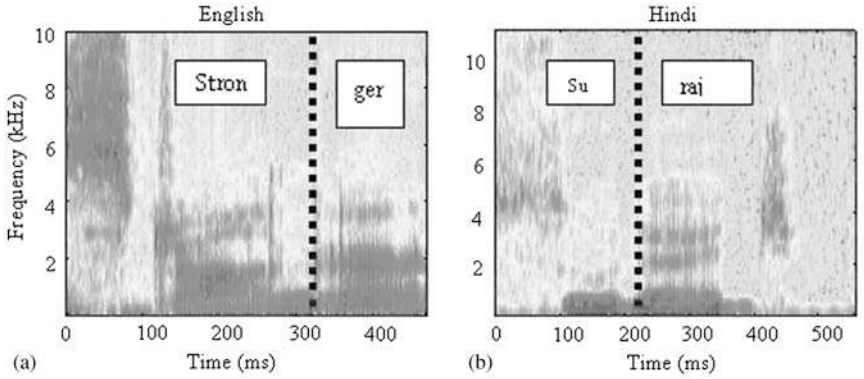


Fig. 1. Representative spectrograms showing the segmentation of uttered words into corresponding syllables. The dotted line in each spectrogram demarcates the syllables. (a) Utterance of the word 'stronger' by an English speaker. Dotted line segments syllables 'stron' and 'ger'. (b) Utterance of the Hindi word 'suraj' by a Hindi speaker with syllabic segmentation 'su' and 'raj'. (See Color Plate 17.1 in color plate section.)

power in the temporal modulations falls to 5% of its maximum value. As seen from Fig. 2a, b, the temporal bandwidth for Hindi was found to be 8.8 Hz while that for English was obtained as 11.8 Hz.

Mean syllable durations (MSDs)

We also study the syllable durations as obtained from spoken utterances for English and Hindi. As indicated earlier these are obtained from manual segmentation of bisyllabic words from both languages. The MSDs for English and Hindi (Fig. 3a) were measured and found to be to be $\sim$170 and $\sim$164 ms respectively (Fig. 3a). However, the co-efficient of variation or the variability (standard deviation/mean) in MSD for English and Hindi (Fig. 3b) were found to be 0.44 and 0.25 respectively, indicating smaller variability for Hindi as compared to English, as postulated by the rhythmic class hypothesis. A plot of the distribution of modulation frequencies as obtained from syllable durations for English and Hindi is plotted in Fig. 4. It shows a much smaller spread for Hindi as compared to English.

We define the inverse of the syllabic duration as the equivalent temporal modulation, and thus also calculate the probability distribution of temporal modulations from syllabic durations. The probability distribution for different temporal modulations from syllabic durations for English and Hindi are shown in Fig. 5. For both languages, the normalized probability distribution peaks around 5.5 Hz, which reflects the fact that the mean syllabic duration is similar, a feature also evident in Fig. 3a. The temporal bandwidth as estimated from the probability distribution provides a value of 10 Hz for English and 8.8 Hz for Hindi which does not differ significantly from the values obtained from the modulation spectrum analysis.

Discussion

This paper investigates the syllabic structure of spoken utterances in two languages, English and Hindi using two methods; one using manual calculations of syllable durations and the other by carrying out spectro-temporal analysis of the amplitude envelopes of the speech signal also called the modulation spectrum analysis. Our study shows that the rhythmic structure for Hindi is different from English.

The modulation spectrum analysis, which characterizes the spectral and temporal structure of spoken language in terms of the fluctuations of the amplitude envelope suggests two results. Firstly, it showed that the distribution of both temporal and spectral energy of Hindi is quite different from that in English. Secondly, the temporal bandwidth for

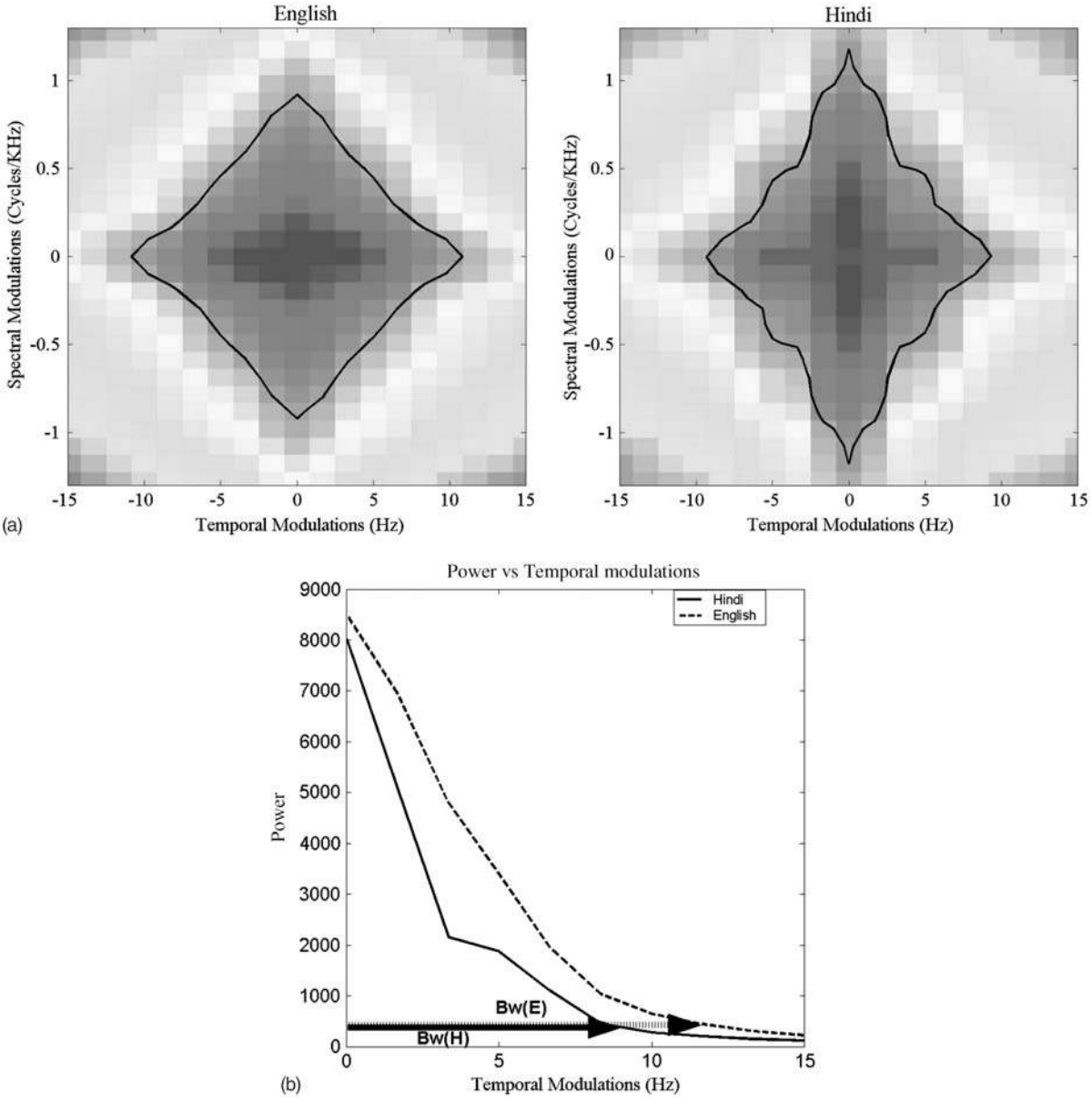


Fig. 2. (a) Modulation spectra for English and Hindi indicate differences in both spectral and temporal structure between the two languages. (b) The energy distribution for different temporal modulations for both English and Hindi, shows a narrower temporal bandwidth for Hindi (8.8 Hz) as compared to English (11 Hz). (See Color Plate 17.2 in color plate section.)

Hindi is narrower than English. Since Hindi has been classified as a syllable-timed language while English is stress-timed, our results are consistent with the rhythmic classification of languages proposed earlier. To compare the results of the MSA with standard procedure for studying syllabic structure of language, we also obtain syllable durations for both English and Hindi by manual demarcation of spoken language utterances. We find that the MSDs of Hindi and English are similar, 170 and 164 ms respectively. We also observe greater variability in the syllable durations

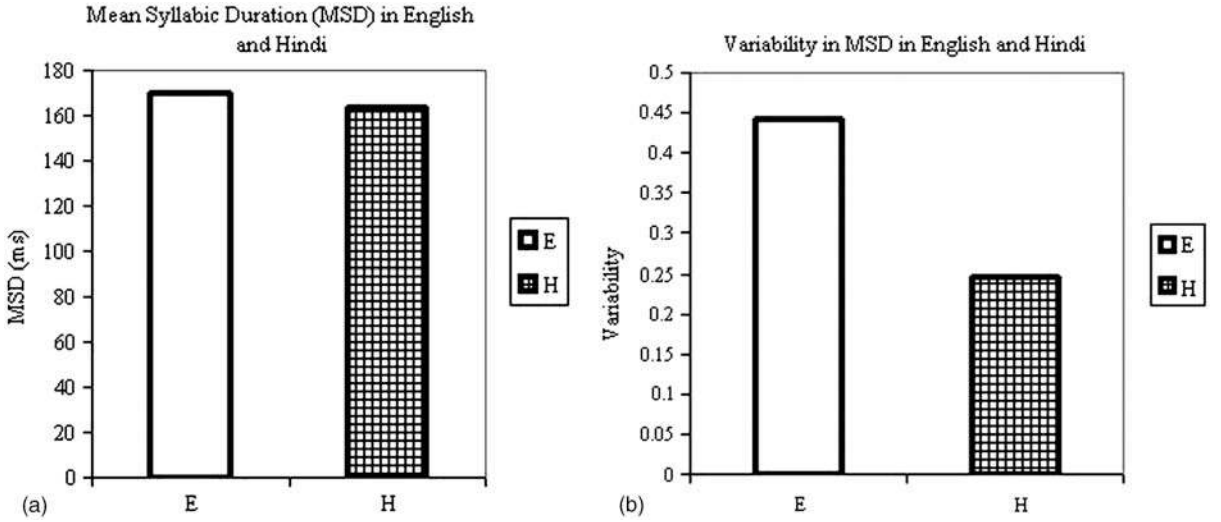


Fig. 3. (a) The plot compares the mean syllabic durations (in milliseconds) of English (E) and Hindi (H) monolinguals. (b) A comparison of variability (standard deviation/mean) in MSD between English and Hindi shows greater variability in syllable durations for English.

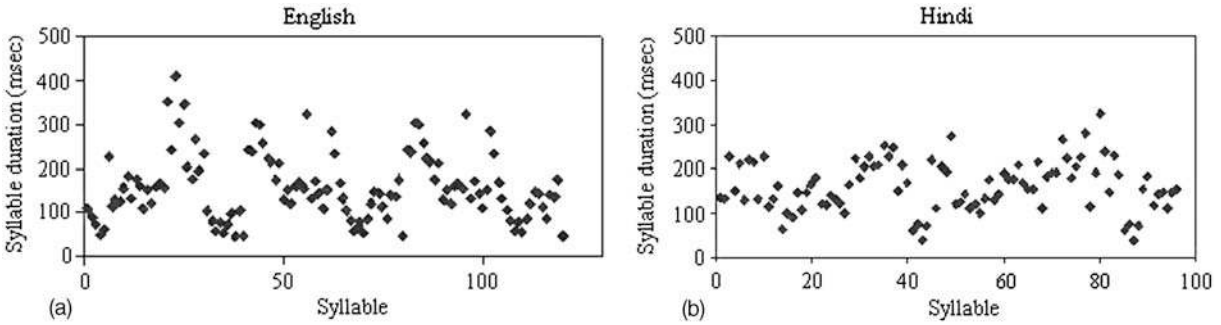


Fig. 4. Scatter plot of syllable durations for English and Hindi. Syllable durations for English are distributed over a wider range as compared to Hindi.

of English as compared to Hindi. We also obtain normalized probability distribution for syllable durations for both languages and observe a narrow bandwidth for Hindi as compared to English.

We therefore relate temporal bandwidth with variability in syllable duration and hypothesize that narrow temporal bandwidth suggests less variability in syllable durations while broader temporal bandwidth suggests greater variability in syllable durations events. Our results are in agreement with the rhythmic classification wherein Hindi has been classified as a syllable-timed language and English as a stress-timed language. We thereby also demonstrate that syllabic information is contained in low frequency temporal modulations.

Though, this paper demonstrates the use of the modulation spectrum to study the temporal structure of only two languages, the results nevertheless indicate that the modulation spectrum presents a promising tool to study syllabic rhythm in spoken language. Since it is based on the distribution of energy in different amplitude fluctuations of the speech envelope (Singh and Theunissen, 2003) in the spectrographic domain it provides a joint distribution of the spectral and temporal modulations in the speech signal. It therefore provides

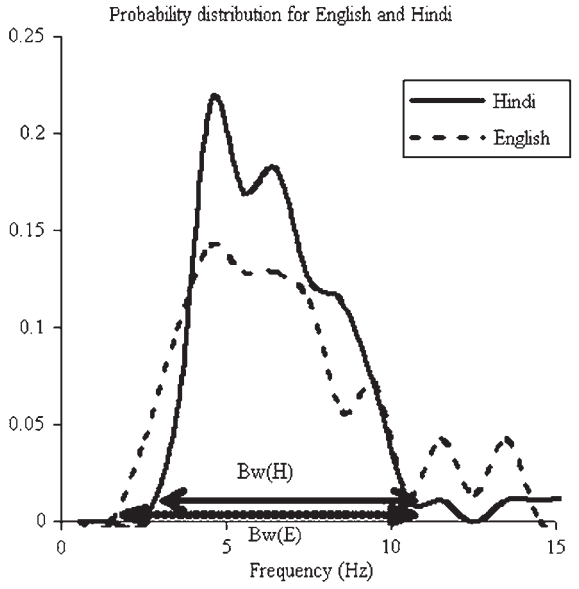


Fig. 5. Normalized probability distribution for equivalent temporal modulation frequency obtained from syllabic durations. The plot shows similar syllabic rates ($\sim$6 Hz) for both English and Hindi with a broader bandwidth for English as compared to Hindi. It also shows a narrower bandwidth for Hindi as compared to English.

insights into the distribution of both spectral and temporal energy in a syllable. The distribution of temporal energy in turn provides information about the underlying temporal structure of a language. The advantage of this analysis is that it can be used to study temporal structures in a large collection of speech samples and can thus be used to study language effects in a population as opposed to the manual demarcation of syllables, which is a tedious process.

Analysis of this kind also provides insights about the segmentation strategies that might be used by speakers of one category (syllable-timed) when attempting to learn languages belonging to a different category (stress-timed). Syllables are the phonological "building blocks" of words and are believed to influence the rhythm of a language, its prosody and its stress pattern. It is believed that speech rhythms organize human vocalizations into temporal events (Fox, 2000) making it possible to predict the rhythms of forthcoming speech events (Lehiste, 1977). As a result listeners tend to focus only on stressed syllables and are able to predict when these stresses might occur in forthcoming speech utterances, thereby removing the need for constant attention to any speech input. Work on infant speech perception (Nazzi et al., 1998) has shown that infants are capable of discriminating rhythmic classes of languages. These results suggest that the temporal properties of one's native language could provide the basis for speech segmentation strategies that would be employed during the perception of a foreign language. The difference in temporal properties of stress-timed and syllable-timed languages could reflect inherent constraints on the temporal processing mechanisms employed in both the perception and production of speech. This would be relevant and interesting for populations learning a second language. In particular an interesting point for study here would be the speech rhythms produced by bilinguals, especially if they spoke languages from both categories. There are hardly any studies that have explored such questions, but the MSA now provides a convenient procedure to initiate such investigations. Though the results are for only two languages, the choice of languages has been such that they belong to different rhythmic classes. More data is required to verify whether this trend is seen for other languages and whether the temporal bandwidth can successfully characterize languages into their respective rhythmic classes. While more data is being collected to verify this for various languages, our analysis suggests that the modulation spectrum presents a promising new approach to study language structure. In addition, this analysis also indicates that differences are seen in both spectral and temporal properties between languages, and therefore a better approach to study rhythmic properties of spoken language would be to define tools that account for both spectral and temporal differences between languages.

Acknowledgments

The authors would like to thank an anonymous reviewer whose comments and criticisms greatly improved the contents of the manuscript. This research was funded by the National Brain

Research Centre and a research grant from the Department of Information Technology, Government of India. The authors also wish to thank all the speakers who participated in the study.

References

Abercrombie, D. (1967) Elements of General Phonetics. Edinburgh University Press, Edinburgh.

Buonomano, D.V. (2003) Timing of neural responses in cortical organotypic slice. Proc. Natl. Acad. Sci. U.S.A., 100: 897–902.

Fox, A. (2000) Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals. Oxford University Press, Oxford.

Grabe, E. and Low, E.L. (2002) Durational variability in speech and the rhythm class hypothesis. In: Papers in Laboratory Phonology, Vol. 7, pp. 515–546.

Lehiste, I. (1977) Isochrony reconsidered. J. Phon., 5: 253–263.

Nazzi, T., Bertoncini, J. and Mehler, J. (1998) Language discrimination by newborns: towards an understanding of the role of rhythm. J. Exp. Psychol. Hum. Percept Perform., 24(3): 756–766.

Pike, K.L. (1945) The intonation of American English. University Press, Michigan.

Ramus, F. (2002) Acoustic correlates of linguistic rhythm: perspectives. Proceedings of Speech Prosody, Aix-en-Provence, pp. 115–120.

Ramus, F., Nespor, M. and Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. Cognition, 73: 265–292.

Rosen, S. (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci., 336: 367–373.

Singh, N.C. and Theunissen, F.E. (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am., 114: 3394–3411.

Steeneken, H.J. and Houtgast, T. (1980) A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am., 67: 318–326.

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 18

Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples

Thomas Metzinger[1,2,*]

[1]*Philosophisches Seminar, Johannes Gutenberg Universität, D-55099 Mainz, Germany*
[2]*Frankfurt Institute for Advanced Studies, D-60438 Frankfurt am Main, Germany*

Abstract: A concise sketch of the self-model theory of subjectivity (SMT; Metzinger, 2003a), aimed at empirical researchers. Discussion of some candidate mechanisms by which self-awareness could appear in a physically realized information-processing system like the brain, using empirical examples from various scientific disciplines. The paper introduces two core-concepts, the ''phenomenal self-model'' (PSM) and the ''phenomenal model of the intentionality relation'' (PMIR), developing a representationalist analysis of the conscious self and the emergence of a first-person perspective.

Keywords: consciousness; self-consciousness; first-person perspective; ownership; agency; self-model; phenomenal transparency; phantom limbs; robotics; rubber-hand illusion; out-of-body experience; evolution of tool-use

SMT: what is the self-model theory of subjectivity?

The goal of this chapter is to give a brief summary of the ''self-model theory of subjectivity'' (SMT) that is accessible to readers who are not professional philosophers.[1] Here, I will use a series of empirical examples from a number of different disciplines to illustrate some core ideas and to demonstrate the explanatory scope as well as the predictive power of SMT. The SMT is a philosophical theory about what it means to be a self. It is also a theory about what it means to say that mental states are ''subjective'' states and that a certain system has a ''phenomenal first-person perspective.'' One of the ontological claims of this theory is that the self is not a substance in the technical philosophical sense of something that could maintain its existence on its own, even if the body, the brain, or everything else disappeared. It is not an individual entity or a mysterious *thing* in the metaphysical sense. No such things as selves exist in the world: selves and subjects are not part of the irreducible constituents of reality. What does exist is the *experience* of being a self, as well as the diverse and constantly changing contents of self-consciousness. This is what philosophers mean when they talk about the ''phenomenal self'': the

*Corresponding author. Tel.: +49-6131-39-23279;
Fax: +49-6131-39-25141; E-mail: metzinger@uni-mainz.de

[1]A short Précis, which deliberately focuses on the conceptual skeleton and ignores bottom-up constraints, is freely available in an electronic version as Metzinger (2005a), at <www.psyche.cs.monash.edu.au/>. See also the *Scholarpedia* entry on "Self Models" at <www.scholarpedia.org>. On the monograph level, an early German language (and meanwhile outdated) version of this theory can be found in Metzinger (1993); for the most comprehensive formulation of the theory to date, see Metzinger (2003a). The standard procedure to learn more about the theory is to go to Section 8.2 in Metzinger (2003a), find the questions most relevant to one's personal interests and work one's way back, using the pointers given there and the index at the end.

DOI: 10.1016/S0079-6123(07)68018-2

way you *appear* to yourself, subjectively, consciously. Under SMT, this conscious experience of being a self is analyzed as the result of complex information-processing mechanisms and representational processes in the central nervous system. Of course, there are also higher-order, conceptually mediated forms of phenomenal self-consciousness that not only have neuronal, but also *social* correlates.[2] This theory, however, begins by focusing on the minimal representational and functional properties that a naturally evolved information-processing system — such as *Homo sapiens* — has to have in order to later satisfy the constraints for realizing these higher order forms of self-consciousness. As most philosophers today would agree, the real problem lies in first understanding the simplest and most elementary form of our target phenomenon. This is the non-conceptual, prereflective and prelinguistic layer in self-consciousness. Therefore, the first question we will have to answer is this: What are minimally sufficient conditions for the emergence of a conscious self?

The self-model theory assumes that the properties in question are representational and functional brain properties. In other words, the psychological property that allows us to become a person in the first place is analyzed with the help of concepts from *subpersonal* levels of description. In philosophy of mind, this type of approach is sometimes called a "strategy of naturalization": a complex and hard-to-understand phenomenon — such as the emergence of phenomenal consciousness and a subjective, inward perspective — is conceptually analyzed in such a way as to make it empirically tractable. By reformulating classical problems from their own discipline, naturalist philosophers try to open them for interdisciplinary investigations and scientific research programs, for instance in the cognitive and neurosciences. These philosophers do not endorse naturalism and reductionism as part of a scientistic ideology; instead, they see them as a rational research strategy. For instance, if it should turn out — as many people believe (see for instance Nagel, 1986, especially Chapter 4, which is also discussed in Metzinger, 1995a) — that there is something about human self-consciousness that lies outside the reach of the natural sciences *in principle*, they would be satisfied with this finding as well. They would have achieved exactly what they set out to do in the first place: they would now have what philosophers like to call "epistemic progress." This type of progress could mean being able to describe, in a much more precise and fine-grained manner and with a historically unprecedented degree of conceptual clarity, *why exactly* science is unable to provide satisfying answers to certain questions, even in principle. Therefore, the most serious and respectable philosophical anti-naturalists will typically also be the ones who show the profoundest interest in recent empirical findings. Naturalism and reductionism are not ideologies or potential new substitutes for religion. It is exactly the anti-naturalist and exactly the anti-reductionist who will have the strongest ambition to make their philosophical case convincingly, in an empirically informed way.

[2]I analyzed the relation between conceptual and non-conceptual contents of self-consciousness in detail in Metzinger (2003c); Metzinger (2003b) is an earlier German version of this text. A hypothesis on the role of the unconscious self-model in the development of non-conceptually mediated forms of social cognition is formulated in Metzinger and Gallese (2003).

Step one: what exactly is the problem?

What we like to call "the self" in folk-psychological contexts is the phenomenal self: that aspect of self-consciousness that is immediately given in subjective experience, as the content of phenomenal experience. The phenomenal self may well be the most interesting form of phenomenal content. It endows our phenomenal space with two particularly fascinating *structural* features: centeredness and perspectivalness. As long as a phenomenal self exists, our consciousness is centered and bound to what philosophers call a "first-person perspective." States inside this center of consciousness are experienced as *my own* states, because they are endowed with a sense of ownership that is prior to language or conceptual thought. In all of my conscious experiences and actions, I engage in constantly changing relations with the environment and with my own mental

states. I experience myself as being *directed* — towards perceptual objects, other human beings, or the contents of my own mental states and concepts. This process gives rise to a subjective inner perspective. The fact that I have such an inner perspective, in turn, is cognitively available to me.[3] In other words, what probably distinguishes human beings from most other animals is that we not only *have* a subjectively experienced inner perspective, but can also consciously conceptualize ourselves *as beings that have such an inner perspective*.

The first problem, however, is that we are not exactly sure what we mean when we talk about these questions in this way. It is not just that we are unable to define concepts like "I", "self", or "subject". The real problem is that these concepts often do not seem to refer to observable objects in the world. Therefore, the first thing we have to understand is how certain structural features of our inner experience determine the way we *use* these concepts. In order to analyze the logic of ascribing psychological properties to ourselves and to understand what these concepts actually refer to, we must first investigate the representational deep structure of conscious experience itself. Three higher order phenomenal properties are particularly interesting in this context:

- "Mineness": This is a higher order property of particular forms of phenomenal content. It is an immediately given, non-conceptual sense of ownership. Here are some examples of how we try to refer to this phenomenal property in folk-psychological discourse, using everyday language: "Subjectively, *my* leg is always experienced as being a part of *me*"; "*My* thoughts and feelings are always experienced as part of *my own* consciousness"; "My volitional acts are always initiated *by myself*."
- "Selfhood": This experientially untranscendable feeling of being a self is the essence, the phenomenal core property we are looking for. Again, a few brief examples can illustrate how we refer to this highly salient feature of our inner experience from the outside, using linguistic tools: "I am *someone*"; "I experience myself as *identical* across time"; "The contents of my self-consciousness form a coherent *whole*"; "Without having the need to engage in any prior cognitive and reflexive operations I am always intimately familiar with the contents of my self-consciousness."
- "Perspectivalness": In the context discussed here, perspectivalness is the dominant structural feature of phenomenal space as a whole: it is centered in an acting and experiencing subject, a self that engages in constantly changing relationships with itself and the world. Examples include: "My world has a fixed center, and *I* am this center"; "Being conscious means having an *individual first-person perspective*"; "In experiencing persons and objects in the world as well as my own mental states, I am always bound to this inward perspective — I am its origin."

The next step consists in a representational and functional analysis of these target properties. We must ask: What functional and representational properties does an information-processing system have to have in order to instantiate the *phenomenal* property in question? Which of these properties are sufficient, and are any of them strictly necessary? What *exactly* does it mean for such a system to experience the world as well as its own mental states from a first-person perspective? What we need is a consistent conceptual background that is sufficiently flexible to continually integrate new empirical findings and at the same is capable of taking the wealth, the heterogeneity, and the subtlety of phenomenal experience into account. Obviously, this is not an easy task. I will now briefly try to sketch the outlines of such a conceptual framework in the remaining five steps.

[3]For a first introduction to the problem of cognitive self-reference as a potential difficulty for philosophical naturalism, see Baker (1998). See also Metzinger (2003a) (Section 6.4.4) and especially Metzinger (2003c). An interesting and lucid criticism of my own account of the cognitive first-person perspective is Baker (2007).

Step two: the self-model

Step two consists in the introduction of a new theoretical entity: the phenomenal self-model (PSM). It is the most important part of the representational basis for instantiating the relevant phenomenal properties (Cummins, 1983). What is a mental "representation"? A representational state, for instance in the brain, is a state that has a certain *content*, because it is directed at something in the world. The brain-state is the physical carrier; the content is the meaning of this state. An inner representation is *about* something: having a correct representation implies *reference*. A representational state often functions as a placeholder for something external, the referent; it represents because it "stands in" for something else. However, this "something" can also be a past event, a potential future outcome, or even a mere possibility — in such cases, we speak of representations as *simulations*. They simulate merely *possible* states of affairs; they represent a possibility, not an actuality. SMT is predominantly a representational theory of consciousness, because it analyzes conscious states as representational states and conscious contents as representational contents.

One of our key questions was: Which set of minimally sufficient *representational* properties does a system have to develop in order to possess the relevant target properties? This is our first, preliminary answer: the system needs a coherent self-representation, a consistent internal model of itself as a whole. In our case, the self-model is an episodically active representational entity whose content is determined by the system's very own properties. Whenever such a self-representation is needed to regulate the system's interactions with the environment, it is transiently activated — for instance in the morning, when we wake up. According to SMT, what happens when you wake up in the morning — when you first *come to yourself* — is that the biological organism, which you are, boots up its PSM: it activates the conscious self-model.

In other words, what we need is a comprehensive theory of the self-model of *Homo sapiens*.[4] Personally, I assume that this will be a predominately neurocomputational theory (see for instance, Churchland, 1989). This means that there is not only a true representational and functional description of the human self-model, but also a true neurobiological description — for instance in terms of being a widely distributed, complex activation pattern in the brain (Damasio, 1999). The PSM is exactly that part of the *mental* self-model that is currently embedded in a highest order integrated structure, the global model of the world (Yates, 1975; Baars, 1988; for a detailed analysis of the criteria for distinguishing different degrees of consciousness, see Metzinger, 2003a, Chapter 3). In other words, certain parts of the self-model can be unconscious and functionally active at the same time. The PSM is a coherent multimodal structure that probably depends on a partially innate, "hard-wired" model of the system's spatial properties. (More about this in the second example; see also the fifth section of O'Shaughnessy, 1995 and his use of the concept of a "*long-term body image*"; and Metzinger, 1993, 1996, 1997; Damasio, 1994, 1999). This type of analysis treats the self-conscious human being as a special type of information-processing system: the subjectively experienced content of the phenomenal self is the representational content of a currently active, dynamic data structure in the system's central nervous system.

Aside from the representational level of description, one can also develop a *functional* analysis of the self-model. Whereas representational states are individuated by their content, a functional state is conceptually characterized by its *causal role*: the causal relationships it bears to input states, output states, and other internal states. An active self-model then can be seen as a subpersonal functional state: a set of causal relations of varying complexity that may or may not be realized at a

[4]The methodological core of psychology — insofar as I may venture this type of metatheoretical observation from my standpoint as a philosophical outsider — can now be analyzed in a fresh and fruitful way. Psychology is *self-model research.* It is the scientific discipline that focuses on the representational content, the functional profile and the neurobiological realization of the human self-model, including its evolutionary history and its necessary social correlates.

given point in time. Since this functional state is realized by a concrete neurobiological state, it plays a certain causal role for the system. For instance, it can be an element in an information-processing account. The perspective of classic cognitive science can help illustrate this point: the self-model is a *transient computational module* that is episodically activated by the system in order to control its interactions with the environment. In other words, what happens when you wake up in the morning, i.e., when the system that you are "comes to itself," is that this transient computational module is activated — the moment of "waking up" is exactly the moment in which this new instrument of intelligent information-processing emerges in your brain. It does so because you now need a conscious self-model in order to achieve sensorimotor integration, generate complex, flexible and adaptive behavior, and attend to and control your body *as a whole*. The development of ever more efficient self-models as a new form of "virtual organ" — and this point should not be overlooked — is also a precondition for the emergence of complex societies. Plastic and ever more complex self-models not only allowed somatosensory, perceptual, and cognitive functions to be continuously optimized, but also made the development of social cognition and cooperative behavior possible. The most prominent example, of course, is the human mirror system, a part of our unconscious self-model that *resonates* with the self-models of other agents in the environment through a complex process of motor-emulation — of "embodied simulation," as Vittorio Gallese (2005) aptly puts it — e.g., whenever we observe goal-directed behavior in our environment. Such mutually coupled self-models, in turn, are the fundamental representational resource for taking another person's perspective, for empathy and the sense of responsibility, but also for metacognitive achievements like the development of a *concept* of self and a *theory of mind* (see for instance, Bischof-Köhler, 1996, 1989; on the possible neurobiological correlates of these basic social skills, which fit very well into the framework sketched above, see Gallese and Goldman, 1998; Metzinger and Gallese, 2003).

The obvious fact that the development of our self-model has a long biological, evolutionary, and (a somewhat shorter) social history can now be accounted for by introducing a *teleofunctionalist background assumption*, as it is often called in philosophy of mind (see for instance Millikan, 1984, 1993; Bieri, 1987; Dennett, 1987; Dretske, 1988, 1998; Lycan, 1996). The development and activation of this computational module plays a role *for* the system: the functional self-model possesses a true evolutionary description, i.e., it was a weapon that was invented and continuously optimized in the course of a "cognitive arms race" (Clark, 1989, p. 61). The functional basis for instantiating the phenomenal first-person perspective can be seen as a specific cognitive achievement: the ability to use a *centered* representational space. In other words, phenomenal subjectivity (the development of a subsymbolic, non-conceptual first-person perspective) is a property that is only instantiated when the respective system activates a coherent self-model and integrates it into its global world-model.

The existence of a stable self-model allows for the development of what philosophers call the "perspectivalness of consciousness": the existence of a single, coherent, and temporally stable reality-model that is representationally centered in a single, coherent, and temporally stable phenomenal subject, a model of the system *in the act of experiencing* (see last section). This structural feature of the global representational space then leads to the episodic instantiation of a temporally extended, non-conceptual first-person perspective. If this global representational property is lost, this also changes the phenomenology and leads to the emergence of different neuropsychological deficits or altered states of consciousness. Some readers may have the impression that all of this is extremely abstract. A self-model, however, is not at all abstract — it is entirely concrete. A first, now classic, example will help demonstrate what — among many other things — I actually mean with the concept of a "self-model."

In a series of fascinating experiments, in which he used mirrors to induce synesthesia and kinesthetic illusions in phantom limbs, Indian neuropsychologist Vilayanur Ramachandran

demonstrated the existence of the human self-model (see Ramachandran and Rogers-Ramachandran, 1996; a popular account can be found in Ramachandran and Blakeslee, 1998, 46ff. The figure was published courtesy of Ramachandran). Phantom limbs are subjectively experienced limbs that typically appear after the accidental loss of an arm or a hand or after surgical amputation. In some cases, for instance following a non-traumatic amputation performed by a surgeon, patients have the subjective impression of being able to control and move their phantom limb at will. The neurofunctional correlate of this phenomenal configuration could consist in the fact that motor commands, which are generated in the motor cortex, continue to be monitored by parts of the parietal lobe and — since there is no contradictory feedback from the amputated limb — are subsequently integrated into the part of the self-model that serves as a *motor emulator* (related ideas are discussed by Grush 1997, 1998, p. 174; see also Ramachandran and Rogers-Ramachandran, 1996, p. 378). In other cases, the subjective experience of being able to move and control the phantom limb is lost. These alternative configurations may result from preamputational paralysis following peripheral nerve damage or from prolonged loss of proprioceptive and kinesthetic "feedback" that could confirm the occurrence of movement. On the phenomenological level of description, this may result in a paralyzed phantom limb.

Ramachandran and colleagues constructed a "virtual reality box" by vertically inserting a mirror in a cardboard box from which the lid had been removed. The patient, who had been suffering from a paralyzed phantom limb for many years, was then told to insert both his real arm and his phantom into two holes that had been cut in the front side of the box. Next, the patient was asked to observe his healthy hand in the mirror. On the level of visual input, this generated the illusion of seeing both hands, even though he was actually only seeing the reflection of his healthy hand in the mirror. So, what happened to the content of the PSM when the patient was asked to execute symmetrical hand movements on both sides? This is how Ramachandran describes the typical outcome of the experiment

> I asked Philip to place his right hand on the right side of the mirror in the box and imagine that his left hand (the phantom) was on the left side. "I want you to move your right and left arm simultaneously," I instructed.
>
> "Oh, I can't do that," said Philip. "I can move my right arm but my left arm is frozen. Every morning, when I get up, I try to move my phantom because it's in this funny position and I feel that moving it might help relieve the pain." But, he said looking down at his invisible arm, "I never have been able to generate a flicker of movement in it."
>
> "Okay, Philip, but try anyway."
>
> Philip rotated his body, shifting his shoulder, to "insert" his lifeless phantom into the box. Then he put his right hand on the other side of the mirror and attempted to make synchronous movements. As he gazed into the mirror, he gasped and then cried out, "Oh, my God! Oh, my God, doctor! This is unbelievable. It's mind-boggling!" He was jumping up and down like a kid. "My left arm is plugged in again. It's as if I'm in the past. All these memories from years ago are flooding back into my mind. I can move my arm again. I can feel my elbow moving, my wrist moving. It's all moving again."
>
> After he calmed down a little I said, "Okay, Philip, now close your eyes."
>
> "Oh, my," he said, clearly disappointed. "It's frozen again. I feel my right hand moving, but there's no movement in the phantom."

Fig. 1. Mirror-induced synesthesia. Making part of a hallucinated self available for conscious action control by installing a virtual source of visual feedback. (Picture courtesy of Vilayanur Ramachandran.) (See Color Plate 18.1 in color plate section.)

> ''Open your eyes.''
>
> (See Ramachandran 1998, 47f. For the clinical and experimental details, see Ramachandran and Rogers-Ramachandran, 1996) (Fig. 1).

By now, it should be clear how these experimental findings illustrate the concept of a ''self-model'' that I introduced above; what is moving in this experiment *is* the PSM. What made the sudden occurrence of kinesthetic movement sensations in the lost subregion of the self-model possible was the installation of an additional source of feedback, of ''virtual information.'' This immediately created a new functional property, let us call it ''availability for selective motor control.'' By providing access to the visual mode of self-simulation, this made the corresponding information available to volition as well. Now, volitional control once again was possible. This experiment also shows how phenomenal properties are determined by computational and representational properties. Bodily self-consciousness is directly related to brain processes.

Let us directly move on to the next example, while staying with the phenomenology of phantom limbs. How ''ghostly'' are phantom limbs? Can we measure the ''realness'' of the conscious self? A recent case study by Brugger and colleagues introduced a vividness rating on a 7-point scale that showed highly consistent judgments across sessions for their subject AZ, a 44-year-old university-educated woman born without forearms and legs. For as long as she remembers, she has experienced mental images of forearms (including fingers) and legs (with feet and first and fifth toes) — but, as the figure below shows, these were not *as* realistic as the content of her non-hallucinatory PSM. Functional magnetic resonance imaging of phantom hand movements showed no activation of the primary sensorimotor areas, but of the premotor and parietal cortex bilaterally. Transcranial magnetic stimulation (TMS) of the sensorimotor cortex consistently elicited phantom sensations in the contralateral fingers and hand. In addition, premotor and parietal stimulation evoked similar phantom sensations, albeit in the absence of motor-evoked potentials in the stump. These data clearly demonstrate how body parts that were never physically developed can be phenomenally simulated in sensory and motor cortical areas. Are they components of an innate body model? Or could they have been ''mirrored

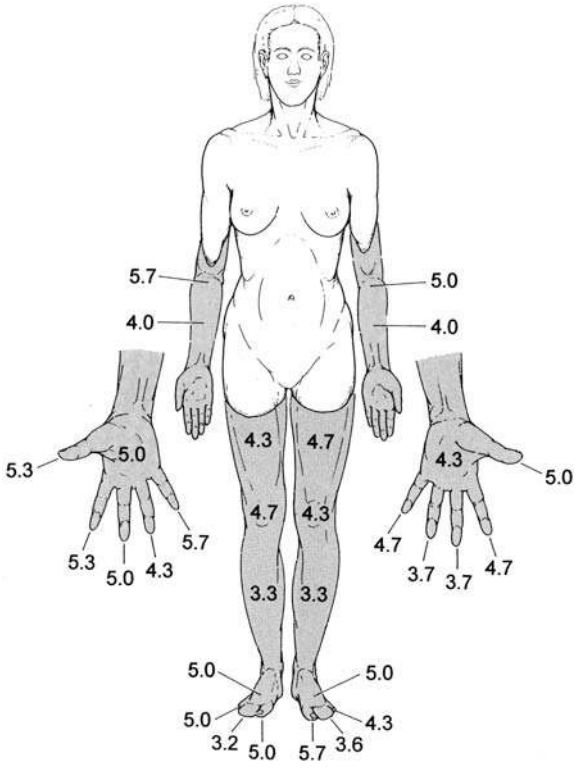


Fig. 2. Evidence for an innate component of the PSM? Phantoms (shaded areas) in a subject with limb amelia. The numbers are vividness ratings for the felt presence of different phantom body parts on a 7-point scale from 0 (no awareness) to 6 (most vivid impression). (Picture courtesy of Peter Brugger, Zürich.)

into'' the patient's self-model through the visual observation of other human beings moving around? As I am a philosopher and not a neuropsychologist, I will refrain from further amateurish speculation at this point (Fig. 2).

However, recent results from research on pain experiences in phantom limbs point to the potential existence of a genetically determined neuromatrix whose activation pattern may form the basis of these rigid parts of the self-model and the more invariant background of bodily self-experience (the ''phylomatrix of the body schema''; see Melzack, 1989; on the concept of a ''neurosignature,'' see Melzack, 1992, p. 93; an important study on phantom limbs following aplasia and early amputation is Melzack et al., 1997). Another interesting empirical result is that more than 20% of children born without an arm or a leg later develop the realistic conscious experience of having a phantom limb. In the context of phenomenal ''realness'' and in terms of the integration of the bodily self-model into the brain's conscious reality model as a whole it may also be interesting to note that, in this case, ''Awareness of her phantom limbs is transiently disrupted only when some object or person invades their felt position or when she sees herself in a mirror.'' (Brugger et al., 2000, p. 6168. For further details concerning the phenomenological profile see ibid; for an interesting experimental follow-up study demonstrating the intactness of the phenomenal model of kinesthetic and postural limb properties, see Brugger et al., 2001).

What do the phenomenologies of Ramachandran's and Brugger's subjects have in common? The transition from stump to phantom limb is *seamless*; subjectively, they are both part of one and the same bodily self, because the quality of ownership is distributed evenly among them. There is no gap or sudden jump in the sense of ownership. The emergence of the bodily self-model is based on a subpersonal, automatic process of *binding features* together, of achieving coherence. But what exactly is it that is being experienced? What is the *content* of experience? Aristotle said that the soul is the *form* of the physical body, which perishes together with it at death (*On The Soul*, II: 412a, 412b–413a). According to Spinoza, the soul is the *idea* that the body develops of itself (*The Ethics*, II: 12 and 13). In more modern terms, we might say that an ''idea'' is simply a mental representation — more precisely a *self*-representation — and that the content of self-consciousness is the introspectively accessible part of this self-representation, namely the PSM postulated by the self-model theory. *Gestalt* properties — like body shape — are *global* properties of an object, and could the self-model then not be a neural mechanism to represent exactly such global properties, a new tool to acquire knowledge about the organism *as a whole*? Plato, however, claimed that some ideas are innate. And this still is an interesting question for today's neuroscience of self-consciousness as well: Does the PSM possess an innate component? Is the conscious body image a

kind of "fixed idea," anchored in an inborn and genetically predetermined *nucleus*?

Let us now turn to example no. 3. It comes from a different scientific discipline altogether, namely from the fascinating new field of evolutionary robotics. It demonstrates a number of further aspects that the conceptual framework of SMT, the self-model theory, predicts and seeks to explain. First, a self-model can be entirely *unconscious*; i.e., it can frequently be seen as the product of an automatic "bottom-up" process of *dynamical self-organization*; second, it is not a "thing" (or a model of a thing) at all, but based on a continuous, ongoing modeling *process*; third, it can exhibit considerable *plasticity* (i.e., it can be modified through learning); and fourth, in its origins it is not based on language or conceptual thought, but very likely on an attempt to organize motor behavior. It is a computational tool to achieve global control. More precisely, a body-model has the function of integrating sensory impressions with motor output in a more intelligent and flexible manner. The unconscious precursor of the PSM clearly was a new form of intelligence.

Bongard et al. (2006) have created an artificial "starfish" that gradually develops an explicit internal self-model. Their four-legged machine uses actuation–sensation relationships to indirectly infer its own structure and then uses this self-model to generate forward locomotion. When part of its leg is removed, it adapts its self-model and generates alternative gaits — it learns to limp. In other words unlike the phantom-limb patients presented in example no. 1 and no. 2 (and like most ordinary patients), it is able to *restructure* its body-representation following the loss of a limb. It can learn. This concept may not only help develop more robust machines and shed light on self-modeling in animals, but is also theoretically interesting, because it demonstrates for the first time that a physical system has the ability, as the authors put it, to "autonomously recover its own topology with little prior knowledge" by constantly optimizing the parameters of its own resulting self-model (Fig. 3a–c).

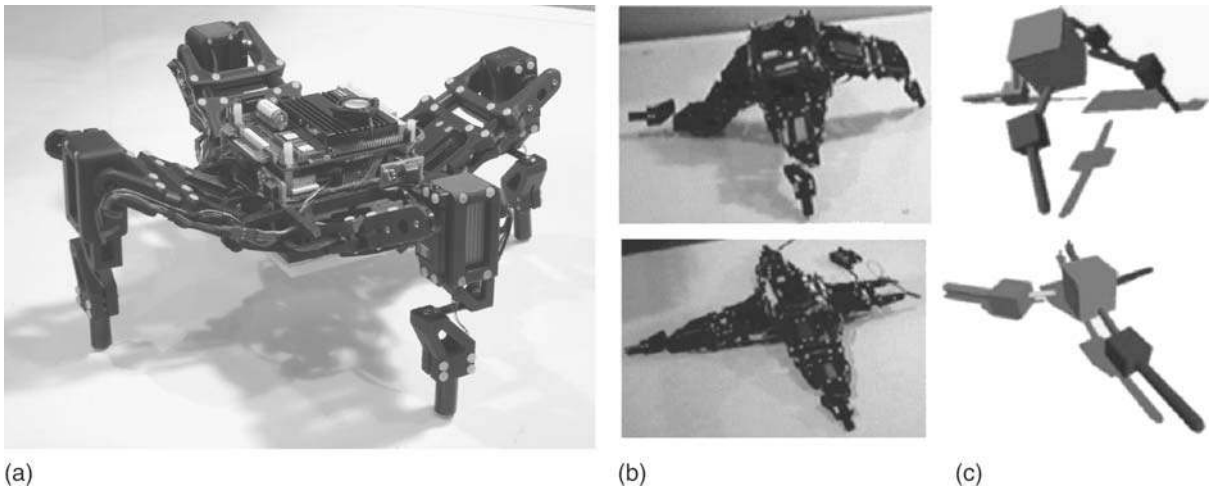


Fig. 3. (a) Starfish, a four-legged physical robot that has eight motorized joints, eight joint angle sensors, and two tilt sensors. (See www.ccsl.mae.cornell.edu/research/selfmodels/morepictures.htm for additional online material.) (b and c) The starfish-robot walks by using an explicit internal self-model that it has autonomously developed and that it continuously optimizes. If he loses a limb, he can adapt his internal self-model. (d) The robot continuously cycles through action execution. (a and b) Self-model synthesis. The robot physically performs an action (a). Initially, this action is random; later, it is the best action found in (c). The robot then generates several self-models to match sensor data collected while performing previous actions (b). It does not know which model is correct. (c) Exploratory action synthesis. The robot generates several possible actions that disambiguate competing self-models. (d) Target behavior synthesis. After several cycles of (a)–(c), the best current model is used to generate locomotion sequences through optimization. (d) The best locomotion sequence is executed by the physical device. (e) (See Color Plate 18.3 in color plate section.)
Figure 3 continued on p. 224.

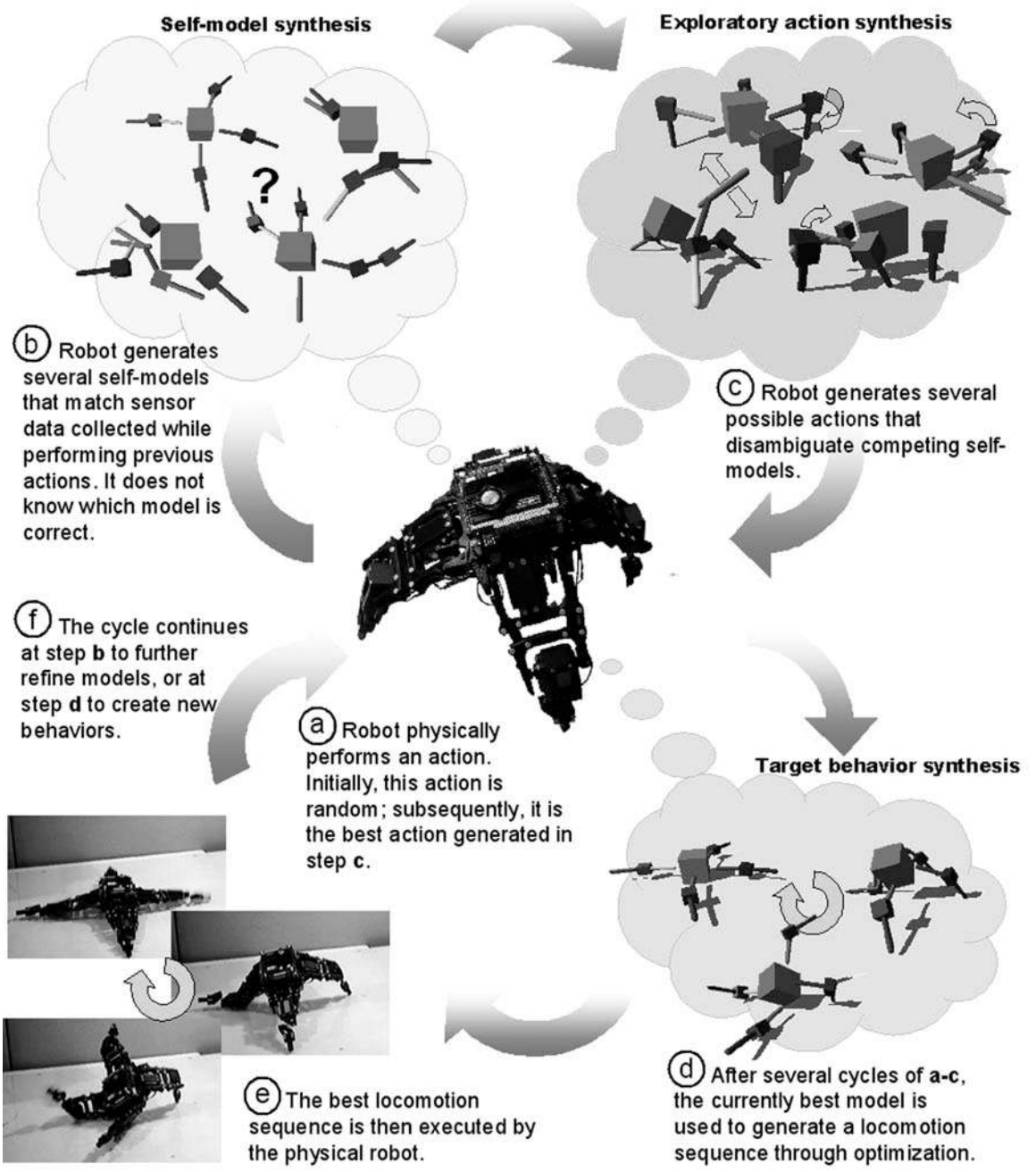


Fig. 3. (*Continued*)

Starfish not only synthesizes an internal self-model, but also uses this self-model to generate intelligent behavior. The next figure gives an overview over this process (Fig. 3d).

As we see, the robot initially performs an arbitrary motor action and records the resulting sensory data. The model synthesis component then synthesizes a set of 15 candidate self-models using stochastic optimization to explain the observed sensory–actuation relationship. The robot then synthesizes an exploratory motor action that causes maximum *disagreement* among the different predictions of these competing self-models. This action is physically carried out, and the 15 candidate self-models are subsequently improved using the new data. When the models converge, the most accurate model is used by the behavior synthesis component to create a desired behavior that can then be executed by the robot. If the robot detects unexpected sensor–motor patterns or an external signal resulting from unanticipated morphological change, it reinitiates the alternating cycle of modeling and exploratory actions to produce new models reflecting this change. The most accurate of these new models is then used to generate compensatory behavior and recover functionality.

Technical details aside — what are the philosophical consequences of example no. 3? First, you do not have to be a living being in order to have a self-model. Non-biological SMT-systems are possible. Second, a self-model can be entirely unconscious, i.e., it does not have to be a PSM. Awareness obviously is a second step (see Metzinger, 1995b, 2000a, for a first overview; Metzinger, 2003a, Section 3.2, for an additional set of ten constraints to be satisfied for conscious experience). Third, a self-model supports planning and fast learning processes in a number of different ways. It clearly makes a system more intelligent. Fourth, it is what I called a virtual model or "virtual organ" above, and one of its major functions consists in appropriating a body by using a global morphological model to control it as a whole. Elsewhere, I have introduced the term "second-order embodiment" for this type of self-control (Metzinger, 2006b). If I may use a metaphor: one of the core ideas is that a self-model allows a physical system to "enslave" its low-level dynamics with the help of a single, integrated, and internal whole-system model, thereby controlling and functionally "owning" it. This is the decisive first step towards becoming an autonomous agent.

Step three: a representationalist analysis of the three target properties

Here, the basic idea is that self-consciousness, first of all, is an *integrative* process: by becoming embedded in the currently active self-model, representational states acquire the higher order property of phenomenal mineness. If this integrative process is disturbed, this results in various neuropsychological syndromes or altered states of consciousness (for case studies, see Chapter 7 in Metzinger, 2003a). Let us take a look at some examples of what happens when phenomenal mineness, the subjective sense of ownership, is selectively lost.

- Florid schizophrenia: Consciously experienced thoughts are no longer *my* thoughts.
- Somatoparaphrenia, unilateral hemi-neglect: My leg is no longer *my* leg.
- Depersonalization, delusions of control: I am a robot, I am turning into a puppet, and volitional acts are no longer *my* volitional acts. (In this case, what philosopher and psychiatrist Karl Jaspers called *Vollzugsbewusstsein*, or "executive consciousness," is selectively lost.)
- Manic disorders: I am the whole world; all events in the world are controlled by *my own* volitional acts.

Subjectively experienced "mineness" is a property of discrete forms of phenomenal content, such as the mental representation of a leg, a thought, or a volitional act. This property, the sense of ownership, is not necessarily connected to these mental representations; i.e., it is not an intrinsic, but a *relational* property. That a thought or a body part is consciously experienced as your own is not an essential, strictly necessary property of the conscious experience of this thought or body part.

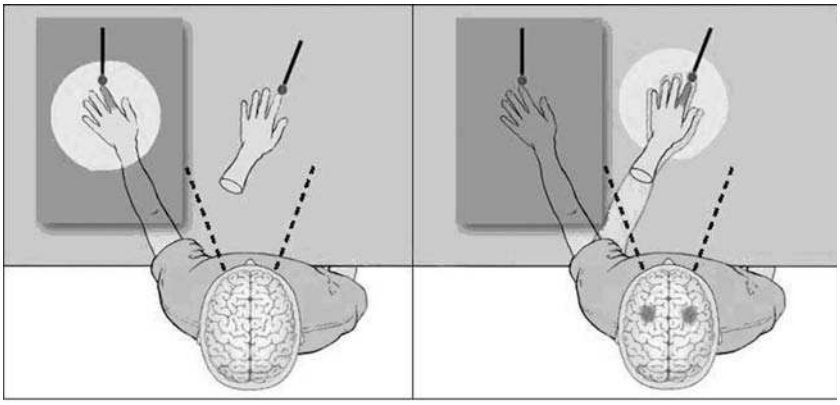


Fig. 4. The rubber-hand illusion. A healthy subject experiences an artificial limb as part of her own body. The subject observes a facsimile of a human hand while one of her own hands is concealed (gray square). Both the artificial rubber hand and the invisible hand are then stroked repeatedly and synchronously with a probe. The yellow and green areas indicate the respective tactile and visual receptive fields for neurons in the premotor cortex. The illustration on the right shows the subject's illusion as the felt strokes (green) are brought into alignment with the seen strokes of the probe (areas of heightened activity in the brain are colored red; the phenomenally experienced, illusory position of the arm is indicated by the blue area). The respective activation of neurons in the premotor cortex is demonstrated by experimental data. (Figure by Litwak illustrations studio 2004.) (See Color Plate 18.4 in color plate section.)

It could have been otherwise, in other phenomenological contexts, mineness disappears. Its distribution over the different elements of a conscious world-model can vary. If the system is no longer able to integrate certain discrete representational contents into its self-model, it is lost. If this analysis is correct, it should be possible, at least in principle, to operationalize this property by searching for an empirically testable metrics for the coherence of the self-model in the respective areas of interest. One could also empirically investigate *how* and in which brain areas a certain type of representational content is integrated into the self-model. Here is a concrete example for what I mean by "mineness," example no. 4 (Fig. 4).

In the rubber-hand illusion (RHI), the sensation of being stroked with a probe is integrated with the corresponding visual perception in such a way that the brain transiently matches a proprioceptive map (of the subject's own-body perception) with a visual map (of what the subject is currently seeing). At the same time, the feeling of "ownership" or phenomenal "mineness" is transferred to the rubber hand. The subject experiences the rubber hand as her *own* hand and feels the strokes *in* this hand. When asked to point to her concealed left hand, her arm movement will automatically swerve in the direction of the rubber hand (Botvinick and Cohen, 1998, p. 756). If one of the fingers of the rubber-hand is "hurt" by being bent backwards into a physiologically impossible position, the subject will also experience her real phenomenal finger as being bent much farther backwards than it is in reality. At the same time, this will also result in a clearly measurable skin conductance response. While only 2 out of 120 subjects reported an actual pain sensation, many subjects drew back their real hands, opened their eyes up widely in surprise, or laughed nervously (Armel and Ramachandran, 2003, p. 1503). Subjects also showed a noticeable reaction when the rubber hand was hit with a hammer. Again, it becomes clear how the phenomenal target property is directly determined by representational and functional brain processes. What we experience as part of our self depends on the respective context and on which information our brain integrates into our currently active self-model (see especially Botvinick and Cohen, 1998, and the neuroimaging study by Botvinick, 2004; Ehrsson et al., 2004). The intriguing question, of course, is this: Could *whole-body* illusions exist as well? The answer is

yes, and we will soon return to this point in example no. 5.

But first, let us take a look at the second target property, at consciously experienced selfhood. Methodologically, it is important to first isolate the simplest form of the target. Phenomenal selfhood corresponds to the existence of a single, coherent, and temporally stable self-model that constitutes the center of the representational state as whole. If this representational module is damaged or disintegrates, or if multiple structures of this type alternate or are simultaneously activated by the system, this will again result in various neuropsychological disturbances or altered states of consciousness

- Ansognosia and anosodiaphoria: Loss of higher order insight into existing deficits, e.g., in cortically blind patients who deny that they are blind (Anton's Syndrome).
- Dissociative Identity Disorder (DID): The system uses different and alternating self-models as a means of coping with extremely traumatic and socially inconsistent situations (for the current diagnostic criteria for DID, see DSM-IV: 300.14).
- *Ich-Störungen*, or identity disorders: A large class of psychiatric disturbances connected to altered forms of experiencing one's own *identity*. Schizophrenia is a classical example, as are Cotard syndrome, reduplicative paramnesia, or delusional misidentifaction (for a discussion on why identity disorders are interesting from a philosophical perspective, see Metzinger, 2004a).

The existence of a stable self-model also almost always gives rise to the ''perspectivalness of consciousness'' in terms of transient subject–object relationships (see step 6 below; see also Nagel, 1986; Metzinger, 1993, 1995a, 2005a, and especially Metzinger, 2006a). This structural feature of the global representational space leads to the episodic instantiation of a temporally extended and non-conceptual first-person perspective. It, too, can be lost.

- Complete depersonalization: Loss of the phenomenal first-person perspective, accompanied by dysphoric states and functional deficits (''dreadful ego-dissolution''; see Dittrich, 1985).
- Mystical experiences: Selfless and non-centered global states, which are experienced viz. described as non-pathological and unthreatening (''oceanic boundary loss,'' ''*The Great View from Nowhere*'').

In order to do justice to the wealth and the diversity of different forms of human experience, one has to acknowledge the existence of certain non-perspectival and selfless forms of conscious experience. Phenomenologically, *non-subjective* consciousness — phenomenal experience that is not tied to a self or an individual first-person perspective — is not only a possibility, but a reality, even if we may find this idea inconceivable. The self-model theory provides the conceptual means to account for these special cases (for additional neurophenomenological case studies, see Metzinger, 2003a, Chapters 4 and 7).

Example no. 5 will demonstrate this principle in another domain. If we have the necessary conceptual instruments, we can not only take the subtleties and the variability of human experience seriously. We can also develop new interdisciplinary research programs that penetrate into ''taboo zones'' and shed light on phenomena that in the past were only the targets of esoteric folklore and metaphysical ideologies. Could there be an integrated kind of bodily self-consciousness, be it of a mobile body fully available for volitional control or of a paralyzed body that in its entirety is a phenomenal confabulation — in short, a *hallucinated* and a *bodily* self at the same time? Is it conceivable that something like a full-body analog of the rubber-hand-illusion or a ''globalized phantom-limb experience'' — the experience of a *phantom body* — could emerge in a human subject? The answer is, yes. There is a well-known class of phenomenal states in which the experiencing person undergoes the untranscendable and highly realistic conscious experience of leaving his or her physical body, usually in the form of an etheric double, and moving around outside of it. In other

words, there is a class (or at least a strong cluster) of intimately related phenomenal models of reality that are classically characterized and defined by a *visual representation* of one's own body from a perceptually impossible, externalized third-person perspective (e.g., seeing oneself from above, lying on the bed, or on the road) plus a *second representation* of one's own body, typically (but not in all cases) freely hovering or floating in space. This second body-model is the locus of the phenomenal self. It not only forms the "true" focus of one's phenomenal experience, but also functions as an integrated representation of all kinesthetic qualia and all non-visual forms of proprioception. This class of phenomenal states is called the "Out-of-body experience" (OBE). Elsewhere (Metzinger, 2005b, for further references see also Lenggenhager et al., 2007), I have argued that our traditional, folk-phenomenological concept of a "soul" may have its origins in accurate and sincere first-person reports about the experiential content of this specific neurophenomenological state-class.

OBEs frequently occur spontaneously while falling asleep, but also following severe accidents or during surgical operations. At present, it is not clear whether the concept of an OBE possesses a clearly delineated set of necessary and sufficient conditions. Instead, the concept of an OBE may turn out to be a cluster concept constituted by a whole range of diverging (and possibly overlapping) subsets of phenomenological constraints, each forming a set of sufficient, but not necessary, conditions. On the other hand, the OBE clearly is something like a phenomenological *prototype*. There is a common core to the phenomenon, as can be seen from the simple fact that many readers will already have heard about this type of experience in one way or another.

One can offer a representationalist analysis of OBEs by describing them as a class of deviant self-modeling processes. On the level of conscious self-representation, a prototypical feature of this class of deviant PSM seems to be the coexistence of (a) a more or less veridical representation of the bodily self as seen from an external visual perspective, which does *not*, however, function as the center of the global model of reality, and (b) a second self-model, which according to subjective experience largely integrates proprioceptive perceptions — although, interestingly, weight sensations are only integrated to a lesser degree — and possesses special properties of shape and form that may or may not be veridical. Both models of the experiencing system are located within the same spatial frame of reference (that is why they are *out-of*-body-experiences). This frame of reference is an *egocentric* frame of reference. Let us now look at two classical phenomenological descriptions of OBEs, as spontaneously occurring in an ordinary non-pathological context

> I awoke at night — it must have been at about 3 a.m. — and realized that I was completely unable to move. I was absolutely certain I was not dreaming, as I was enjoying full consciousness. Filled with fear about my current condition, I had only one goal, namely to be able to move my body again. I concentrated all my will-power and tried to roll over to one side: Something rolled, but not my body — something that was me, my whole consciousness including all of its sensations. I rolled onto the floor beside the bed. While this happened, I did not feel bodiless, but as if my body consisted of a substance in between the gaseous and the liquid state. To the present day, I have never forgotten the combination of amazement and great surprise that gripped me when I felt myself falling onto the floor, but without the expected thud. Had the movement actually unfolded in my normal physical body, my head would have had to collide with the edge of my bedside table. Lying on the floor, I was overcome by terrible fear and panic. I knew that I possessed a body, and I only had one great desire — to be able to control it again. With a sudden jolt, I regained control, without knowing how I managed to get back into it. (Waelti, 1983, p. 25; English translation TM)

The prevalence of OBEs ranges from 10% in the general population to 25% in students, with extremely high incidences in certain subpopulations like, to name just one example, 42% in schizophrenics (Blackmore, 1986; see also Blackmore, 1982; for an overview and further references see Alvarado, 1986, 2000, p. 18; Irwin, 1985, p. 174). However, it would be false to assume that OBEs typically occur in people suffering from severe psychiatric disorders or neurological deficits. Quite the contrary, most OBE-reports come from ordinary people in everyday life situations. Let us therefore stay with non-pathological situations and look at another paradigmatic example, again reported by Swiss biochemist Ernst Waelti

> I went to bed in a dazed state at 11 p.m. and tried to go to sleep. I was restless and turned over frequently, causing my wife to grumble briefly. Now, I forced myself to lie in bed motionless. For a while I dozed before feeling the need to pull up my hands, which were lying on the blanket, in order to bring them into a more comfortable position. At the same instant, I realized that I was absolutely unable to move and that my body was lying there in some kind of paralysis. Nevertheless, I was able to pull my hands out of my physical hands, as if the latter were just a stiff pair of gloves. The process of detachment started at the fingertips, in a way that could be clearly felt, almost with a perceptible sound, a kind of crackling. It was exactly the movement that I had actually intended to carry out with my physical hands. With this movement, I detached from my body and floated out of it head first. I moved into an upright position, as if I was almost weightless. Nevertheless, I had a body consisting of real limbs. You have certainly seen how elegantly a jellyfish moves through water. I could now move around with the same ease. I lay down horizontally in the air and floated across the bed, like a swimmer, who has pushed himself from the edge of a swimming pool. A delightful feeling of liberation arose within me. But soon, I was seized by the ancient fear common to all living creatures, the fear of losing my physical body. It sufficed to drive me back into my body. (Waelti, 1983, p. 25; English translation TM) (Figs. 5 and 6)

Sleep paralysis is not a necessary precondition for OBEs. They frequently occur during extreme sports, for instance, in high-altitude climbers or marathon runners.

> A Scottish woman wrote that, when she was 32 years old, she had an OBE while training for a marathon. "After running approximately 12–13 miles … I started to feel as if I wasn't looking through my eyes but from somewhere else. … I felt as if something was leaving my body, and although I was still running along looking at the scenery, I was looking at myself running as well. My 'soul' or whatever, was floating somewhere above my body high enough up to see the tops of the trees and the small hills." (Alvarado, 2000, p. 184)

The classic OBE contains two self-models, one visually represented from an external perspective and one forming the center of the phenomenal world from which the first-person perspective originates. What makes the representationalist and functionalist analysis of OBEs difficult and at the same time challenging is the fact that many *related* phenomena exist, e.g., autoscopic phenomena during epileptic seizures in which only the first criterion is fulfilled (for a neurological categorization see Brugger et al., 1997). Devinsky et al. (1989, p. 1080) have differentiated between autoscopy in the form of a complex hallucinatory perception of one's own body as being external with "the subject's consciousness … usually perceived within his body" and a second type, the classic OBE, which includes the feeling of leaving one's body and viewing it from another vantage-point. The incidence of autoscopic seizures is

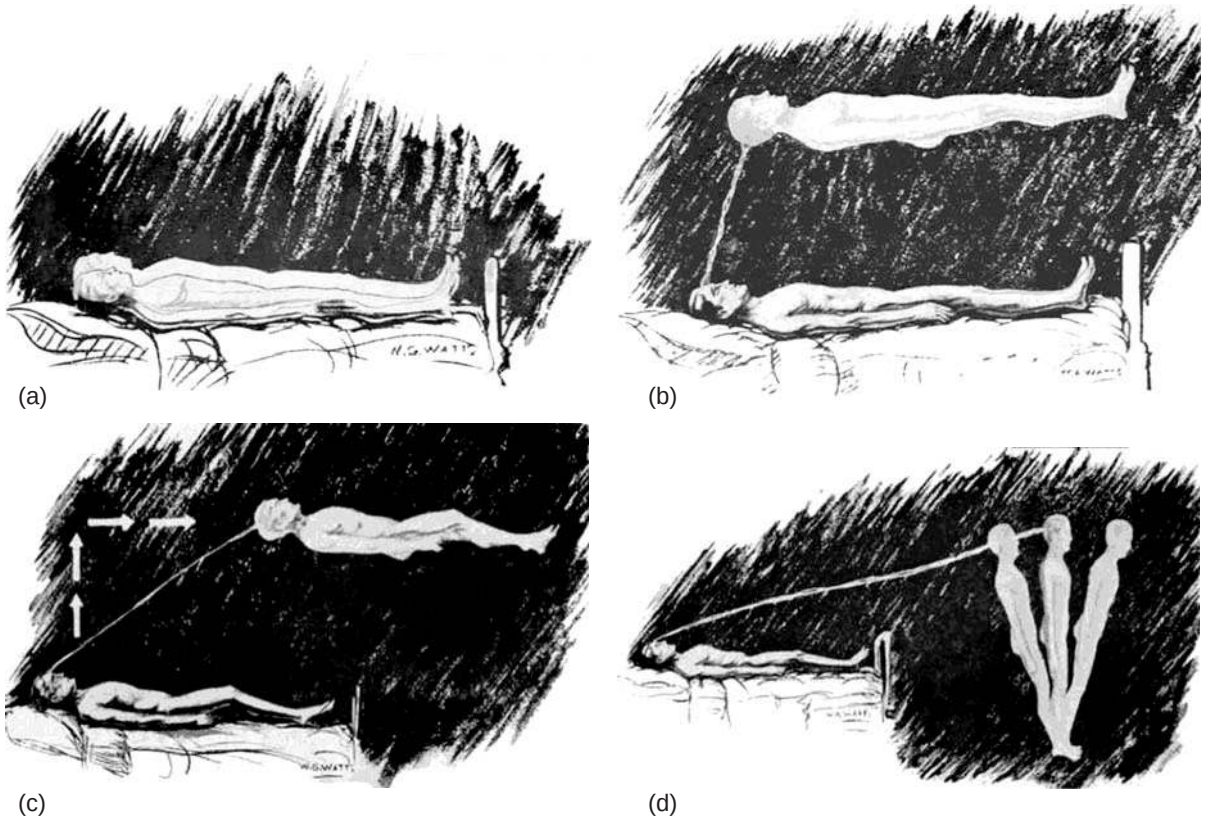


Fig. 5. Kinematics of the PSM during OBE-onset: The classical Muldoon-scheme. From Muldoon S. and Carrington, H. (1929). The Projection of the Astral Body. Rider & Co., London.

Fig. 6. Kinematics of the phenomenal body-image during OBE onset. An alternative, but equally characteristic motion pattern, as described by Swiss biochemist Ernst Waelti (1983).

possibly higher than previously recognized, and the authors found a 6.3% incidence in their patient population (Devinsky et al., 1989, p. 1085). Seizures involving no motor symptoms or loss of consciousness, which may not be recognized by the patient, may actually be more frequent than commonly thought for a case study of a patient who first experienced OBEs for a number of years and only later suffered from generalized seizures (see Vuilleumier et al., 1997, p. 116).

What function could this type of experience have *for* the organism as a whole? Here is a speculative proposal by Devinsky and colleagues

> There are several possible benefits that dissociative phenomena, such as autoscopy, may confer. For example, when a prey is likely to be caught by its predator, feigning death may be of survival value. Also, accounts from

> survivors of near-death experiences in combat or mountaineering suggest that the mental clarity associated with dissociation may allow subjects to perform remarkable rescue manoeuvres that might not otherwise be possible. Therefore, dissociation may be a neural mechanism that allows one to remain calm in the midst of near-death trauma. (Devinsky et al., 1989, p. 1088)

It is not at all inconceivable that there are physically or emotionally stressful situations in which an information-processing system is forced to introduce a ''representational division of labour'' by distributing different representational functions into two or more distinct self-models (for instance in what in the past was called ''multiple personality disorder,'' see Metzinger, 2003a, Section 7.2.4). The OBE may be an instance of transient functional modularization, of a ''purposeful,'' i.e., functionally adequate, separation of levels of representational content in the PSM. For instance, if the system is cut off from somatosensory input or flooded with stressful signals and information threatening the overall integrity of the self-model as such, it may be advantageous to integrate the ongoing conscious representation of higher cognitive functions like attention, conceptual thought, and volitional selection processes into a *separate* model of the self. This may allow for a high degree of integrated processing, i.e., of ''mental clarity,'' by functionally encapsulating and thereby *modularizing* different functions like proprioception, attention, and cognition in order to preserve at least some of these functions in a life-threatening situation. Almost all necessary system-related information is still globally available, and higher order processes like attention and cognition can still operate on this information as it continues to be presented in an integrated manner, but its distribution across specific subregions of phenomenal space as a whole changes dramatically. Only one of the two self-models is truly ''situated'' in the overall scene; only one of them is immediately embodied and virtually self-present in the sense of being integrated into an internally simulated behavioral space.

It has long been known that OBEs not only occur in healthy subjects, but in certain clinical populations (e.g., epileptics) as well. In a recent study, Olaf Blanke and colleagues were able to localize the relevant brain lesion or dysfunction in the temporo–parietal junction (TPJ) in five out of six patients. It was also possible, for the first time, to induce an OBE-type state by direct electrical stimulation. These researchers argue that two separate pathological conditions may be necessary to cause an OBE. First, a disintegration in the self-model or ''personal space'' (brought about by a failure to integrate proprioceptive, tactile, and visual information regarding one's own body) plus an additional, second disintegration between external, ''extrapersonal'' visual space, and the internal frame of reference created by vestibular information. The experience of seeing one's own body in a position that does not coincide with its felt position could therefore be caused by cerebral dysfunction at the TPJ, causing both types of functional disintegration and thereby leading to the representational configuration described above (Figs. 7 and 8).

Using evoked potential mapping, these authors also showed that a selective activation of the TPJ takes place 330–400 ms after healthy volunteers mentally imagined themselves being in a position and taking a visual perspective characteristic of an OBE. At the same time, it is possible to impair this mental transformation of the bodily self-model by interfering at this specific location with TMS. In an epileptic patient with OBEs caused by damage at the TPJ, it could be shown that by mimicking the OBE-PSM (i.e., by mentally simulating an OBE like the ones she had experienced before), there was a partial activation of the seizure focus (Blanke et al., 2005). Therefore, there exists an anatomical bridge overlap between these three very similar types of phenomenal mental content.

What is most needed at the current stage is an experimental design that makes OBEs a controllable and repeatable phenomenon in healthy subjects, under laboratory conditions. Achieving this interim goal would be of high relevance, not only from an empirical, but also from a philosophical perspective. Studying the functional fine structure of embodiment by developing a convincing representationalist analysis of phenomenal

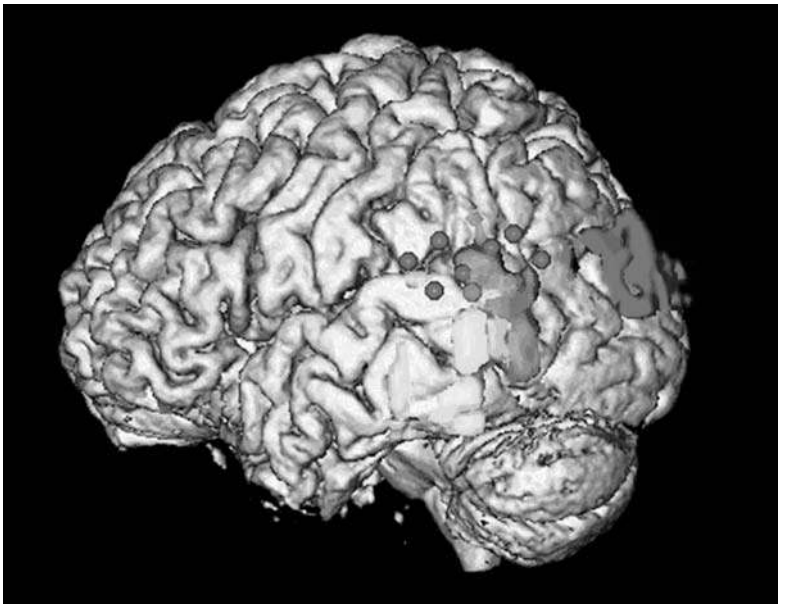


Fig. 7. The figure shows results of a mean lesion overlap analysis in five patients with OBEs. The analysis centers on the TPJ. The blue dots show the locus of electrical cortical stimulation in the patient in whom an OBE-like phenomenal state was artificially induced. (Figure courtesy of Olaf Blanke, cf. Blanke et al., 2004.) (See Color Plate 18.7 in color plate section.)

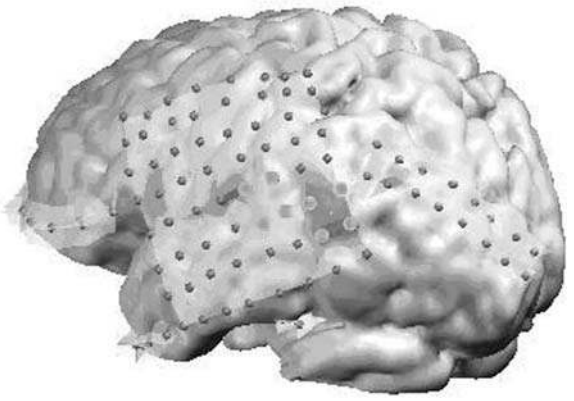


Fig. 8. The next figure shows another set of patient data, in which MRI was performed with implanted electrodes in the left hemisphere. The epileptic focus, where the discharge induced an OBE, is indicated by eight turquoise electrodes at the TPJ. (Figure courtesy of Olaf Blanke, cf. Blanke et al., 2005.) (See Color Plate 18.8 in color plate section.)

disembodiment would certainly shed new light on the issue of non-conceptual self-awareness and the origin of a conscious first-person perspective. In particular, it would be of high theoretical relevance if one could empirically demonstrate the possibility of minimal selfhood *without an agency component*. Let me therefore give you a brief example of my own recent research. Example no. 5 is a study based on interdisciplinary cooperation between neuroscience and philosophy of mind and, specifically, on an experimental design originally developed from philosophical considerations (for details see Lenggenhager et al., 2007).

The classical RHI (example no. 4) only tells us something about the target property of ''ownership'' (for body parts), but not about ''selfhood'' (ownership for the *whole* body). To manipulate attribution and localization of the *entire* body and to study selfhood per se we designed an experiment based on clinical data in neurological patients with out-of-body experiences. These data suggest that the spatial unity between self and body may be disrupted leading in some cases to the striking experience that the conscious self is localized at an extracorporeal position. Therefore, the aim of the present experiments was to induce out-of-body experiences in healthy participants in order to investigate the phenomenal target

property of selfhood. We hypothesized that under adequate experimental conditions participants would experience a visually presented body as if it was their own, inducing a drift of the subjectively experienced bodily self to a position outside one's bodily borders. Can one create a whole-body analog of the RHI, an illusion during which healthy participants experience a virtual body as if it were their own and localize their self outside their body boundaries at a different position in space?

We applied virtual reality to examine the possible induction of out-of-body experiences by using multisensory conflict. In the first experiment participants viewed the back of their body filmed from a distance of 2 m and projected onto a 3D-video head-mounted display (HMD; see Fig. 9). The participants' back was stroked during 1 min either synchronously or asynchronously with respect to the virtually seen body. Global self-attribution of the virtual character was measured by a questionnaire that was adapted from the RHI. Global self-localization was measured by passively displacing the blindfolded participants immediately after the stroking and asking them to return to their initial position (Fig. 9).

While being stroked, the subjects were either shown their own back (''own body condition''), the back of a mannequin (''fake body condition''), or an object (''object condition'') being stroked and projected directly (synchronously) or with a time lag (asynchronously) onto a HMD. After being stroked, the subjects were passively displaced and then asked to return to their initial position and fill out a modified ''rubber-hand-questionnaire.'' Results of the questionnaire showed that for the synchronous ''own body'' and ''fake body'' conditions, subjects often felt as if the observed virtual figure were their own body. This impression was less likely to occur in the ''object condition'' and in all of the asynchronous conditions. The synchronous experimental conditions also showed a significantly larger shift towards the projected real or fake body than the asynchronous and control conditions. These data suggest that self-location — due to conflicting visual-somatosensory input — is as prone to misidentification and mislocalization as was previously reported for body parts, as in the RHI.

Illusory self-localization to a position outside one's body shows that bodily self-consciousness and selfhood can be dissociated from an accurate representation of one's physical body position. This differs from the RHI where the aspect of selfhood remained constant and only the attribution and localization of the stimulated hand was manipulated. Does illusory self-localization to a position outside one's body mean that we have experimentally induced full-blown out-of-body experiences? No, this was only a first step. But it is quite clear what the next steps will have to be. Out-of-body experiences are characterized by disembodiment of the self to an extracorporeal location, an extracorporeal visuo-spatial perspective, and seeing of one's own body from this extracorporeal self-location. As the present illusion was neither associated with overt disembodiment nor with a change in visuo-spatial perspective, we argue that we have induced only some aspects of out-of-body experiences or rather the closely related experience of heautoscopy that has also been observed in neurological patients (see original publication for further references).

To give just one example, I believe that an additional necessary condition involved in generating full-blown out-of-body experiences and the complete transfer of selfhood to the illusory body is a transient episode of visual-vestibular disintegration. At least two spatial frames of reference must be functionally dissociated, in order to not only have a ''teleportation-OBE,'' but a realistic exit phenomenology, a gradual motion path through phenomenal space. This general principle should hold for our experimental setup as well as for OBEs in epileptic patients or ''gifted subjects'' in the healthy population. Why is this principle relevant from a theoretical perspective, and why is it difficult to test experimentally? In standard situations, and as opposed to all other conscious model of aspects of reality, the human PSM is anchored in the brain through a continuous flow of self-generated input. There exists a persistent causal link into the physical body itself. In order to understand the SMT better, we must turn to this point now — it explains why our conscious model of reality is a *centered* model of reality.

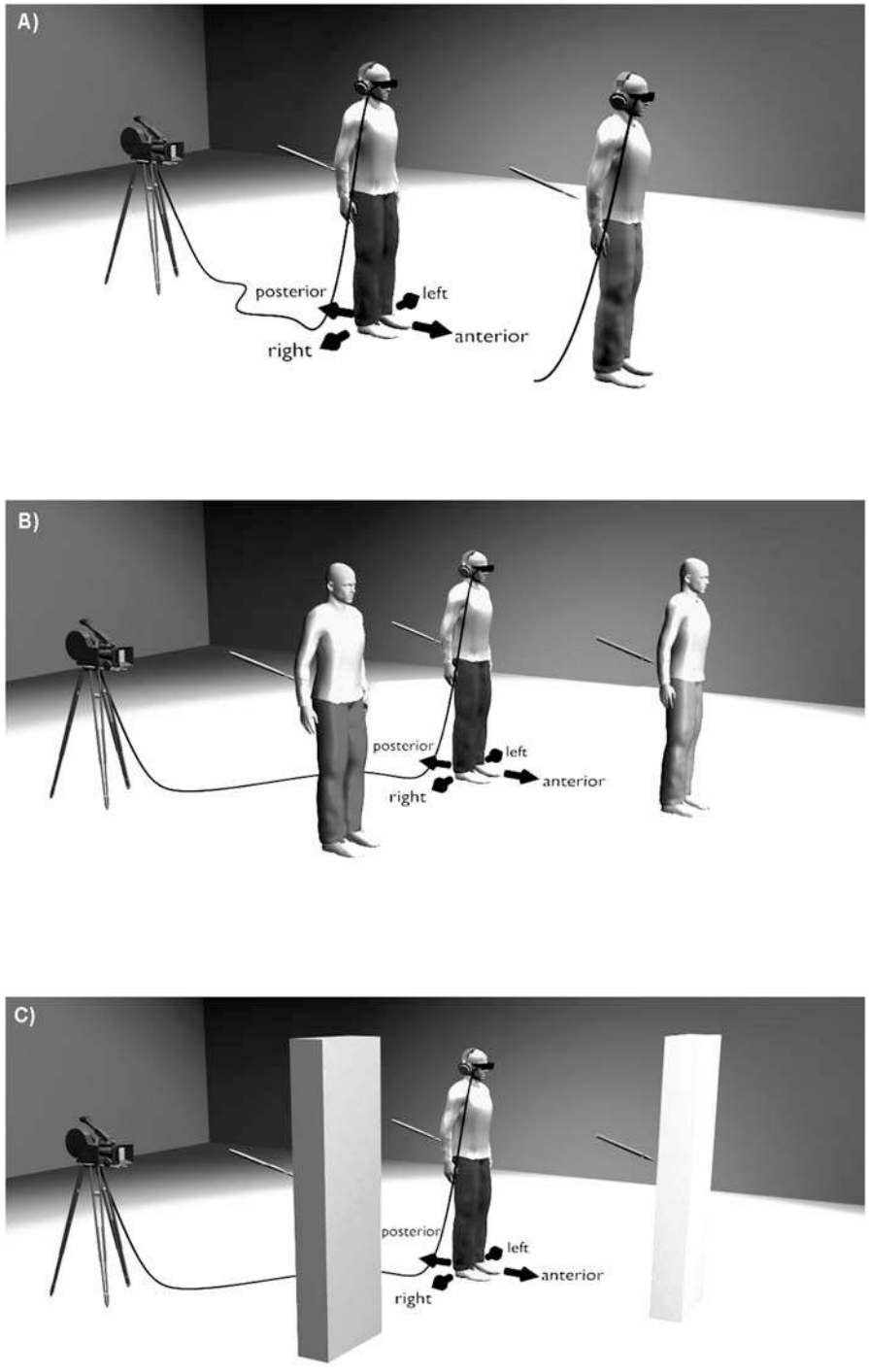


Fig. 9. (A) Participant (in dark blue trousers) sees through a HMD his own virtual body (light blue trousers) in 3D, standing 2 m in front of him and being stroked synchronously or asynchronously at the participant's back. In other conditions (Study II) the participant sees either (B) a virtual fake body (light red trousers) or (C) a virtual non-corporeal object (light gray) being stroked synchronously or asynchronously at the back. Dark colors indicate the actual location of the physical body/object, whereas light colors represent the virtual body/object seen on the HMD. Illustration by Martin Boyer, published in Science. (See Color Plate 18.9 in color plate section.)

Step four: the bodily self as a functional anchor of phenomenal space

Above, I drew attention to the distinction between the representational and the functional analysis of the first-person perspective. The central theoretical problem on the functional level of description can be summed up by the following question: What exactly is the difference between the PSM and the other phenomenal models that are currently active in the system? Is there a characteristic causal mark of the PSM? Which *functional property* is responsible for turning it into the stable center of phenomenal representational space?

This is my first, preliminary, answer. The self-model is the only representational structure that is anchored in a *continuous source of internally generated input* in the brain. Let us call this the "persistent causal link hypothesis." Whenever conscious experience arises (i.e., whenever a stable, integrated model of reality is activated), this continuous source of internal proprioceptive input also exists. The human self-model possesses an enduring causal link in the brain. It has parts, which in turn are realized by *permanent* forms of information processing on *permanent* forms of self-generated input and low-level autoregulation. To put this general point differently, the body, in certain of it aspects, is the only perceptual object from which the brain can never run away. Again, I will not enter into any amateurish empirical speculation here, but offer a number of obvious candidates for sources of high invariance. Basically, there are four different types of internally generated information that during conscious episodes, constitute a persistent functional link between the PSM and its bodily basis in the brain

- Inputs from the vestibular organ: the sense of balance.
- Inputs from the autonomously active, invariant part of the body schema: the continuous "background feeling" in the spatial model of the body, which is independent of external input, e.g., via motion perception.
- Inputs from the visceral sensors, but also from the blood vessels, for instance from the cardiovascular mechanosensors: "gut feelings" and somatovisceral forms of self-presentation.
- Inputs from certain parts of the upper brain stem and hypothalamus: background emotions and moods, which are anchored in the continuous homeostatic self-regulation of the "internal milieu," the biochemical landscape in our blood.

Philosophically, it is not as much the neurobiological details that are crucial as the highly plausible assumption that there is a certain part of the human self-model that is characterized by a high degree of stimulus correlation and that depends exclusively on *internally* generated information. This layer of the PSM is directly and permanently anchored in stimuli from the inside of the body. Do you still remember patient AZ from example no. 2? The weaker degree of phenomenological "vividness" or "realness" in her phantom limbs may reflect exactly the absence of permanent bottom-up stimulation that in normal situations is caused by existing physical limbs. In this context, Marcel Kinsbourne has spoken of a "*background 'buzz' of somatosensory input*" (Kinsbourne, 1995, p. 217). To capture the phenomenology involved in this sheer "raw feel of embodiment" on the representationalist level of description I like to distinguish between self-presentation and self-representation.[5] Phenomenologically, the first concept is related to the purely sensory feeling of bodily presence, which so interestingly goes along with a subjective sense of temporal immediacy and the experiential certainty of possessing direct, non-inferential self-knowledge. What exactly is this deepest layer

[5]For an extensive theoretical treatment of the subject and numerous recent empirical results on the body as an anchor of conscious experience, see Damasio (1999). Antonio Damasio uses the term of a *core self*, and elsewhere (Metzinger, 1993, p. 156ff; Metzinger, 2003a, Section 5.4) I introduced the technical concept of "phenomenal self-presentation" (as opposed to self-*re*presentation). On the level of body-representation, self-presentation is what AZ lacks in her phantom limbs, whereas self-*re*presentation is what she actually has — although, as the referent of this representation never existed, this obviously is also a form of *mis*representation.

of the phenomenal self? Why is it the origin of the first-person perspective? My hypothesis is that the constant self-organizing activity of those regions of the bodily self that are independent of external input constitutes the functional *center* of phenomenal representational space.

As our first example of how to understand the concept of a self-model, we used the experiment in which Ramachandran managed to mobilize a paralyzed phantom limb. A self-*presentation* is exactly that part of the phantom limb that remains conscious independently of the occurrence of movement. If *this* part is lost, you also lose the subjective experience of bodily presence — you turn into a "disembodied being."[6] But there may even be other, more general empirical perspectives from which the self-model is necessarily related to the baseline of brain activity per se, as it can be observed in the resting state (see Raichle et al., 2001; Gusnard, 2005).

Step five: autoepistemic closure — transparency and the naïve-realistic self-misunderstanding

Back on the *representational* level of analysis, the central theoretical problem is that one might easily accuse me of mislabeling the actual problem by introducing the concept of a "self-model." First, a self-model, of course, is not a model of a mysterious *thing* that we then call the self. It is a continuous and self-*directed* process tracking global properties of the organism. Second, at least according to certain modal intuitions, there appears to be no necessary connection between the fundamental functional and representational properties on the one hand and the *phenomenal* target properties of "mineness," "prereflexive/preagentive selfhood," and "perspectivalness" on the other hand. All this could easily occur without resulting in a real phenomenal self or a subjective inner perspective; it is conceivable that biological information-processing systems could develop and successfully employ a representational space centered by a self-model *without* also developing self-consciousness. More interestingly, even *given the phenomenal level*, i.e., even in a system that is already conscious, it is not obvious or self-evident that the specific phenomenology of *selfhood* should emerge. What would, by logical necessity, bring about an ego? A "self-model" is by no means a self, but only a representation of the system as a whole — it is no more than a *system-model*. If the functional property of centeredness and the representational property of having a self-model are to lead to the phenomenal property of perspectivalness, the conscious system-model must turn into a phenomenal self. The decisive philosophical question is this: How does the existence of a functionally centered representational space necessarily lead to the emergence of a conscious self and what we commonly call a phenomenal first-person perspective? In other words, how does the system-model turn into a *self*-model?

My answer is that a genuinely conscious self emerges at exactly the moment when the system is no longer able to recognize the self-model it is currently generating *as* a model on the level of conscious experience. So, how does one get from the functional property of "centeredness" and the representational property of "self-modeling" to the phenomenal target property of "prereflexive self-intimacy"? The solution has to do with what philosophers call "phenomenal transparency" (for a short explanation of the concept of "phenomenal transparency," see Metzinger, 2003c; Metzinger, 2003b is the German precursor). The conscious representational states generated by the system are *transparent*, i.e., they no longer represent the very fact that they *are* models on the level of their content. Consequently — and this is a phenomenological metaphor only — the system simply looks right "through" its very own representational structures, as if it were in direct and immediate contact with their content. Please note how this is only a statement about the system's *phenomenology*. It is not a statement about epistemology, about the possession of knowledge: you can be completely deluded and have no or very little knowledge about reality (or your own mind) and

[6]Again, the corresponding phenomenological state classes exist. In Metzinger (1993) and Metzinger (1997), I discussed Oliver Sacks' example of the "disembodied lady". In this context, see also the famous case of Ian Waterman, which is discussed in Metzinger (2003a).

at the same time enjoy the phenomenology of certainty, of knowing that you know. Phenomenal transparency is not *epistemic* transparency, or Descartes' classical — and now falsified — idea that we can not be wrong about the contents of our own mind. Transparency, as defined in this context, is exclusively a property of *conscious* states. Unconscious states are neither transparent nor opaque. Phenomenal transparency also is not directly related to the second technical concept in philosophy, to "referential transparency." Non-linguistic creatures incapable of conceptual thought can have phenomenally transparent states as well. Naïve realism is not a belief or an intellectual attitude, but a feature of phenomenal experience itself.

I have two causal hypotheses about the micro-functional underpinnings and the evolutionary history of transparent phenomenal states. First, in a very small time-window, the neural data structures in question are activated so quickly and reliably that the system is no longer able to recognize them as such, for instance due to the comparatively slow temporal resolution of *metarepresentational* functions. Introspectively, the construction process is invisible. Second, in a much larger explanatory time-window, there apparently was no evolutionary pressure on the respective parts of our functional architecture in the process of natural selection. For biological systems like us, naïve realism was a functionally adequate background assumption. We needed to know "Careful, there is a wolf nearby!" but not "A wolf-representation is active in my brain right now!"

Transparency is a special form of darkness. It is a lack of knowledge. Epistemologically speaking, it is an implicit, not an explicit lack of knowledge. As Franz Brentano ([1874] 1973, 165f) and Daniel Dennett (1991, 359) pointed out, the representation of absence is not the same thing as the absence of representation. In transparent states, there is no representation of earlier processing stages. In the phenomenology of visual awareness, it means not being able to see something. Phenomenal transparency *in general*, however, means that the representational character of the contents of conscious experience itself is not accessible to subjective experience. This analysis can be applied to all of the sensory modalities, especially to the integrated phenomenal model of the world as a whole. Because the very *means* of representation cannot be represented as such, the experiencing system necessarily becomes entangled in naïve realism; it experiences itself as being directly in contact with the contents of its own conscious experience. It is unable to experience the fact that all of its experiences take place in a *medium* — and this is exactly what we mean by the "immediacy" of phenomenal consciousness. In a completely transparent representation, the very mechanisms that lead to its activation as well as the fact that its contents depend on a concrete inner state as a carrier can no longer be recognized by way of introspection. As philosophers like to say: "Only content properties are introspectively accessible, vehicle properties are inaccessible." Therefore, the phenomenology of transparency is the phenomenology of naïve realism.

Many phenomenal representations are transparent because their content and its very existence appear to be fixed in all possible contexts. According to subjective experience, the book you are currently holding in your hands will always stay the same book — no matter how the external perceptual conditions vary. You never have the experience that an "active object emulator" in your brain is currently being integrated into your global reality-model. You simply experience the *content* of the underlying representational process: the *book* as effortlessly given, here and now. The best way to understand the concept of transparency is to distinguish between the vehicle and the content of a representation, between representational carrier and representational content (see also Dretske, 1998, p. 45ff).

The representational carrier of your conscious experience is a particular brain process. This process — that itself is in no way "book-like" — is not consciously experienced; it is transparent in the sense that phenomenologically, you look right through it. What you look *at* is its representational content, the perceptually mediated existence of a book, here and now. In other words, this content is an abstract property of a concrete representational state in your brain. If the representational carrier is a good and reliable instrument for the generation

of knowledge, its transparency allows you to ''look right through'' it out into the world, at the book in your hands. It makes the information it carries globally available without your having to worry about *how* this actually happens. What is special about most phenomenal representations is that you experience their content as maximally *concrete* and unequivocal, as directly and immediately given even when the object in question — the book in your hands — does not really exist at all, but is only a hallucination. Phenomenal representations appear to be exactly that set of representations for which we cannot distinguish between representational content and representational carrier on the level of subjective experience.

Of course, there are counterexamples, and they may help further illustrate the concept of ''transparency.'' For instance, *opaque* phenomenal representations arise when the information *that* their content is the result of an internal representational process suddenly becomes globally available. If you suddenly discover that the book in your hands does not really exist, the hallucination turns into a pseudohallucination. The information that you are not looking at the world, but rather ''at'' an active representational state that apparently is not functioning as a reliable instrument for the generation of knowledge at this moment, now also becomes available, and it does so on the level of subjective experience itself. The phenomenal book state becomes opaque. You lose *sensory* transparency. You become aware of the fact that your perceptions are generated by your sensory system and that this system is not always completely reliable. Not only do you now suddenly experience the book as a representation, you also experience it as a *mis*representation.

Let us further assume that you suddenly discover that not only your perception of the book, but all of your philosophical thoughts about the problem of consciousness are taking place in a dream. Then, this dream would turn into a lucid dream (for a discussion of the reasons for regarding lucid dreams as a philosophically relevant class of conscious states, see Metzinger, 2003a, Section 7.2.5; more on the topic can be found in Windt and Metzinger, 2007). The fact that you are currently not experiencing a world, but only a *world-model* would become globally available; now, you could use this information to control your actions, thoughts, and the direction of attention. You would lose *global* transparency. The interesting point, however, is that cognitive availability alone is not sufficient to dissolve the naïve realism of phenomenal experience. You cannot simply ''think'' yourself out of your phenomenal model of reality by changing your opinions about this model: the transparency of phenomenal representations is cognitively impenetrable; here, phenomenal knowledge is not the same as conceptual/propositional knowledge.

Now, the final step is to apply this insight to the self-model. Here is my key claim— we are systems that are experientially unable to recognize our own subsymbolic self-model *as* a self-model. For this reason, phenomenologically, we operate under the conditions of a ''naïve-realistic self-misunderstanding''; we experience ourselves as being in direct and immediate epistemic contact with ourselves. By logical necessity, a phenomenally transparent self-model will create the experience of *being infinitely close to yourself*. The core of the self-model theory is that this is how the basic sense of selfhood arises and how a phenomenal self that is untranscendable for the respective system comes about. The content of non-conceptual self-consciousness is the content of a transparent PSM. It also commits me to a specific prediction: Were the PSM to loose its transparency and become opaque, were the organism as a whole capable of recognizing its current self-model *as* a model, then the phenomenal property of selfhood would disappear. In standard phenomenological configurations, however, the entity that looks at the book in its hands is itself a form of transparent phenomenal content. And this is also true of the ''at''-ness inherent in this act of visual attention, of the relation that seems to connect subject and object.

Step six: the PMIR — the phenomenal model of the intentionality relation

Let us take one more step before we close. The experience of selfhood is intimately related not only to the sense of ownership, but also to the

experience of agency; it is not only a question of having a transparent self-model, but also of directedness, of being dynamically related to target objects and goal states. Here are two further examples, this time from yet another academic discipline — experimental neuroscience using macaque monkeys as subjects.

Classical neurology hypothesized about a ''body schema,'' an unconscious, but constantly updated map of body shape and posture in the brain.[7] Recent research shows how Japanese macaque monkeys can be trained to use tools even though they only rarely exhibit tool-use in their natural environment (see Maravita and Iriki, 2004, for a good review). During successful tool-use, changes in specific neural networks in their brains take place — a finding that suggests that the tools are temporarily integrated into their body schema. When a food pellet is dispensed beyond the reach of their hands and they skillfully use a rake to pull it closer, one can observe a change in their bodily self-model in the brain. In fact, it looks as if their conscious model of their hand had been expanded towards the tip of the tool. A more precise way of describing what happens is to say that, on the level of the monkey's conscious model of reality, properties of the hand are now transferred to the distant tip of the tool. We know the same effect in human beings. In our own case, repeated practice can turn the tip of a tool into a part of our own hand, a part that can be used just as ''sensitively'' and as skillfully as our own fingers.

In other words, recent neuroscientific data clearly support the view that tools not only enable us to extend our reaching space. They show that any successful extension of behavioral space is also mirrored in the neural substrate of the body image in the brain. The brain constructs an ''internalized'' image of the tool by swiftly assimilating it into the existing body image as a whole. Of course, we do not know if monkeys actually have the conscious experience of ownership or only the unconscious mechanism. But we do know about several similarities between macaques and humans that make this assumption seem plausible. This may be the very beginning of mentally *simulating* yourself as currently being directed at a target object or goal state. And this leads us to second major aspect of selfhood: besides global *ownership* what we need to understand is *agency* — global control.

One exciting aspect of these new data is that they shed light on the evolution of tool-use. A necessary precondition of expanding your space of action and your capabilities by using tools clearly seems to be the ability to integrate them into a preexisting self-model. You can only engage in goal-directed and intelligent tool-use if your brain temporarily represents them as part of your own self. Intelligent tool-use was a major achievement in human evolution. One may plausibly assume that some elementary building block of human tool-use abilities already existed in the brains of our ancestors. Then, due to some not-yet-understood evolutionary pressure, it rapidly expanded into what we see in humans today.[8]

There is a new, rapidly growing field of research in which engineers and neuroscientists work together: brain–machine interfaces (for a brief overview, see Lebedev and Nicolelis, 2006). One application of this general idea consists in driving and controlling artificial limbs or robotic manipulators with the help of ensembles of cortical neurons, allowing a machine to carry out motor commands generated in the brain. The following figure shows an example from the Duke University Center for Neuroengineering, demonstrating the general principle (Fig. 10).

In our context, the perhaps most interesting observation in this experiment (see Carmena et al., 2003, for details) is how the monkey gradually begins to neglect his original arm, which is, after all, a part of his biological body. That is, as he now tries to control feedback in a new kind of motor task and with a different goal-state, optimizing a

[7]The terminology was never entirely clear, but it frequently differentiated between an unconscious ''body schema'' and a conscious ''body image.'' For a philosophical perspective on the conceptual confusion surrounding both notions, see Gallagher (2005); for an excellent review of the empirical literature, see Maravita (2006).

[8]See Iriki et al. (1996); Maravita and Iriki (2004).

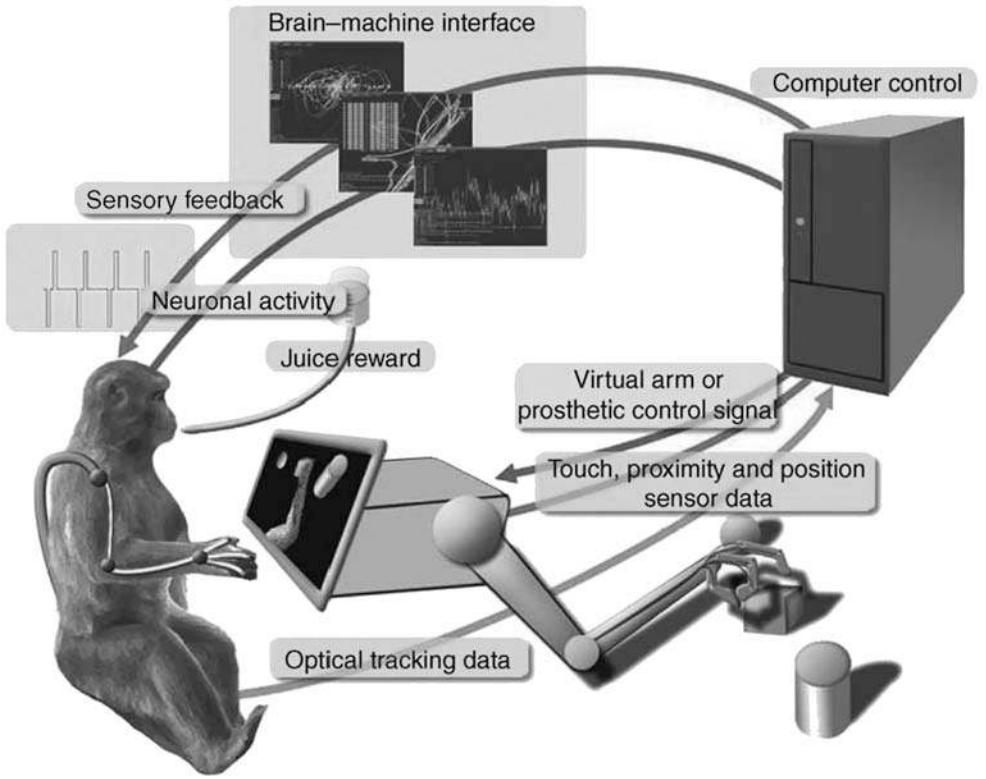


Fig. 10. A BMI with multiple feedback loops that is currently being developed at the Duke University Center for Neuroengineering. A rhesus macaque operates an artificial robotic manipulator that reaches for and grasps different objects. The manipulator is equipped with touch, proximity, and position sensors. Signals from the sensors are delivered to the control computer (right), which processes them and converts them to microstimulation pulses delivered to the sensory areas in the monkey's brain, providing it with feedback information (red loop). A series of microstimulation pulses is illustrated in the inset on the left. Neuronal activity is recorded in multiple brain areas and translated into commands to the actuator via the control computer and multiple decoding algorithms (blue loop). The arm position is monitored using an optical tracking system that tracks the position of several markers mounted on the arm (green loop). The hypothesis is that the continuous operation of this interface would lead to the incorporation of the external actuator into the representation of the body in the brain. (Figure designed by Nathan Fitzsimmons.) (See Color Plate 18.10 in color plate section.)

new set of motor parameters by trying to control a real-world robot arm or even a virtual arm he sees on the screen in front of him, his brain seems to undergo certain changes — the ''tuning properties'' of neurons change. Here is how Lebedev and Nicolelis (2006, p. 542) describe the effect: ''Remarkably, after these animals started to control the actuator directly using their neuronal activity, their limbs eventually stopped moving, while the animals continued to control the actuator by generating proper modulations of their cortical neurons. The most parsimonious interpretation of this finding is that the brain was capable of undergoing a gradual assimilation of the actuator within the same maps that represented the body.''

From the perspective of SMT, the self-model theory, the most plausible interpretation is that, once the monkey has successfully embedded an internal representation of this new actuator into his conscious self-model, the representations of his old body parts lose certain functional properties, they transiently becomes less and less available for attentional processing and gradually recede from conscious experience. These examples teach us two further important insights. Very obviously, the PSM is an important part of a *control hierarchy*; it is a means to monitor certain critical aspects of the

process by which the organism generates flexible, adaptive patterns of behavior; second, it is highly plastic in the sense that multiple representations of objects *outside* the body can transiently be integrated into it. This is not only true of rubber-hands, but even more so of tools in the most general sense — extensions of bodily organs which must be successfully controlled in order to generate intelligent, goal-directed behavior. The self-model is the functional window through which the brain can interact with the body as a whole, and vice versa. If the body is augmented by sticks, stones, rakes, or robot arms, the self-model has to be extended. If an integrated representation of body-plus-tool is in existence, the extended system of body-plus-tool can become part of the brain's control hierarchy. After all, how *could* one learn to intelligently use a tool *without* integrating it into the conscious self? The conscious self-model is a virtual organ that allows us to *own* feedback loops, to initiate, sustain, and flexibly adapt control processes. Some elements of the control loop are physical (such as the brain and tools); others are virtual (such as the self-model and goal-state simulation).

In passing, let me briefly emphasize one further point. In human beings (and some other animals as well), it is frequently the behavior or mental state of *another* person that is to be controlled. We ''instrumentalize'' and ''appropriate'' each other. Human beings constantly augment themselves not only with sticks, stones, rakes, or robot arms — but also with the brains and bodies of *other* human beings (Metzinger and Gallese, 2003). Clearly, the transition form biological to cultural evolution is intimately connected with the appearance of new and specific functional properties in the primate-PSM. This is one of the most interesting questions for the future: What exactly was the change in the PSM of Homo, as opposed to the PSM of the chimpanzee, which lead to the explosion of culture and the emergence of complex societies? Here, my own speculative working hypothesis would be that it was not complex tool use *per se*, but the ability to take a much larger part of the control hierarchy *offline*, to use it in simulation, while at the same time generating an opaque (i.e., a non-transparent) PSM. It was the ability to consciously represent yourself *as* representing, *as* being directed at a goal state. It was the difference between having a first-person perspective and the mental capacity to explicitly represent this very fact.

Now, let us take a look at the representational architecture underlying the subjective experience of directedness in general. Phenomenologically, a transparent world-model gives rise to a reality. A transparent system-model gives rise to a self that is embedded in this reality. If there is also a transparent model of the transient and constantly changing relations between the perceiving and acting self and the objects and persons in this reality, this results in what I called a ''phenomenal first-person perspective'' above. A genuine inner perspective arises if only and only if the system represents itself as currently interacting with the world *to itself*, and if it does not recognize this representation *as* a representation. Now, it has a conscious model of the intentionality relation (a PMIR). It represents itself as directed towards certain aspects of the world. Its phenomenal space is a *perspectival* space, and its experiences are *subjective* experiences.

The intentionality relation is primarily an epistemic relation between subject and object. A mental state becomes a carrier of knowledge in virtue of being directed at something other than itself — like an arrow pointing from a person's mind to an object in the real or even just in a possible world. Philosophers say that this type of mental state has *intentional content*. Its content is what the arrow is pointing at. This may be an image, a proposition, or even the goal of an action — as philosophers say, there is ''practical intentionality'' in terms of your being directed at certain ''satisfaction conditions'' (e.g., an action goal), and there is ''theoretical intentionality'' in terms of being directed at the ''truth conditions'' (e.g., of a sentence). If many of these arrows are consciously available, represented by the brain on the functional level of global availability, this results in a temporally extended first-person perspective. In short, it is one thing to be a biological organism that represents the world, and it is another thing to consciously represent yourself *as representing*, in ''real-time'' and while this is actually happening. SMT wants to understand the latter case. Now,

there is not only a neurobiologically anchored core self, a self-presentation, but also a dynamic phenomenal simulation of the *self as subject* embedded in the world via constantly changing epistemic relations and agentive interactions. Of course, there is much more to be said about the central notion of a PMIR.[9] But the core idea is as follows: a conscious human being is a system that is capable of dynamically *co-representing* the representational relation while representational acts are taking place, and the instrument it uses for this purpose is the PMIR. The phenomenal model of the intentionality relation (PMIR), is just another naturally evolved virtual organ, just like the PSM. The content of higher order forms of self-consciousness is always relational: the self *in the act of knowing* (Damasio, 1999, p. 168ff), the *currently acting* self. The ability to co-represent this intentional relationship itself while actively constructing it in interacting with a world is what it means to be a subject.

Of course, the way we subjectively experience this subject–object relation is a simplified version of the actual processes — in a sense, it is a functionally adequate confabulation. Once again, evolution favored a simple, but elegant solution. The virtual self-moving through the phenomenal world does not have a brain, a motor system, or sensory organs: certain parts of the environment appear directly in its mind; the perceptual process is experienced as effortless and immediate. Body movements also appear to be caused ''directly.'' Such effects are typical for *our* type of subjective experience and — seen as a neurocomputational strategy — they have the advantage of creating a user-friendly interface. What was defined as ''transparency'' above is a way of describing the *closed* structure of this multimodal, high-dimensional user interface — the brain's user surface. The phenomenal self is the part of this interface that the system uses to experience *itself* as a whole, to represent itself as a thinking self and an agent. This virtual agent ''sees with his eyes'' and ''acts with his hands.'' He does not know that he has a visual or a motor cortex. The PSM is the interface that the system uses to functionally appropriate its own hardware, to control its own low-level dynamics and to become *autonomous*. The intentional arrows connecting this agent to objects and other selves in the currently active reality-model are phenomenal representations of transient subject–object relations — and frequently, they too cannot be recognized *as* representational processes. In standard situations, the consciously experienced first-person perspective is the content of a transparent PMIR.

All this takes place within a phenomenal window of presence. The contents of phenomenal experience not only create a world; they also create a *present* (see Metzinger, 2003a, Section 3.2.2). In a sense, the core of phenomenal consciousness is just the creation of an island of presence within the physical flow of time (see Ruhnau, 1995 and the references given there, especially to the work of Ernst Pöppel). Experiencing means ''being there,'' and this necessarily includes ''being now.'' It means processing information in a very specific way. It means repeatedly and continuously binding discrete events that have already been represented as such into temporal *gestalts*, into a consciously experienced moment. Many recent empirical data clearly demonstrate that in a certain sense, the consciously experienced present is a *remembered* present (see for instance, Edelman, 1989). In this sense, even the phenomenal ''Now'' is a representational construct, a *virtual* present. And this finally helps understand what it means to say that phenomenal space is a virtual space: its content is a *possible* reality.[10] The realism of phenomenal experience arises because it represents a possibility — the best hypothesis there is at a given moment — as an untranscendable reality, or an *actuality*. In other words, the mechanisms creating temporal

[9]Of course, the theory of the PMIR is more complex than I can explain in this brief overview. Apart from Metzinger (2003a), I recommend Section 4 of Metzinger (2005a, p. 26ff) for readers interested in the idea. A more detailed discussion, specifically applied to the representational architecture of conscious volitional acts, can be found in Metzinger (2006a).

[10]My own ideas on this point are very similar to those discussed by Antti Revonsuo: *Virtual reality* is simply the best technological metaphor for phenomenal consciousness we currently have. See Revonsuo (1995, 2000), and especially Revonsuo (2006).

experience and our subjective sense of presence are transparent as well. Then, finally, this point also has to be applied to the special case of self-modeling because the virtual character of both the self-model *and* the window of presence are not available on the level of subjective experience itself, the system they represent turns into a *currently present subject*.

SMT solves the homunculus problem, because we can now see how no "little man in the head" is needed to interpret and "read out" the content of mental representations. It is also maximally parsimonious, as it allows us to account for the emergence of self-consciousness without assuming the existence of a substantial self. Does all this mean that the self is only an illusion? On second glance, the popular concept of the "self-illusion" and the metaphor of "mistaking oneself for one's inner picture of oneself" contain a logical error: *Whose* illusion could this be? Speaking of illusions presupposes someone *having* them. But something that is not an epistemic subject in a strong sense of conceptual/propositional knowledge is simply *unable* to confuse itself with anything else. Truth and falsity, reality and illusion do not exist for biological information-processing systems at the developmental stage in question. So far, we only have a theory of the phenomenology of selfhood, not a theory of self-knowledge. Here, I have only very briefly sketched how a *phenomenal* first-person perspective can be the product of natural evolution. Subjectivity in an *epistemic* sense, an epistemic first-person perspective is yet another step. Of course, the phenomenology of selfhood, of non-conceptual self-consciousness, is the most important precondition for this step, because it is the precondition for genuinely *reflexive*, conceptual self-consciousness. In a way, this is the whole point behind the theory: if we want to take high-level forms of subjectivity and intersubjectivity seriously, we must be modest and careful at the beginning, focusing on their origins on the level of non-conceptual content and self-organizing neural dynamics. And readers will not be surprised that the author of this chapter holds that subjective, first-person *knowledge* is precisely knowledge associated with a specific inner mode of presentation, namely as knowledge *under a PMIR*. Subjectivity in the epistemological sense can be naturalized as well — but only if we can tell a convincing evolutionary and neuroscientific story about how this representational architecture, this highly specific, indexical inner mode of presentation, could actually have developed in a self-organizing physical universe in the first place. Ultimately, and obviously, every single instance of the PSM/PMIR is identical with a specific time-slice in the continuous, dynamical self-organization of coherent activity taking place in an individual biological brain. In this ongoing process on the subpersonal level there is no agent — no evil demon that could count as the *creator* of an illusion. And there is no entity that could count as the *subject* of the illusion either. There is nobody *in* the system who could be mistaken or confused about anything — the homunculus does not exist. On the level of phenomenology, as well as on the level of neurobiology, the conscious self is neither a form of knowledge nor an illusion. It just is what it is.

Acknowledgment

I am greatly indebted to Jennifer M. Windt for help with the English version and to Rahul Banerjee for very stimulating discussion and a number of helpful critical comments.

References

Alvarado, C.S. (1986) Research on spontaneous out-of-body experiences: a review of modern developments, 1960–1984. In: Shapin B. and Coly L. (Eds.), Current Trends in PSI Research. Parapsychology Foundation, New York.

Alvarado, C.S. (2000) Out-of-body experiences. In: Cardeña E., Lynn S.J. and Krippner S. (Eds.), Varieties of Anomalous Experience: Examining the Scientific Evidence. American Psychological Association, Washington, DC.

Armel, K.C. and Ramachandran, V.S. (2003) Projecting sensations to external objects: evidence from skin conductance response. Proc. R. Soc. Lond. Biol., 270: 1499–1506.

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

Baker, L.R. (1998) The first-person perspective: a test for naturalism. Am. Philos. Q., 35: 327–346.

Baker, L.R. (2000) Die Perspektive der ersten Person: Ein Test für den Naturalismus. In: Keil G. and Schnädelbach H. (Eds.), Naturalismus-Philosophische Beiträge. Suhrkamp, Frankfurt am Main.

Baker, L.R. (2007) Naturalism and the first-person perspective. In: Gasser G. (Ed.), How Successful is Naturalism? Publications of the Austrian Ludwig Wittgenstein Society. Ontos Verlag, Frankfurt am Main.

Bieri, P. (1987) Evolution, Erkenntnis und Kognition. In: Lütterfelds W. (Ed.), Transzendentale oder Evolutionäre Erkenntnistheorie? Wissenschaftliche Buchgesellschaft, Darmstadt.

Bischof-Köhler, D. (1989) Spiegelbild und Empathie. Huber, Bern, Nachdruck.

Bischof-Köhler, D. (1996) Ichbewusstsein und Zeitvergegenwärtigung. Zur Phylogenese spezifisch menschlicher Erkenntnisformen. In: Barkhaus A., Mayer M., Roughley N. and Thürnau D. (Eds.), Identität, Leiblichkeit, Normativität. Neue Horizonte anthropologischen Denkens. Suhrkamp, Frankfurt am Main.

Blackmore, S. (1982) Beyond the Body: An Investigation of Out-of-the-Body-Experiences. Granada, London.

Blackmore, S.J. (1986) Spontaneous and deliberate OBEs: a questionnaire survey. J. Soc. Psych. Res., 53: 218–224.

Blanke, O., Landis, T., Spinelli, L. and Seeck, M. (2004) Out-of-body experience and autoscopy of neurological. Brain, 127: 243–258.

Blanke, O., Mohr, C., Michel, C.M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T. and Thut, G. (2005) Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. J. Neurosci., 25(3): 550–557.

Bongard, J., Zykov, V. and Lipson, H. (2006). Resilient machines through continuous self-modeling. Science, 314: 1118. In particular, see also free online support material at http://www.sciencemag.org/cgi/content/full/314/5802/1118/DC1

Botvinick, M. (2004) Probing the neural basis of body ownership. Science, 305: 782–783.

Botvinick, M. and Cohen, J. (1998) Rubber hand "feel" touch that eyes see. Nature, 391: 756.

Brentano, F. (1973) [1874]. Psychologie vom empirischen Standpunkt. Erster Band. Meiner, Hamburg.

Brugger, P., Regard, M. and Landis, T. (1997) Illusory reduplication of one's own body: phenomenology and classification of autoscopic phenomena. Cognit. Neuropsychiatry, 2: 19–38.

Brugger, P., Kollias, S.K., Müri, R.M., Crelier, G., Hepp-Reymond, M.-C. and Regard, M. (2000) Beyond remembering: phantoms sensations of congenitally absent limbs. Proc. Natl. Acad. Sci. U.S.A., 97: 6167–6172.

Brugger, P., Regard, M. and Shiffrar, M. (2001) Hand movement observation in a person born without hands: is body scheme innate? J. Neurol. Neurosurg. Psychiatry, 70: p. 276.

Carmena, J.M., Lebedev, M.A., Crist, R.E., O'Doherty, J.E., Santucci, D.M., Dimitrov, D.F., Patil, P.G., Henriquez, C.S. and Nicolelis, M.A.L. (2003) Learning to control a brain–machine interface for reaching and grasping by primates. PLoS Biol., 1: 193–208.

Churchland, P.M. (1989) A Neurocomputational Perspective. MIT Press, Cambridge, MA.

Clark, A. (1989) Microcognition-Philosophy, Cognitive Science, and Parallel Distributed Processing. MIT Press, Cambridge, MA.

Cummins, R. (1983) The Nature of Psychological Explanation. MIT Press, Cambridge, MA.

Damasio, A. (1994) Descartes' Error. Putnam/Grosset, New York.

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

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

Dennett, D.C. (1991) Consciousness explained. Little, Brown and Company, Boston, Toronto and London.

Devinsky, O., Feldmann, E., Burrowes, K. and Bromfield, E. (1989) Autoscopic phenomena with seizures. Arch. Neurol., 46: 1080–1088.

Dittrich, A. (1985) Ätiologie-unabhängige Strukturen veränderter Wachbewusstseinszustände. Enke, Stuttgart.

Dretske, F. (1988) Explaining Behavior: Reasons in a World of Causes. MIT Press, Cambridge, MA.

Dretske, F. (1998) Die Naturalisierung des Geistes. Mentis, Paderborn.

Edelman, G.M. (1989) The Remembered Present: A Biological Theory of Consciousness. Basic Books, New York.

Ehrsson, H.H., Spence, C. and Passingham, R.E. (2004) That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. Science, 305: 875–877.

Gallagher, S. (2005) How the Body Shapes the Mind. Oxford University Press, Oxford.

Gallese, V. (2005) Embodied simulation: from neurons to phenomenal experience. Phenomenol. Cognit. Sci., 4: 23–38.

Gallese, V. and Goldman, A. (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci., 2: 493–501.

Grush, R. (1997) The architecture of representation. Philos. Psychol., 10: 5–25.

Grush, R. (1998) Wahrnehmung, Vorstellung, und die sensomotorische Schleife. In: Heckmann H.-D. and Esken F. (Eds.), Bewusstsein und Repräsentation. Mentis, Paderborn.

Gusnard, D. (2005) Being a self: considerations from functional imaging. Conscious. Cogn., 14(4): 679–697.

Iriki, A., Tanaka, M. and Iwamura, Y. (1996) Coding of modified body schema during tool-use by macaque postcentral neurons. Neuroreport, 7: 2325–2330.

Irwin, H. (1985) Flight of mind. Scarecrow Press, Metuchen, NJ and London.

Kinsbourne, M. (1995) Awareness of one's own body: an attentional theory of its nature, development, and brain basis. In: Bermúdez J.L., Marcel A. and Eilan N. (Eds.), The Body and the Self. MIT Press, Cambridge, MA.

Lebedev, M.A. and Nicolelis, M.A.L. (2006) Brain–machine interfaces: past, present and future. Trends Neurosci., 29(9): 536–546. (http://www.science-direct.com/science/article/B6T0V-4KFV367-2/2/1c479d63267ad95d2a57070bfd516003)

Lenggenhager, B., Tadi, T., Metzinger, T. and Blanke, O. (2007) Video ergo sum: manipulating bodily self-consciousness. Science, 317(5841).

Lycan, W.G. (1996) Consciousness and Experience. MIT Press, Cambridge, MA.

Maravita, A. (2006) From "body in the brain" to "body in space": sensory and intentional components of body representation. In: Knoblich G., Thornton I., Grosjean M. and Human Shiffrar M. (Eds.), Body Perception from the Inside Out. Oxford University Press, New York.

Maravita, A. and Iriki, A. (2004) Tools for the body (schema). Trends Cogn. Sci., 8: 79–86.

Melzack, R. (1989) Phantom limbs, the self and the brain: The D.O. Hebb memorial lecture. Can. Psychol., 30: 1–16.

Melzack, R. (1992) Phantom limbs. Sci. Am., 266: 90–96.

Melzack, R., Israel, R., Lacroix, R. and Schultz, G. (1997) Phantom limbs in people with congenital limb deficiency or amputation in early childhood. Brain, 120(Pt. 9): 1603–1620.

Metzinger, T. (1993; ^{2}1999). Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewusstseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation. Mentis, Paderborn.

Metzinger, T. (1995a). Perspektivische Fakten? Die Naturalisierung des "Blick von nirgendwo". In: Meggle G. and Nida-Rümelin J. (1997) (Eds.), ANALYOMEN 2: Perspektiven der Analytischen Philosophie. de Gruyter, Berlin, pp. 103–110.

Metzinger, T. (Ed.) (1995b). Conscious Experience. Imprint Academic, Thorverton, UK.

Metzinger, T. (1996) Niemand sein. In: Krämer S. (Ed.), Bewusstsein-Philosophische Positionen. Suhrkamp, Frankfurt am Main.

Metzinger, T. (1997) Ich-Störungen als pathologische Formen mentaler Selbstmodellierung. In: Northoff G. (Ed.), Neuropsychiatrie und Neurophilosophie. Mentis, Paderborn.

Metzinger, T. (2000). The subjectivity of subjective experience: a representationalist analysis of the first-person perspective. In: Metzinger T. (Ed.), Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press, Cambridge, MA. Revised version (2004): Networks, 3–4: 33–64.

Metzinger, T. (2003a; ^{2}2004). Being No One. The Self-Model Theory of Subjectivity. MIT Press, Cambridge, MA.

Metzinger, T. (2003b) Phänomenale Transparenz und kognitive Selbstbezugnahme. In: Haas-Spohn U. (Ed.), Intentionalität zwischen Subjektivität und Weltbezug. Mentis, Paderborn, pp. 411–459.

Metzinger, T. (2003c) Phenomenal transparency and cognitive self-reference. Phenomenol. Cogn. Sci., 2: 353–393. doi: 10.1023/B:PHEN.0000007366.42918.eb

Metzinger, T. (2004a) Why are identity-disorders interesting for philosophers? In: Schramme T. and Thome J. (Eds.), Philosophy and Psychiatry. de Gruyter, Berlin.

Metzinger, T. (2004b). Appearance is not knowledge: the incoherent strawman, content–content confusions and mindless conscious subjects. Invited commentary for Alva Noë und Evan Thompson: "Are there neural correlates of consciousness?" in a special issue of J. Conscious. Stud., 11(1): 67–71.

Metzinger, T. (2005a). Précis of "Being No One". In: PSYCHE: An Interdisciplinary Journal of Research on Consciousness, 11(5): 1–35. URL: www.psyche.cs.monash.edu.au/

Metzinger, T. (2005b) Out-of-body experiences as the origin of the concept of a "soul". Mind Matter, 3(1): 57–84.

Metzinger, T. (2005c) Die Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung in sechs Schritten. In: Herrmann C.S., Pauen M., Rieger J.W. and Schicktanz S. (Eds.), Bewusstsein: Philosophie, Neurowissenschaften, Ethik. UTB/Fink, Stuttgart.

Metzinger, T. (2006a) Conscious volition and mental representation: towards a more fine-grained analysis. In: Sebanz N. and Prinz W. (Eds.), Disorders of Volition. MIT Press, Cambridge, MA, S19–S48.

Metzinger, T. (2006b). Reply to Gallagher: different conceptions of embodiment. In: PSYCHE: An Interdisciplinary Journal of Research on Consciousness, 12(4). URL: www.psyche.cs.monash.edu.au/symposia/metzinger/reply_to_;Gallagher.pdf

Metzinger, T. and Gallese, V. (2003). The emergence of a shared action ontology: building blocks for a theory. In: Knoblich G., Elsner B., von Aschersleben G. and Metzinger T. (Eds.), Self and Action. Special issue of Conscious. Cogn., 12(4): 549–571.

Millikan, R.G. (1984) Language, Thought, and other Biological Categories. MIT Press, Cambridge, MA.

Millikan, R.G. (1993) White Queen Psychology and Other Essays for Alice. MIT Press, Cambridge, MA.

Muldoon, S. and Carrington, H. (1929) The projection of the astral body. Rider and Co., London.

Nagel, T. (1986) The View from Nowhere. Oxford University Press, New York.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A. and Shulman, G.L. (2001) A default mode of brain function. Proc. Natl. Acad. Sci. U.S.A., 98: 676–682.

Ramachandran, V.S. and Blakeslee, S. (1998) Phantoms in the Brain. William Morrow and Company, Inc., New York.

Ramachandran, V.S. and Rogers-Ramachandran, D. (1996) Synaesthesia in phantom limbs induced with mirrors. Proc. R. Soc. Lond. B, 377–386.

Revonsuo, A. (1995) Consciousness, dreams, and virtual realities. Philos. Psychol., 8: 35–58.

Revonsuo, A. (2000) Prospects for a scientific research program on consciousness. In: Metzinger T. (Ed.), Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press, Cambridge, MA.

Revonsuo, A. (2006) Inner Presence. MIT Press, Cambridge, MA.

Ruhnau, E. (1995). Time-Gestalt and the observer. In: Metzinger T., (Ed.), Conscious Experience. Imprint Academic, Thorverton.

O'Shaughnessy, B. (1995) Proprioception and the body image. In: Bermúdez J.L., Marcel A. and Eilan N. (Eds.), The Body and the Self. MIT Press, Cambridge, MA.

Vuilleumier, P., Despland, P.A., Assal, G. and Regli, F. (1997) Héautoscopie, exta-se et hallucinations expérientielles d'origine épileptique. Rev. Neurol., 153: 115–119.

Waelti, E. (1983) Der dritte Kreis des Wissens. Ansata, Interlaken.

Windt, J.M. and Metzinger, T. (2007) The philosophy of dreaming and self-consciousness: what happens to the experiential subject during the dream state? In: Barrett D. and McNamara P. (Eds.), The New Science of Dreaming. Praeger Imprint/Greenwood Publishers, Westport, CT.

Yates, J. (1975) The content of awareness is a model of the world. Psychol. Rev., 92: 249–284.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 19

Vipassana: the Buddha's tool to probe mind and body

Dhananjay V. Chavan*

Vipassana Research Institute, Igatpuri 422403, India

Abstract: The chapter reviews certain practical aspects related to Vipassana , a first-person method taught by the Buddha, to probe one's stream of conscious experience, its contents and related material events. Although references to Vipassana are found scattered throughout the Pali literature the most comprehensive and detailed account with regard to its practice is contained in the *Satipatthana Sutta*. This chapter summarizes some salient features of Vipassana practice as found in the Sutta. It could be possible that formal first-person methods will be found necessary to track the succession of internal subjective states even in the scientific study of consciousness. A review of classical methods could provide directions for further studies.

Keywords: first-person methods; Vipassana; concentration; meditative states; Abhidhamma; awareness

Introduction

When young Siddhartha Gautama left his home on a historic journey 26 centuries ago, he set out to seek an end to suffering which he saw was a universal characteristic of human existence. He had little interest in metaphysical speculation. During his efforts to find a practical solution, he discovered that an understanding of the nature of body, mind and their mutual interaction would be necessary to resolve the issue. This understanding became central to his enquiry. After 6 years of relentless effort he elucidated and subsequently taught a clear and comprehensive set of first-person methods to eradicate unhappiness and gain direct insight into the processes underlying conscious experience. He consistently maintained that this understanding could be developed by one and all, by means of a specialized first-person technique, namely Vipassana, implemented within the framework of the mind and body.

Since human unhappiness is ubiquitous, the Buddha was in search of a remedy which would find universal application without regard to race, creed, gender or a specific cultural milieu. The history of this attempt has been fairly well documented and is arguably ''the first historical attempt to map the human mind in a thorough and realistic way without admixture of metaphysics and mythology'' (Nyanaponika Thera, 1998). Thus in spirit and approach Buddha was essentially 'scientific', if by the term we denote a free enquiry undertaken without recourse to either dogma or preconceived metaphysical beliefs, the results of which can be cross-validated by peers. This is borne out in an oft-repeated dialogue between Buddha and the Kalamas, a tribal group residing in northern India, who were quite

*Corresponding author. Tel.: +91 94235 78087;
Fax: +91 2553 244176; E-mail: dhananjay65@gmail.com

DOI: 10.1016/S0079-6123(07)68019-4

bewildered by the plethora of philosophical views in vogue during that time.

> 'Come, Kalamas, do not simply believe whatever you are told, or whatever has been handed down in tradition, or what is common opinion, or whatever the scriptures say. Do not accept something as true merely by deduction or inference, or by considering outward appearances, or by partiality for a certain view, or because of its plausibility, or because your teacher tells you it is so. But when you yourselves directly know, ''These principles are unwholesome, blameworthy, condemned by the wise; when adopted and carried out they lead to harm and suffering,'' then you should abandon them. And when you yourselves directly know, ''These principles are wholesome, blameless, praised by the wise; when adopted and carried out they lead to welfare and happiness for me and for others,'' then you should accept and practise them'. Anguttara Nikaya (1.3.66, 1998a)

As here so elsewhere, Buddha's emphasis was invariably on direct knowledge from a first-person perspective ('…when you yourselves directly know.'), the effort being made independently by each individual.

> 'You yourselves have to walk on the path and make efforts. The Buddhas can only show the path'. Dhammapada (276, 1998)

The last decade has seen the reemergence of consciousness as a discipline of scientific study. However, one aspect of conscious experience which distinguishes it from all other branches of science, is its subjectivity. Apparently, the only stream of consciousness we have direct access to is our own. Whatever may be one's view of consciousness in the natural order, there is probably a need for systematic first-person methods to study our subjective mental states and correlate them to physical states (brain states) which can be empirically characterized. Given the bewildering variety and range of conscious mental states it seems unlikely that any methodical observation can be made of one's subjectivity without proper training and grounding in formalized first-person methods. In such a scenario, a review of classical meditative traditions such as Vipassana, may provide fruitful directions for further development.

Before the details of Vipassana practice are discussed it would be appropriate to mention that Buddha's approach to the mind–body problem was essentially interactionist. Advanced Vipassana meditation led him to conclude that an intricate network of causal relationships governed the interaction between different mental states as well as between physical and mental states. Buddha consistently maintained that to understand the chain of causation underlying mental and physical events was the primary aim of Vipassana.

> 'Anyone who comprehends the chain of causation, comprehends the teaching; anyone who comprehends the teaching, comprehends the chain of causation'. Majjhima Nikaya (1.3.306, 1998a)

Vipassana

Vipassana is the development of insight into the reality of mind and body. Though the description of the principles and practice of Vipassana are found throughout the Pali[1] literature, the Buddha gave the most comprehensive and succinct account of the subject in the *Satipatthana Sutta* (Mahasatipatthana Sutta, 1998) (literally the discourse on the establishment of mindfulness) wherein he gave the assurance of effective results upon proper practice. Vipassana and *Satipatthana* are synonymous. Anyone who gives this form of mental training a trial will soon discover that despite its

[1]All references to Pali texts are from Chattha Sangayana edition of Pali Tipitaka published by Vipassana Research Institute (www.tipitaka.org, www.dhamma.org).

apparent simplicity, it is an arduous undertaking requiring prolonged application.

Vipassana practice is generally preceded by training in morality (*sila*) and concentration (*samadhi*). All meditators are required to refrain from unlawful means of livelihood (which involves business in arms etc.), killing, stealing, sexual misconduct and untruthfulness. The method proposed by Shakyamuni Buddha is actually a three-tier system, wherein morality serves as a base for concentration which in turn supports Vipassana. Vipassana practitioners are forewarned that without a proper grounding in morality all subsequent practices would prove futile.

Buddha gave about 30 meditative subjects for concentrative practice, the development of which has been discussed in great detail in classical texts such as the *Vishuddhimagga* (Nanamoli, 1975). All forms of 'right' (*samma*) concentration involve prolonged undivided attention being given to the meditative object, attended with wholesome mental concomitants. The subjects range from visual objects (color *kasinas*) to topics requiring sustained discursive thought (death, repulsiveness with regard to the body etc.) or simply natural physiological phenomena such as the breath. Meditations are also given to generate wholesome emotional states of goodwill, compassion, sympathetic joy, equanimity and inhibit their unwholesome contraries. Although some meditative topics are suitable for all, at times Buddha would select a specific set of subjects depending on the mental constitution of the student. Again, the maximum degree of mental absorption possible, differs from one subject to another. The function of concentration here is to provide the necessary mental tranquility and sustained focus for successful Vipassana practice.

As mentioned earlier, a certain level of proficiency in morality and concentration is a prerequisite for Vipassana. Although concentration serves as a basis for Vipassana, it should not be viewed merely as another form of concentrative practice. Essentially, every form of Vipassana involves the sustained witness-like observation of natural phenomena within the framework of the body. Literally, the word Vipassana means to 'see' in a special way, that is to see with awareness and equanimity. In 'concentration' an effort is made to sustain attention on a preconceived subject. In Vipassana, all such 'effort' is wholly abandoned and the emphasis is on the equanimous observation of natural phenomena, 'as it is', whatever it may be, without any imaginative, conceptual or discursive mental activity on the part of the meditator. Such sustained observation over prolonged periods of time does lead to insight or penetration into the true nature of phenomena pertaining to the body and mind. As an eminent teacher and authority on Vipassana meditation Shri S.N. Goenka puts it

> 'The practice is to explore within oneself, and to experience the reality of one's own physical and mental structure. This is what Siddhartha Gautama did to become a Buddha. Leaving aside all preconceptions, he examined himself to discover the true nature of the physical and mental structure. Starting from the level of superficial, apparent reality, he penetrated to the subtlest level, and he found that the entire physical structure, the entire material world, is composed of subatomic particles, called in Pali language *attha kalapa*. … These particles, he found, are the basic building blocks of matter, and they are themselves constantly arising and passing away, with great rapidity — trillions of times within a second. In reality there is no solidity in the material world; it is nothing but combustion and vibrations'. Goenka (2004)

Although Vipassana is designed to probe every facet of the body–mind complex, initially one aspect is made the focus of attention. Formally, Buddha identified four bases as focal points to initiate Vipassana practice

- Awareness of body.
- Awareness of sensations (feelings).
- Awareness of mind or consciousness.
- Awareness of mental contents.

Awareness of body

Awareness of breath (Anapana Sati)

Within the framework of the body which in this context probably serves as a laboratory, the awareness of respiration is the first step in the initiation of Vipassana. One starts the training with simple awareness of incoming and outgoing-breaths. Only, the natural breath is made the focus of attention and no attempt is made to regulate the breath in any manner as in some other spiritual practices. As one progresses, the attention is gradually extended to include all physical and mental states. Respiration is a natural phenomenon which can serve as an object both for concentration (samadhi) and Vipassana.

This process of awareness of respiration has an interesting history. Buddha had tried various meditative methods in vogue during that time. He mastered various concentration techniques but was still left dissatisfied. He then turned to self-mortification but the most severe austerities also proved to be futile. It was at this point that he remembered an incident in his childhood when he was in complete solitude. At that time, he experienced a very peaceful and joyous state as he had watched his breath coming in and going out. He remembered that state as being bereft of any sense desire or unwholesome thoughts. It was this memory of peaceful absorption under a rose-apple tree that made him give up self-mortification and embark on a series of objective observations within his own mind and body culminating in the complete elucidation of the Vipassana process.

When one sits quietly in solitude, among the first bodily function that attracts attention is the breath, and this is where one begins.

> If the breath is deep, it is deep. If it is shallow, it is shallow. Don't interfere with the natural breath. Just observe. Do nothing. It is mere observation, bare observation. One does nothing. It is like a person sitting on the bank of a river and looking at the flow of the river without interfering with it in any way. In Indian languages today it is called tatastha; literally, sitting on the bank of the river. Goenka (2004)

Though the Buddha gave many objects of concentration, his object of choice was respiration as it is intimately related to both physical and mental states. Such a facility does not exist with other objects. On the other hand, even as one starts the practice of awareness of respiration, one simultaneously and automatically begins to monitor concomitant mental and physical states. However, the importance of all the four bases of mindfulness was also stressed, demonstrating the importance he gave to understanding the totality of the mind and body phenomenon at the experiential level.

Awareness of sensations

Contact of sense objects (sight, sound, smell, taste, touch, mental objects) with their respective sense doors (eyes, ear, nose, tongue, body and mind) produces sensations within the body. Normally unskilled reaction to these sensations produces *craving* (the Pali word *tanha* includes both craving and aversion) which in turn leads to addictive attachment. It was Buddha's insight that one does not crave directly for the external object per se, but for the secondary sensations the perception of the object induces in one. For a Vipassana practitioner this is a crucial link in the chain of causation. If one maintains proper awareness of sensations (that is awareness with equanimity), the habit of *craving* can be largely attenuated. Sensations arise both due to physical and mental causes and thus can be used to track both material and mental phenomena.

> Having experienced as they really are, the arising of sensations, their passing away, the relishing of them, the danger in them, the release from them, the Buddha is fully liberated, free from all attachments. Digha Nikaya (1.1.36, 1998)

The practice of *Anapana Sati*, the awareness of breath, prepares the mind for the more advanced task regarding the awareness of sensations. One

observes the breath in the narrow area below the nostrils and above the upper lip, the practice of which endows the mind with precision, focus and subtlety, necessary for the next stage which is to observe sensations with equanimity, throughout the body. In practice the attention is moved in a stylized manner from feet to head, making a choiceless observation of all sensations one encounters in different parts of the body during the passage. Thus the whole process is aptly described as

> 'And you proceed further … you find that there are sensations throughout the body. Something is happening throughout the body. Some biochemical reaction or some electro-magnetic reaction. Whatever the sensation may be — heat, cold, dryness, heaviness, numbness, pain, pressure, tension, throbbing, pulsing, vibrating, itching, tickling or something else that you cannot name or label; whatever it may be — just observe the reality as it is. Vipassana helps you to go deeper and deeper within. You start feeling sensations throughout the body. You reach a stage where you find that the entire structure is nothing but vibrations. Mind and body — nothing but vibrations. Arising, passing, arising, passing… See that you don't react to any sensation, gross or subtle. Always maintain equanimity. Goenka (2004)

And again

> 'The word passana means "seeing", the ordinary sort of vision… Vipassana means a special kind of vision: observation of the reality within oneself. This is achieved by taking as the object of attention one's own physical sensations. The technique is the systematic and dispassionate observation of sensations within oneself'. Hart (1999)

In Buddha's view, lack of awareness of internal sensations throughout the body constitutes ignorance. The subconscious craving for pleasant sensations and aversion for unpleasant ones is the root cause of mental unrest and addictions. As the systematic observation of sensations is continued, the awareness proceeds from gross to subtle sensations and to their relationship with physical and mental states. A Vipassana meditator understands that "a pleasant (or an unpleasant or a neutral) sensation has arisen in me". It is compounded, gross and dependent upon causal conditions. Equanimity is maintained as a result of direct experience of the impermanent nature of sensations, leading to the weakening and eventual eradication of the conditioning of craving, aversion and ignorance. Thus

> When his (or her) underlying conditionings of craving for pleasant sensation, of aversion toward unpleasant sensation, and of ignorance toward neutral sensation are eradicated, the meditator is totally free of underlying conditionings, has seen the truth, has realized the illusory nature of ego and has made an end of suffering. Samyutta Nikaya (2.4.251, 1998a)

The Buddha maintained that *Anapana Sati* (awareness of breath) was not complete without the awareness of sensations. Observation of sensations naturally follows the awareness of breath. Similarly Vipassana with the proper awareness of sensations naturally involves awareness of mind and mental contents. Actually all the four sections of awareness are interrelated and overlapping. The practice of one naturally leads to the practice of all four (Majjhima Nikaya 3.2.144–152, 1998b).

Awareness of mind (*Citta*) and mental contents (*Cetasikas*)

Everything that arises in the mind is accompanied by sensation (Anguttara Nikaya 3.8.83, 3.10.58, 1998b).

The awareness of the stream of consciousness and mental contents (concomitant mental factors)

requires diligent training. Sensations help us by giving a continual hold on the reality of the mind and body. One elementary scheme classifies mental states to be associated with craving, aversion or delusion. The sensations experienced in the body due to the three are also phenomenally different. Since one conscious state rapidly succeeds another a considerable degree of awareness and equanimity has to be maintained to observe the fluctuating sensations associated with each state. Fairly elaborate schemes classify different types of mental states to aid the observation. A few such schemes are listed below

Awareness of Hindrances: Understanding that "at this moment there is present (or absent) in me sense desire or aversion or sloth and torpor or agitation or doubt."
Awareness of Aggregates: Awareness of the five aggregates — matter (body), sensation, perception, conditioning (all three included in mental contents) and consciousness (mind).
Awareness of Sense Spheres: The Buddha included mind here along with eye, ear, tongue, nose and body. For the sense sphere of mind, mental contents are the sense objects.
Awareness of Factors of Enlightenment: Awareness training as to whether these factors are present or not. These factors are awareness, investigating faculty, effort, rapture, tranquility, concentration and equanimity.
Understanding of the Four Noble Truths: Detailed analysis of the four noble truths — suffering, origin of suffering, cessation of suffering and way leading to the cessation of suffering. It should be understood that use of any of the schemes does not involve either conceptual or discursive thinking. Direct access to and observation of the succession of conscious states without the aid of sensations is also possible, but is rarely achieved, involving a high level of proficiency and skill.

Conclusion

In concluding the *Satipatthana Sutta* Buddha assured that all those who would properly practice the fourfold awareness would definitely find lasting contentment and peace. He consistently maintained that the problem of human happiness was his primary concern and a mere intellectual account of the world held little interest for him. In his own words, what he taught was a minuscule fraction of what he really knew.

> Which is more: the few leaves that I have picked up in my hand or the leaves on the trees in the wood?… The things that I have known by direct knowledge are more: the things that I have told you are only a few. Why have I not told them? Because they are of no benefit in eradication of suffering, in attaining happiness… Samyutta Nikaya (3.5.1101, 1998b)

An elaborate model of the conscious mind and its causal relation to physical states, based upon cross-validated experiences arising out of the application of the Vipassana technique, was developed by Shakyamuni. The details of the model were sketched out in a series of seven treatises referred to as the Abhidhamma (Bhikkhu Bodhi, 1999) in the Pali Canon (Tipitaka). In this chapter only the practical aspects of Vipassana have been described. It is our belief that the Abhidhamma insights into the nature of the conscious mind have the potential to enrich modern thought related to the philosophy and psychology of mind.

References

Anguttara Nikaya 1.3.66 Kesamuttisutta. (1998a) Tipitaka, Vol. 35, Vipassana Research Institute, Igatpuri.
Anguttara Nikaya 3.8.83, 3.10.58 Mulaka Sutta. (1998b) Tipitaka, Vol. 40, Vipassana Research Institute, Igatpuri.
Bhikkhu Bodhi. (1999) Comprehensive Manual of Abhidhamma. Buddhist Publication Society, Sri Lanka.
Dhammapada 276. (1998) Tipitaka, Vol. 47, Vipassana Research Institute, Igatpuri.
Digha Nikaya 1.1.36 Brahmajala Sutta. (1998) Tipitaka, Vol. 1, Vipassana Research Institute, Igatpuri.
Goenka, S.N. (2004) Ten Day Discourses. Audio CDs. Vipassana Research Institute, Igatpuri.
Hart, W. (1999) The Art of Living (Vipassana Meditation as taught by S.N. Goenka). Vipassana Research Institute, Igatpuri.

Mahasatipatthana Sutta being English translation of Satipatthana Sutta. (1998). Vipassana Research Institute, Igatpuri.

Majjhima Nikaya 1.3.306 Mahahatthipadopama Sutta. (1998a) Tipitaka, Vol. 12, Vipassana Research Institute, Igatpuri.

Majjhima Nikaya 3.2.144–152. Anapanasati Sutta. (1998b) Tipitaka, Vol. 14, Vipassana Research Institute, Igatpuri.

Nanamoli, (1975) Visuddhimagga by Buddhaghosha. Buddhist Publication Society, Sri Lanka.

Nyanaponika Thera, (1998) Abhidhamma Studies. Wisdom Publications, Boston, MA.

Samyutta Nikaya 2.4.251 Pahana Sutta. (1998a) Tipitaka, Vol. 26, Vipassana Research Institute, Igatpuri.

Samyutta Nikaya 3.5.1101 Sisapavana Sutta. (1998b) Tipitaka, Vol. 28, Vipassana Research Institute, Igatpuri.

This page intentionally left blank

R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 20

Buddha and the bridging relations

Rahul Banerjee*

Saha Institute of Nuclear Physics, Sector 1, Block AF, Bidhan Nagar, Calcutta 700064, India

Abstract: The chapter reviews a classical Indian model of consciousness found in the Abhidhamma, a collection of seven treatises in the Pali Canon Tipitaka. The model was based on observations made during advanced vipassana practice, a first-person method taught by the Buddha. The climax of the model consists in the elucidation of 24 'Bridging Relations' causally linking the stream of consciousness, its contents and associated physical events. Review of such a model based on a specialized first-person technique could prove to be a resource of useful ideas providing directions for further research.

Keywords: first-person methods; hard problem; bridging relations; Abhidhamma; vipassana

Introduction

The last decade has seen the reemergence of consciousness as a subject suitable for scientific study. The biological basis of consciousness and how it arises from the matrix of molecular events in the brain has been identified as an important unresolved scientific question (Miller, 2005). Although identifying the neural correlates of consciousness (NCC) forms an indispensable part of the experimental paradigm to solve this problem, there is also a growing consensus that simply to study the neural correlates of conscious experiences in a variety of modes will not be wholly sufficient. Rather a complete theory of consciousness is necessary (Baars, 2003), which will demonstrate in trenchant terms the status of conscious experience in the natural order (Crick and Koch, 1998).

Relating the inner qualitative subjective 'feel' of conscious experience to either cognitive processing or molecular events probably forms the most difficult aspect of the problem. Designated as the 'Hard Problem' (Chalmers, 1996), the question has generated a great deal of intellectual activity in the philosophic and scientific literature. One view, put forward by David J. Chalmers, argues in favor of the irreducibility of the phenomenal aspects of consciousness to any other physical principle. It thus follows that consciousness is a fundamental reality in its own right, and a further set of psychophysical laws or 'bridging relations' will have to be found, in order to resolve the issue.

The present paper reviews a set of 24 conditional (bridging) relations found in a classical model of consciousness attributed to the Buddha (Narada, 1997). The model in its most elaborate form can be found in the Abhidhamma, a collection of seven treatises contained in the Pali Canon Tipitaka. The model is based on a series of insights experienced through the practice of vipassana, a first-person method taught by the Buddha in order to probe the dynamics of one's own stream of

*Corresponding author. Tel.: +91-33-2337-5345;
Fax: +91-33-2337-4637; E-mail: rahul.banerjee@saha.ac.in

DOI: 10.1016/S0079-6123(07)68020-0

consciousness, its contents and associated material events (Nanamoli, 1975). Buddha consistently maintained that the insights gained by him were accessible to all. Thus the Abhidhamma gives a rational form to a series of cross-validated observations based on a specialized first-person technique.

The picture of mind and matter which emerges in the Abhidhamma is wholly interactionist. Consciousness, its contents and associated material events are irreducible fundamental realities, yet causally conditioned by each other in a complex and intricate interaction. The 24 conditional relations attempt to describe the modes by which this mutual conditioning and interaction occurs. In a world which still searches for a comprehensive natural and scientific theory of consciousness, the relations might provide illuminating guidelines for further studies. A few of the principles central to the Abhidhamma are reviewed prior to the discussion of the relations.

The stream of consciousness and its contents

The central concept in the Abhidhamma (Nārada, 1979) is that of the chitta. Generally it can be defined as a transient pulse of 'experience'. The practice of vipassana meditation enables the meditator to resolve mental events which are closely spaced in time. In advanced vipassana practice it becomes evident that there is a certain 'graininess' involved in our stream of consciousness, which is actually composed of a succession of rapid, discrete and transient pulses. Each such individual pulse is called a chitta. The temporal duration for which a chitta exists is indeed very short and the classical texts remark that a 'million chittas rise and fall with the blinking of an eye'.

To resolve our stream of consciousness into a succession of chittas is possible only in a highly advanced stage of vipassana. Yet the terms 'mind' or 'consciousness' in the Abhidhamma actually refers to this stream. So fundamental is the notion of chitta in the Abhidhamma, that the duration for which a chitta exists, called mind moment, serves as a natural temporal unit for mental events. Thus three mind moments implies the passage of three chittas. In an individual stream of consciousness there is no superposition of chittas and a unique chitta is in existence at any given moment. Chittas are therefore inherently dynamic with each transient pulse being replaced by its successor in a continuous series.

The contents of (a unit pulse of) consciousness or the contents of a chitta are termed 'concomitant mental factors' or cetasiks (Gorkom, 1997). Every chitta will invariably be accompanied by a set of cetasiks, which also serve to identify the chitta. Both the chitta and its cetasiks are in absolute synchrony, rising and falling simultaneously. There are seven cetasiks which are universally associated with every chitta: (1) contact (phassa); (2) feeling or sensation (vedana); (3) volition (cetana); (4) psychic life energy (jivitindriya); (5) perception (sanna); (6) attention (manasikara); and (7) one-pointedness (ekaggata).

'Contact' refers to the fact that an object has impinged or has come into contact with anyone of the sense bases and is now being experienced by the chitta. Perception in this context is a cognitive function which has to do with taking 'note' of the perceptual cues pertaining to the object and also recognizing what has been previously noted. The Atthasalini (Davids, 1999) defines it as 'The noting of an object as blue-green, etc., is perception. It has the characteristic of noting and the function of recognizing what has been previously noted. There is no such thing as perception in the four planes of existence without the characteristic of noting.' Every chitta will have a corresponding object and attention, one-pointedness refer to the exclusive focus of the chitta on the object of its experience. Volition directs and coordinates the cognitive functions of all other cetasiks with respect to the object. Although psychic energy provides the power to sustain the surge of the chitta and its cetasiks, it is volition which directs the energy in addition to coordinating the cognitive function of other associated cetasiks. In defining volition the Atthasalini (Davids, 1999) sums it up as 'There is no such thing as volition in the four planes of existence without the characteristic of coordinating, all volition has it …… It has directing as manifestation. It arises directing associated states

(read cetasiks) ……. as the general, fighting himself, makes other soldiers take part in the battle ….. even so, when volition starts work on its object, it sets associated states to do each its own work.' Volition and psychic energy can perhaps be fruitfully combined to give volitional energy, the character of which is unique for every chitta.

Apart from the seven universal cetasiks, additional cetasiks referred to as 'sankharas' distinguish one chitta from another. In fact apart from perception and feeling all other concomitant mental factors are collectively grouped together as sankhara. Sankharas are habitual mental formations (activities), which give a specific character to all the universal cetasiks of the chitta. For example, hatred and goodwill are two sankharas which form the contents of two distinct chittas respectively

1. chitta, unprompted, accompanied by displeasure, and connected with ill will;
2. chitta, unprompted, accompanied by pleasure, associated with knowledge.

Here 'unprompted' refers to the fact that the chitta arises spontaneously and is not instigated or 'prompted' by some external agency. The set of cetasiks which can be found in synchronous association with a particular chitta is fixed. As has been stated previously, although volitional energy is universally present with all chittas, its specific character is conditioned by the constellation of sankharas in simultaneous existence with the chitta. Jealousy, hatred, sloth, goodwill and tranquility are sankharas associated with their characteristic volitional energies and thus specific type of chitta. Actually, the chitta and its cetasiks are mutually supportive and reciprocally condition each other.

According to one classification scheme there are 89 distinct chittas. Although it is beyond the scope of this essay to describe the status of each individual chitta, they are divided into four categories of which two are skillful (kusala) and unskillful (akusala). The akusala chittas will be invariably 'rooted' in the sankharas of delusion, aversion and greed, though aversion and greed are not simultaneously instantiated in the same chitta. In other words, distinctly different akusala chittas support aversion and greed as cetasiks. Likewise kusala or moral chittas are rooted in knowledge, detachment and goodwill.

Here, 'delusion' refers to the inability to penetrate and determine the true nature of the object, which is effectively done by 'knowledge' or 'wisdom' (panna).

The full set of sankharas for the skillful and unskillful classes can be expanded to: (1) sincerity; (2) mindfulness; (3) moral shame; (4) moral dread; (5) detachment; (6) goodwill; (7) equanimity; (8, 9) calmness; (10, 11) uprightness; (12, 13) proficiency; (14, 15) pliancy; (16, 17) wieldiness; (18, 19) lightness *and* (1) delusion; (2) moral shamelessness; (3) moral fearlessness; (4) restlessness; (5) greed; (6) self-conceit; (7) false view; (8) aversion/hatred; (9) jealousy; (10) avarice; (11) anxiety; (12) sloth; (13) torpor; (14) doubt, respectively.

Calmness along with a few other cetasiks (uprightness, proficiency etc.) are counted twice, as calmness pervades both the chitta and its associated cetasiks. In addition sympathetic joy, compassion and wisdom can also be found in association with kusala chittas. The first 19 kusala sankharas can be found in all the skillful chittas. There are a few other sankharas which can be found in one or both classes.

Resultants, life continuum and pathways

During the course of advanced vipassana practice it was observed that some chittas appeared as a consequence or causal effect of previously instantiated chittas. These were termed resultants (vipāka) and put in a separate category in addition to kusala and akusala chittas. Essentially skillful and unskillful chittas stand in a cause–effect relationship with respect to their resultants. A chitta and its resultant need not be temporally contiguous and may appear after a considerable lapse of time when conditions are favorable for their instantiation. For some particular cases however (lokuttara chittas) the chitta and its resultant follow each other immediately. For example for one chitta rooted in aversion

1. chitta, unprompted, accompanied by displeasure, and connected with ill will, could have as one of its resultants;
2. body consciousness associated with pain.

The volitional energy of a resultant can be considered to be passive in contrast to the active volition of its parent chitta.

One noteworthy feature of the Abhidhamma model is the complete absence of unconscious mental states. Under normal circumstances the stream of chittas cannot be arrested and there is no break in their rapid succession. Even in the case of deep dreamless sleep the flow continues. It thus becomes imperative to characterize the stream in both states to account for their cognitive and phenomenological differences. Actually, most often the stream is constituted of vipāka chittas which flow without any definite order or sequence. This random flow of vipāka chittas is termed bhavanga (life continuum), which gives continuity to our psychic existence. Upon contact of an object with any of the sense bases this irregular succession of vipāka chittas gets checked and bhavanga gets arrested. A train of chittas is then instantiated in which specific chittas follow each other in a regular and definite temporal sequence. This succession of specific chittas constitutes a pathway. On the completion of a pathway the stream relapses back into bhavanga. Thus the stream moves back and forth between bhavanga and different pathways contingent upon the objects to be experienced. In our waking state two pathways are separated by bhavanga whereas the stream is dominated by bhavanga in sleep.

Most often, the pathway instantiated is in response to an object which impinges with sufficient strength upon one of the sense bases. An example of such a pathway in response to a visual object making contact with a visual sense base is given below (the terms being explained after the pathway).

Arrest of Bhavanga	
↓	
Sense door adverting consciousness (Pañcadvārāvajjana chitta)	(1 mind moment)
↓	
Eye consciousness (Cakkhu viññāna chitta)	(1)
↓	
Receiving consciousness (Sampaticchana chitta)	(1)
↓	
Investigating consciousness (Santirana chitta)	(1)
↓	
Determining consciousness (Votthapana chitta)	(1)
↓	
Javana	(7)
↓	
.	
.	
↓	
Javana	
↓	
Registering consciousness (Tadārammana chitta)	(2)
↓	
Registering consciousness	
↓	
Relapse into Bhavanga	

For all the five sense bases the basic structure of the given pathway is conserved. As has been previously stated every pathway involves a regular succession of specific chittas. The volitional energy of each chitta in the temporal order performs (and directs associated cetasiks to perform) a definite cognitive function which enables the recognition and representation of the object in experience. Occasionally the chitta is named after the function performed by its volitional energy. On the arrest of bhavanga (see pathway given above) a momentary chitta (sense door adverting consciousness) orients the stream to the sense base which has come into contact with the sense object. This is followed by a chitta which is specific to the sense base, in this case eye consciousness. Subsequent to the passing away of eye consciousness, arise two chittas in succession whose volitional energies perform the functions of 'receiving' and 'investigating', followed by determining consciousness, by which time some definite features of the object have been determined.

The first part of the pathway from 'sense door adverting ' to 'determining' can be said to be passive wherein the nature of the object is sought to be 'investigated' and 'determined'. Eye consciousness, receiving and investigating consciousnesses are all vipāka chittas. 'Receiving' and 'investigating' are functions which can be performed by only two and three specific chittas respectively. Apart from kusala, akusala and vipāka chittas there is yet

another class which are neither resultants nor do they give rise to resultants as effects. Sense door adverting and mind door adverting chitta (which performs the function of determining) consciousnesses belong to this fourth category called kiriya.

Javana refers to the reactive stage of the pathway in which either kusala or akusala chittas instantiate and repeat themselves seven times in succession. The skillful and unskillful chittas in javana are essentially the reaction of the stream in response to the object. Vipāka chittas are excluded from appearing in javana. The pathway finally terminates with two mind moments of 'registering' which leaves a memory trace, 11 distinct vipāka chittas being capable of performing the said function.

Several different pathways exist which get triggered in different cognitive situations. In order to have a full-blown experience of an object complete with all its perceptual details selected pathways in correct sequence have to repeat several times. The process is cumulative. That is initially the stream experiences partial representations of the object which gradually mature to synthesize the entire object due to the repeated iterations of specific pathways in the correct sequential order (Davids, 2003).

Conditional relations

The Abhidhamma considers phenomena which are conditioned and causally bound to be constituted of chittas, cetasiks and matter or material events referred to as rupa. The 24 conditional relations describe the manner in which the three elements causally condition each other. To start with a clear distinction is made between material modifications which results from energy transactions exclusively within the material domain, independent of mind (i.e., to say chittas and cetasiks) and material phenomena conditioned by mind. Thus only a subset of material events are causally closed with respect to mind and the same principle does not apply to the totality of phenomena involving chitta, cetasiks and rupa. Given the limited scope of this essay the 24 conditional relations will not be dealt with exhaustively, but only some salient features of a few relations will be discussed to give an overview of the attempt to model the mind and causally link them to material events.

A conditional relation is a link between two states, a conditioning state (paccaya dhamma) and a conditioned state (paccayuppanna dhamma). A chitta, its constellation of cetasiks or a flux of material events could qualify as a state. A state could be determined by several conditioning states which is to say several conditional relations could be in simultaneous operation as determinants of a state. The complete enumerations of the 24 relations are: (1) root; (2) object; (3) dominance; (4) contiguity; (5) immediate contiguity; (6) conascence; (7) reciprocity; (8) prenascence; (9) postnascence; (10) repetition; (11) dependence; (12) strong dependence; (13) kamma; (14) vipāka; (15) nutriment; (16) faculty; (17) jhana; (18) path; (19) association; (20) dissociation; (21) presence; (22) absence; (23) disappearance; (24) non-disappearance. The same relation can be used in different contexts to relate a wide range of conditioning and conditioned states. In the following discussion the 24 relations will be factored to relate: (1) chitta to chitta; (2) chitta to cetasiks; and (3) chitta, cetasiks to rupa. 'Chitta to chitta' refers to the fact that in this case chittas are both conditioning and conditioned states. For categories 2, 3 the reverse relation is also implied, that is, cetasiks conditioning chitta and rupa conditioning both chitta, cetasiks.

Chitta to chitta

In bhavanga there is no short- or long-range correlation between chittas as it is a random flow of resultants. A pathway on the other hand has a well-defined temporal sequence. Within the context of a pathway the preceding chitta exercises control or conditions the subsequent chitta which is temporally contiguous to it (*contiguity*, *immediate contiguity*). Since the flow is linear there can be only one chitta at a time. The chitta which has just collapsed strongly conditions the character of its immediate successor (*absence, disappearance*). Thus the principles of *contiguity* and *absence* (refers to the chitta which has just collapsed and therefore absent) determines the ordered sequence

(adverting, eye consciousness, receiving, investigating, etc.) of a pathway.

Repetition refers to the propensity of a kusala or akusala chitta to repeat itself seven times in succession during javana. It is not mandatory for a chitta to go through the full seven cycles. If the impact of the sense object to the sense base is not sufficiently strong both the javana and consequently the pathway can terminate with a number of repetitions less than seven.

Chitta to cetasiks

A chitta is invariably accompanied by its constellation of cetasiks (*association*), with which it rises and falls in absolute synchrony (*conascence*). No element constituting the chitta–cetasik complex can stand in isolation and each element mutually supports and conditions each other (*reciprocity*, *dependence*). The analogy given is that of a bunch of sticks propping each other up in order to stay erect. Although there is reciprocal dependence, yet one sankhara can play a key role in terms of conditioning the rest (*root*). Thus akusala and kusala chittas are said to be rooted in delusion, greed, aversion or wisdom, detachment and goodwill respectively. In addition there could also arise a situation where a sankhara exercises a dominant causal influence over its associated conascent partners (*dominance*). *Dominance*, probably lies at the basis of meditative (*Jhana*) and vipassana (*Path*) practice whereby a selected set of sankharas are trained to exert a preponderant causal influence.

Kusala and akusala chittas will infallibly give rise to resultants (*vipāka*). The Abhidhamma holds volition to be the connecting link between a chitta and its resultant (*kamma*). One meaning of the word kamma denotes volition. Kamma is a multisignificant term yet the most profound meaning attached to it is the arising of vipāka chittas as a consequence of the volitions of previously instantiated kusala or akusala chittas.

Chitta, cetasik to rupa

Before describing the modes by which matter (rupa) and mind (chitta and cetasiks) condition each other, it would be proper to briefly review the Abhidhamma analysis of matter. Intense application of vipassana reduces the stream of consciousness into discrete chittas. Likewise the body of the meditator is resolved into rapid, transient and ephemeral events called kalapas. Thus by virtue of the first-person technique the body is seen as an aggregate of kalapas. The temporal duration for which a kalapa exists is postulated to be longer than that of a chitta, though within comparable orders of magnitude. With prolonged vipassana practice four types of kalapas arise in the immediate awareness of the meditator: (1) those responsible for inertia, heaviness and extension; (2) temperature fluctuations of heat and cold; (3) oscillation; and (4) cohesion. In any material event within the framework of the body all four types of kalapas will be simultaneously present, even though one may dominate, with the expression of the other three held in abeyance. The condition of *conascent reciprocity* used to relate the chitta to its cetasiks, also applies to the mutual conditioning of the four kalapas amongst themselves. All other higher forms of material organization termed 'derived matter' are thus due to kalapic aggregations. Of all the forms of derived matter two deserve special mention: (1) the five sense bases (pasada rupas), most probably referring to the sensitive surface of the physical organ which finally receives the impact of the sense object and a vital life energy referred to as rupa jivitindriya possibly the physical equivalent of the cetasik psychic life energy which is a concomitant mental factor.

According to the Abhidhamma mind and matter causally condition each other in three principal modes. Every chitta will invariably be focused on some object (*object*). The sense base which receives the impact of the sense object exists prior to (*prenascence*) and conditions the subsequent pathway selected by the stream of chittas (*dependence*, *presence*, *non-disappearance*). Trivially, the physical object must also exist prior to its contact with the sense base (*object prenascence*). The principle of *dominance* also finds use in this context to describe the dominant causal influence exerted by a physical object on mental events. The difference between the relations of *contiguity* and *dependence* is that in the

former, the causal control is exercised only over the succeeding chitta whereas in the latter the conditioning extends over the whole pathway. Thus the contact of an object with the corresponding sense base leads to the arrest of bhavanga, reorientation of the stream and instantiation of a chitta (eye consciousness, ear consciousness, body consciousness ….. etc. as the case may be) which has the same sense base as its place of origin. Other chittas of the pathway will then follow initially conditioned by this primary physical event.

In the previous case, material events condition the flow of mind, whereas in the following instances mind causally conditions matter. It is an observation that the pulse of a chitta (*conascence*, *dependence*) or a surge in its volition (*kamma*) can lead to an emanation of kalapas. The relationship between the chitta and the kalapa is *conascent* but not *reciprocal* as the existence of the chitta is not dependent on the emanating kalapas. When volition is responsible for the projected kalapas, physical life energy or rupa jivitindriya will also be contained in the kalapic aggregates. Volition is thus related to two categories of vital energies one psychic and the other physical, though the conditioning relations with respect to both are different. The conditional relations of *root*, *jhana* and *path* not only relate the chitta to its concomitant mental factors but also to the flux of kalapas which the flow of chittas or the fluctuation of its volitions generate. There can even be a situation where the chitta–cetana (volition) stabilizes a material aggregate in prior existence (*postnascence*).

Conclusion

The Abhidhamma attempts to integrate the phenomenal and causal aspects of consciousness in addition to linking them to the flux of associated material events. The phenomenal aspect of consciousness is contained in the concept of chitta whereas its causal role (in terms of cognitive functions) is played out by its configuration of cetasiks, primarily volitional energy. The fact that the phenomenal and causal aspects of consciousness cannot be separated remains one of its central insights. The model takes an unabashed interactionist stand denying the causal closure of matter as a uniform and general principle. This stand is taken on the basis of a series of observations based on a specialized first-person technique. Thus it admits two modes of material changes, one without the intervention of mind and second, chitta or volition-related material mutations. The fact that a surge in volitional energy can lead to an emanation of kalapas stands perhaps as its second most important insight. The whole issue probably boils down to the question as to what a kalapa really is, or in other words, in rigorous scientific terms what is it that is experienced as a kalapa. Until this question is fully sorted out the scientific applications of the model will remain hampered. It is also important to note that the meaning of the terms 'volition' or 'perception' differs substantially from its current usage in the western psychology. For example in western terms 'volition' presupposes a conscious mental state, subsequent to preconscious involuntary processes. In contrast, every mental state is a conscious state in the Abhidhamma system, with its associated volitional energy. Thus for any trans-cultural studies on consciousness with its differences in approach, terminology and methods, a committed effort will possibly have to be made to bridge the cultural gaps. Even though the Abhidhamma model probably does not resolve the question concerning the genesis of qualia, the logical structure of the 24 conditional relations could provide directions for further studies.

Acknowledgments

The author wishes to thank Dr. Dhananjay Chavan, Prof. Max Velmans, Prof. Thomas Metzinger for constructive suggestions to improve the quality and contents of the manuscript.

References

Baars, B.J. (2003) Introduction: treating consciousness as a variable: the fading taboo. In: Baars B.J., Banks W.P. and Newman J.B. (Eds.), Essential Sources in the Scientific Study of Consciousness. The MIT Press, MA, pp. 1–9.

Chalmers, D.J. (1996) The Conscious Mind in Search of a Fundamental Theory. Oxford University Press, Oxford.

Crick, F. and Koch, C. (1998) Consciousness and neuroscience. Cereb. Cortex, 8: 97–107.

Davids, R. (1999) The Expositor (Atthasalini) Vols. I, II (Buddhaghosha's Commentary on the Dhammasangani the First Book of the Abhidhamma Pitaka). Pali Text Society, Oxford, pp. 145–150.

Davids, R. (2003) Compendium of Philosophy (Being a Translation from the Original Pali of the Abhidhammattha Sangaha). Kessinger Publishing, Whitefish, MT.

Gorkom, N.V. (1997) Abhidhamma in Daily Life. Triple Gem Press, London.

Miller, G. (2005) What is the biological basis of consciousness. Science, 309: 79.

Nanamoli, (1975) Visuddhimagga by Buddhaghosha. Buddhist Publication Society, Sri Lanka.

Nārada, (1979) A Manual of Abhidhamma (Being Abhidhammattha Sangaha of Bhadanta Anurudhhācariya). Buddhist Missionary Society, Malaysia.

Narada, U. (1997) Conditional Relations (Paṭṭhāna), Vol. 1, The Pali Text Society, Oxford.

Subject Index