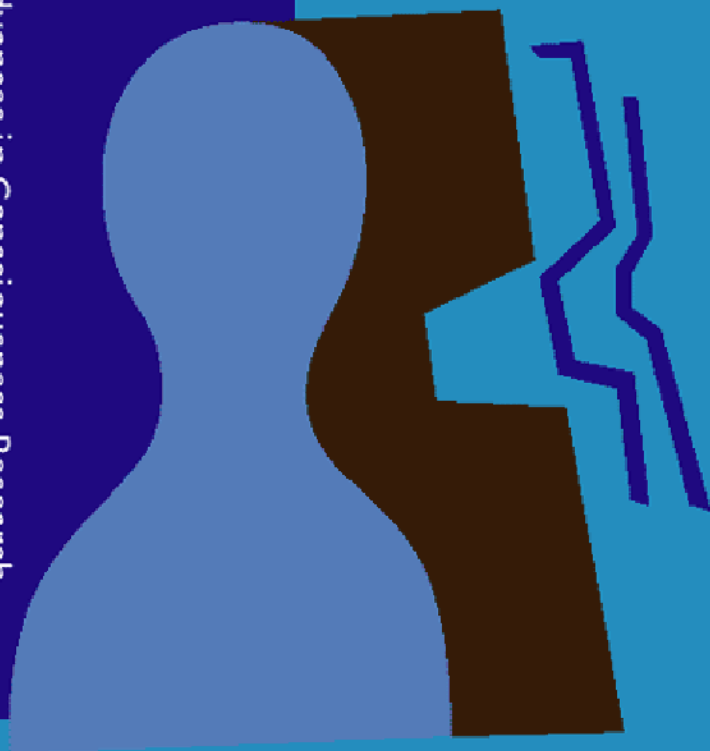


# Consciousness Evolving

Edited by  
James H. Fetzer

Advances in Consciousness Research



# Consciousness Evolving

# Advances in Consciousness Research

Advances in Consciousness Research provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series will include (but not be limited to) the various areas of cognitive science, including cognitive psychology, linguistics, brain science and philosophy. The orientation of the Series is toward developing new interdisciplinary and integrative approaches for the investigation, description and theory of consciousness, as well as the practical consequences of this research for the individual and society.

Series A: Theory and Method. Contributions to the development of theory and method in the study of consciousness.

## Editor

Maxim I. Stamenov  
*Bulgarian Academy of Sciences*

## Editorial Board

David Chalmers, *University of Arizona*  
Gordon G. Globus, *University of California at Irvine*  
Ray Jackendoff, *Brandeis University*  
Christof Koch, *California Institute of Technology*  
Stephen Kosslyn, *Harvard University*  
Earl Mac Cormac, *Duke University*  
George Mandler, *University of California at San Diego*  
John R. Searle, *University of California at Berkeley*  
Petra Stoerig, *Universität Düsseldorf*  
Francisco Varela, *C.R.E.A., Ecole Polytechnique, Paris*

## Volume 34

Consciousness Evolving  
Edited by James H. Fetzer

# Consciousness Evolving

*Edited by*

James H. Fetzer

University of Minnesota, Duluth

**John Benjamins Publishing Company**  
Amsterdam/Philadelphia



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

## Library of Congress Cataloging-in-Publication Data

Consciousness Evolving / edited by James H. Fetzer.

p. cm. (Advances in Consciousness Research, ISSN 1381-589X ; v. 34)

Includes bibliographical references and indexes.

1. Consciousness. 2. Evolution. 3. Consciousness--Physiological aspects. 4. Evolution (Biology) I. Fetzer; James H., 1940-II. Series.

B808.9.C665 2002

126--dc21

2002016363

ISBN 90 272 5154 1 (Eur.) / 1 58811 108 3 (US) (Hb; alk. paper)

© 2002 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands  
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

*To Greg Mulhauser*



# Contents

|                                                                                                                                              |            |
|----------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Contributors                                                                                                                                 | ix         |
| Introduction                                                                                                                                 | xiii       |
| <b>Prologue</b>                                                                                                                              | <b>1</b>   |
| Turing indistinguishability and the blind watchmaker<br><i>Stevan Harnad</i>                                                                 | 3          |
| <b>Part I: Natural consciousness</b>                                                                                                         | <b>19</b>  |
| Consciousness, adaptation and epiphenomenalism<br><i>Tom Polger and Owen Flanagan</i>                                                        | 21         |
| The function of consciousness<br><i>David Cole</i>                                                                                           | 43         |
| Sensations and grain processes<br><i>George Graham and Terry Horgan</i>                                                                      | 63         |
| <b>Part II: Special adaptations</b>                                                                                                          | <b>87</b>  |
| Evolution, consciousness, and the language of thought<br><i>James W. Garson</i>                                                              | 89         |
| Why did evolution engineer consciousness?<br><i>Selmer Bringsjord, Ron Noel and David Ferrucci</i>                                           | 111        |
| Nothing without mind<br><i>Stephen Clark</i>                                                                                                 | 139        |
| <b>Part III: Artificial consciousness</b>                                                                                                    | <b>161</b> |
| The emergence of grounded representations: The power and limits<br>of sensory- motor coordination<br><i>Stefano Nolfi and Oraxio Miglino</i> | 163        |
| Ago Ergo Sum<br><i>Dario Floreano</i>                                                                                                        | 181        |

|                                                                                          |     |
|------------------------------------------------------------------------------------------|-----|
| Evolving robot consciousness: The easy problems and the rest<br><i>Inman Harvey</i>      | 205 |
| <b>Epilogue</b>                                                                          | 221 |
| The future with cloning: On the possibility of serial immortality<br><i>Neil Tennant</i> | 223 |
| Subject index                                                                            | 239 |
| Name index                                                                               | 247 |

# Contributors

**Selmer Bringsjord**, Professor of Philosophy, Psychology and Cognitive Science and of Computer Science at Rensselaer Polytechnic Institute, also serves as Director of its Minds & Machines Laboratory. Author of *What robots can & can't be* (Kluwer 1992) and of *Artificial intelligence and literary creativity* (Earlbaum 1999), his new *Superminds: a defense of non-computable cognition*, is forthcoming from Kluwer.

**Stephen R. L. Clark** is Professor of Philosophy at the University of Liverpool. A former fellow of All Souls and lecturer at Glasgow University, he has delivered Gifford, Stanton, Wilde, Scott Holland, and Read-Tuckwell lectures in philosophy of religion. His most recent books include *God, religion, and reality* (1998), *The political animal: biology, ethics, and politics* (1999), and *Biology and christian ethics* (2000).

**David Cole** is Head of the Department of Philosophy at the University of Minnesota, Duluth. His publications include articles on natural meaning and intentionality, language and thought, and computers, consciousness, inverted spectra and qualia. The senior editor of *Philosophy, mind, and cognitive inquiry* (Kluwer 1990), he has recently co-authored a book on the evolution of technology.

**James H. Fetzer**, McKnight University Professor at the University of Minnesota, Duluth, is the author or editor of 22 books, including *AI: Its scope and limits* (1990), *Philosophy and cognitive science* (1991; 2nd edition, 1996), and *The philosophy of evolution* (forthcoming), and the author of more than 100 articles in the philosophy of science and on the theoretical foundations of computer science, artificial intelligence, and cognitive science. He is the founding editor of the journal, *Minds and machines*.

**Owen Flanagan**, James B. Duke Professor of Philosophy at Duke University, has written or edited seven books, including *The science of the mind* (1984; 2nd edition, 1991), *Consciousness reconsidered* (1992), and *The nature of consciousness*, edited with Ned Block and Guven Guzeldere (1998). His most recent book, *Dreaming souls: sleep, dreams, and the evolution of the conscious mind*, was published by Oxford University Press in the fall of 1999.

**Dario Floreano** is Senior Researcher at the Swiss Federal Institute of Technology in Lausanne (EPFL). He works in the fields of Evolutionary Robotics, Neural Networks, and Autonomous Agents. He has published two authored and co-authored books and two edited books on these subjects. He organized ECAL99 (The 5th European Conference on Artificial Life) and co-organized SAB2000 (The 6th International Conference on the Simulation of Adaptive Behavior).

**James W. Garson**, Professor and Chair of the Department of Philosophy at the University of Houston, received his Ph.D. from the University of Pittsburgh and has held visiting appointments in Computer Science (University of Illinois at Chicago) and Psychology (Rice University). The author of articles in logic, semantics, formal linguistics, computerized

education, and cognitive science, his recent concern has been to explore connectionist and dynamical models of cognition, with special emphasis on their implications for the nature of mental representations.

**George Graham** is Professor of Philosophy and Psychology at the University of Alabama at Birmingham. He is the author, co-author, or co-editor of six books, including *When self-consciousness breaks* (MIT Press 2000). His current research focuses on the philosophy of psychopathology and also, in collaboration with Terence Horgan, on the phenomenology of intentionality.

**Stevan Harnad**, Professor of Cognitive Science at Southampton University, is the founding editor of the journal *Behavioral and brain sciences*, *psycoloquy*, an electronic journal sponsored by the American Psychological Association, and the *Cogprints electronic preprint archive in the cognitive sciences*, and author or contributor to over 100 publications, including *Categorical perception* (Cambridge 1987), *The selection of behavior* (Cambridge 1988), and *Icon, category, symbol* (forthcoming).

**Inman Harvey** has been researching in the Evolutionary and Adaptive Systems Group at the University of Sussex for the past 11 years while pursuing a doctorate in the development of artificial evolution for design problems, including a series of studies in evolutionary robotics, where the “brain” and other aspects of the “body” of a robot are designed through methods akin to Darwinian evolution, which raises philosophical as well as scientific issues.

**Terence Horgan** is Professor of Philosophy and William Dunavant University Professor at the University of Memphis. He has published numerous articles (many collaborative) in metaphysics, philosophy of mind, philosophy of psychology, philosophy of language, metaethics, and epistemology. He is co-editor (with John Tienson) of *Connectionism and the philosophy of mind* (Kluwer 1991) and co-author (with John Tienson) of *Connectionism and the philosophy of psychology* (MIT 1996).

**Orazio Miglino** is an experimental psychologist who teaches Theories and Systems of Artificial Intelligence at the University of Naples II, Italy. His research interests are in the fields of Cognitive and Educational Psychology, in which he undertakes the construction and the validation of artificial models (computer simulations and real mobile robots) of real life phenomena (based upon artificial life and connectionist approaches).

**Ron Noel**, Assistant Professor of Psychology and Assistant Director of the Minds & Machines Laboratory at Rensselaer Polytechnic Institute, specializes in cognitive engineering and the study of cognitive, biological, and machine design systems. He has received many awards for his work, ranging from a national design award for artwork to an award by the United States Army for the creation of original electronic hardware.

**Stefano Nolfi** is currently coordinator of the Division of Neural Systems and Artificial Life of the Institute of Psychology, National Research Council, Rome. His research interests focus on neurocomputational studies of adaptive behavior in natural organisms and artificial agents, especially situated and embodied in interaction with their environments in order to understand how living organisms change, phylogenetically and ontogenetically, as they adapt to their surroundings.

**Thomas Polger** (Ph.D. 2000, Duke University), currently assistant professor at the University of Cincinnati, is a former William Bernard Peach Instructor in the Department of Philosophy at Duke University. He has contributed to *Where biology meets psychology* (MIT Press 1999) and *Dennett's philosophy: a comprehensive assessment* (MIT Press 2000).

**Neil Tennant**, Professor of Philosophy and Adjunct Professor in Cognitive Science at The Ohio State University, received his Ph.D. from Cambridge University in 1975. He is the author or coauthor of *Natural logic* (Edinburgh 1978), *Philosophy, evolution and human nature* with F. von Schilcher (Routledge and Kegan Paul 1984), *Anti-realism and logic* (Oxford 1987), *Autologic* (Edinburgh 1992), and *The taming of the true* (Oxford 1997).



# Introduction

An adequate understanding of the evolution of consciousness presupposes an adequate understanding of evolution, on the one hand, and of consciousness, on the other. The former, alas, appears to be more readily achieved than the latter. There are many accounts of evolution that have earned wide-spread acceptance, at least in general, while the nature of consciousness remains a matter of considerable dispute. Some envision consciousness as awareness, some as awareness with articulation (or the ability to describe that of which one is aware), others as self-awareness or else as self-awareness with articulation. A central problem appears to be the character of subjective experience itself with respect to its phenomenal properties (or qualia).

Evolution understood as a biological process should be characterized in terms of three principles, namely: that more members are born live into each species than survive to reproduce; that crucial properties of offspring are inherited from their parents; and that several forms of competition between the members of a species contribute to determining which of them succeeds in reproducing. The mechanisms that produce genetic variation, moreover, include genetic mutation, sexual reproduction, genetic drift, and genetic engineering, while those that determine which members tend to survive and reproduce include natural selection, sexual selection, artificial selection, and group selection, which is a process that some theoreticians deny.

Since genetic engineering and artificial selection involve deliberate intervention by humans to affect the course of evolution, they may be differentiated from those that ordinarily do not, which may be broadly envisioned as mechanisms constituting “natural selection” in *the broad sense* as opposed to competition between conspecifics as “natural selection” in *the narrow sense* more commonly associate with Darwin. But modes of communication, such as language, and other aspects of culture, such as tools, clothing, shelter, and types of transportation, for example, also exert an influence on the course of evolution. And species can ultimately be distinguished based upon their varying degrees of cognitive versatility, which contribute to their behavioral plasticity.

As a function of changes in gene pools across time, therefore, biological evolution may be understood, first, as a set of causal mechanisms (specifically, those identified above or an alternative set); second, as a set of historical

explanations (in which those mechanisms are applied to specific historical conditions); and, third, as a branching tree structure that reflects the evolution of species (as a manifestation of the cumulative effect of those mechanisms operating across time). But it should come as no surprise if an adequate understanding of biological evolution should entail an adequate understanding of cultural evolution, where an adequate theory of *gene-culture co-evolution* presupposes an account of mentality and its connections to consciousness. Evolution and consciousness are explored in the chapters that appear in this volume.

## Prologue

As Stevan Harnad contends in the Prologue, the theory of evolution appears to be hard pressed to explain the adaptive value and causal contribution of consciousness in human and non-human animals. One problem is that — unless we embrace dualism and treat it as some sort of independent and non-physical force — consciousness may or may not have an independent adaptive function of its own, over and above those behavioral and physiological functions it might supervene upon, because evolution is completely blind to the behavioral differences between conscious organisms and their functionally equivalent non-conscious counterparts, who are referred to as “zombies”. And this is the case because natural selection itself operates at the level of behavior.

As Harnad expresses it, “the Blind Watchmaker” (natural selection), a functionalist if ever there were one, is no more a mind reader than we are. Hence, if we designate behavioral similarity as “Turing indistinguishability” in the spirit of Turing’s test (TT), then it follows that *Turing-Indistinguishability equals Darwinian-Indistinguishability*. Even though organisms that are Turing-Indistinguishable would also be indistinguishable regarding their behavioral responses and therefore possess all and only the same adaptive capabilities, it (somewhat surprisingly) does not follow that human behavior is therefore adequately explainable on the basis of zombie physical determinism alone.

We are conscious and, more importantly, our consciousness somehow “piggy-backs” on a vast complex of unobservable internal activity — call it “cognition” — that appears to be responsible for generating most of our behavioral responses. Apart from those brought about by irrational or non-rational forces, such as instinctual sexual desires, where distal Darwinian factors continue to exert proximal influence, it is roughly as sensible to seek Darwinian rather than cognitive explanations for most of our current behavior

as it is to seek cosmological rather than engineering explanations of an auto's performance. Evolutionary theory can explain what has shaped our cognitive capacity, but cognitive theory must explain response behavior when it is affected by cognition.

## Part I: Natural consciousness

As Tom Polger and Owen Flanagan observe, consciousness and evolution are rather complex phenomena. It is sometimes thought that, if adaptation explanations for some varieties of consciousness, say, conscious visual perception, can be secured, then we may be reassured that at least those kinds of consciousness are not merely "epiphenomena" as phenomena that accompany other processes yet are causally inert. But what if other varieties of consciousness, such as dreams, for example, are not adaptations? Polger and Flanagan sort out the various and subtle connections among evolution, adaptation, and epiphenomenalism in an attempt to demonstrate that the consequences of epiphenomenalism for understanding consciousness are not so dire as some have supposed.

David Cole contends that consciousness is not a single phenomenon and cannot be adequately understood as having a unique function. He distinguishes various aspects or species of consciousness (including what he calls "creature consciousness", "metaconsciousness", "propositional thought", and "qualia"), while suggesting that each has its own distinctive function. He focuses upon qualia, which he envisions as "the hard problem", where the main obstacles are coping with inverted spectrum alternatives and with missing qualia (or zombie) possibilities. He offers a "two-factor" theory of qualia, according to which there is a way things are for us accounted for by metaconsciousness, while the specific qualitative content of how things are for us is explained by the functional role of imagistic representations in complex connectionist systems.

Terry Horgan and George Graham define an interdisciplinary research program for consciousness. First, they identify the "causal grain" of phenomenal states as neurophysical and functional-representational levels, where scientific progress is expected. Second, they discuss three key philosophical puzzles about phenomenal consciousness, which concern its ontological status, its causal role, and its explainability. Third, they argue that, from the perspective of our current epistemic existential situation, even if the causal grain of phenomenal consciousness were to become fully understood within

cognitive science, various theoretical options about how to understand qualia that are presently “live-options” in philosophical discussion would continue to be live-options.

## **Part II: Special adaptations**

As James W. Garson observes, the hypothesis of the existence of an innate, species-specific mental language (or “language of thought”) seems to be an attractive thesis to account for the propositional nature of higher-order consciousness. But this prospect raises questions about the compatibility of an innate mental language relative to the presumption that it should be the product of an evolutionary process. This chapter examines research in genetic programming in order to ascertain whether or not the language of thought hypothesis is compatible with evolution. Indeed, two different problems arise here. The first concerns the evolution of the causal mechanisms that embed systems of symbols in such a causal role. Research on genetic programming does not support their evolvability. The second concerns the acquisition of symbol systems during the life of the organism. Here the evidence seems more favorable.

Selmer Bringsjord and Ron Noel remark that you (the reader) and I (the editor), the two of them, Plato, Darwin, and our neighbors have not only been conscious, but at some point have decided to continue to live in order to continue to be conscious (of that rich chocolate ice cream, that lover’s tender touch, that joyful feeling of “Eureka!” when that theorem is finally proved). For us, consciousness is, somewhat barbarically, “a big deal”. But is it for evolution? Apparently. After all, we evolved. But why did evolution bother to give us consciousness? The authors refine this question, and then proceed to give what they regard as the only truly satisfactory answer, which appeals to what they take to be an intimate connection between consciousness and creativity.

Stephen Clark suggests the difficulty with “dualistic theories” is that once “matter” has been divided from “mind”, it becomes impossible to conceive of anything but an extrinsic, brute relationship between the two, about which we could only learn from experience. Unfortunately, no such link between mind and matter could itself be the object of experience. It follows from this, from the absurdity of epiphenomenalist and of materialist accounts of thinking — and the contentlessness of merely mathematical descriptions of things — that we ought to abandon the hypothesis of a material world distinct from experi-

enced reality, since such a world could not be experienced and offers no explanation of our actual experience. The unexpected conclusion is that a coherent account of the world must show how the phenomenal world has evolved.

### **Part III: Artificial consciousness**

Stefano Nolfi and Oraxio Miglino explore the power and limits of sensory-motor coordination, which they consider to be one of the lowest, if not the lowest, levels of consciousness. Acknowledging that the word “consciousness” tends to be used with various different meanings, they attempt to investigate the conditions under which internal representations might be expected to emerge, especially in relation to the internal dynamics of organisms in populations of conspecifics that must interact with changeable external environments. They contrast the approach they favor with one known as “behavior based robotics”, where interactions with the environment tend to be reactive and internal representations/internal dynamics have more limited roles.

Dario Floreano explores the hypothesis that some of today’s robots might possess a form of consciousness whose substrate is a mere algorithm. First, consciousness is defined within an evolutionary framework as awareness of one’s own state in relation to the external environment. Then the basic prerequisites for conscious activity, thus understood, are discussed, namely: embodiment, autonomy, and suitable adaptative mechanism. Artificial evolution, rather than evolutionary optimization, is presented as a viable methodology to create conscious robots, whose behavior they exemplify. And he contends that what might be thought to be problematical with the concept of “robot consciousness” is not the notion of a robot, but the concept of consciousness.

Inman Harvey contends that the design of autonomous robots has an intimate relationship with the study of autonomous animals and humans, where robots afford convenient “puppet shows” that illustrating current myths about cognition. Whether we like it or not, any approach to the design of autonomous robots is invariably underpinned by some philosophical position of the designer. While philosophical positions are typically subjected to rational criticism, in building situated robots, philosophical positions first affect design decisions and are later tested in the real world by “doing philosophy of mind with a screwdriver”. He distinguishes various kinds of “problems of consciousness” and suggests that easy types of consciousness can be possessed

by robots and validated by objective tests, while the hard kind, which concerns qualia or subjectivity, reflects a confusion that can be dissolved in the fashion of Wittgenstein.

## Epilogue

Neil Tennant, finally, explores the future of cloning as an extreme form of artificial selection. He speculates how human sexuality might be subjected to an evolutionary re-configuration by access to the technology of cloning. Present psychological differences between the sexes, after all, have arisen from the pressure of natural selection (including sexual selection) in the past, operating on a system of sexual reproduction with differing parental investments. Genetically variable dispositions toward sexual behavior, therefore, might be dramatically changed were cloning to become available as an alternative to sexual reproduction. He considers implications of such a prospect for the “war between the sexes” and how sibling rivalry might be affected, concluding with a stimulating and enjoyable discussion of the possibilities for serial immortality.

The very idea that I might be “my own grandpa”, as the country and western song has it, is now as fascinating as it is amusing, but encounters serious difficulties, some of which are only beginning to be appreciated. Human cloning seems inevitable, but of course these clones are only genetically identical (with respect to their genetically-relative traits, the development of which can be affected in the interuterine environment), and not with respect to their environmentally-relative acquisitions of specific attributes (which may include the formation of basic emotional and mental traits as a function of experience), as the film, “The Boys from Brazil”, illustrates with respect to (hopefully, imaginary) attempts to replicate additional instantiations of Adolf Hitler.

Twin studies suggest that many traits are ones toward which humans specifically are strongly predisposed without being disposed, where a rough “rule of thumb” has it that approximately 60% of many traits are inherited while 40% of those same traits are acquired. But the very prospect of cloning confronts unexpected problems relative to the “copying errors” that occur during the process of cloning itself, where, because of the influence of polygenic and pleiotropic effects, extremely minute differences in genes and proteins can bring about substantial and unexpected differences in phenotypes. The result is that even creating genetic replicas of original organisms appears to be vastly more subtle and complex than has heretofore been as-

sumed, where our greatest hopes and most dire fears may be contrained by limitations of that process.

Theories of consciousness are not complete theories of mind, of course, as some of the contributors to this volume have observed. The focus on consciousness may even appear distracting to a degree, since what we need is a theory about the nature of the mind that brings consciousness along with it as an essential but natural phenomenon. If we are evolutionary successors of *E. coli* bacteria, for example, whose receptor proteins combine with chemotactic substances to affect locomotion (with twelve specific attractants and eight specific repellants), however, then consciousness and mentality may indeed be inextricably intertwined. Other readers are therefore likely to find, as I have found, that the essays gathered together here afford stimulating perspectives on a central problem about the human species and its place in nature, one upon which — thanks to these authors and to others of similar inclination — we are making and continue to make important progress.

J. H. F.



## **| Prologue**



# Turing indistinguishability and the blind watchmaker

Stevan Harnad  
Southampton University

## Consciousness cannot be an adaptation

Here's an argument to try out on those of your adaptationist friends who think that there is an evolutionary story to be told about the "survival value" of consciousness: Tell me whatever you think the adaptive advantage of doing something consciously is, including the internal, causal mechanism that generates the capacity to do it, and then explain to me how that advantage would be lost in doing exactly the same thing unconsciously, with *exactly the same causal mechanism*.

Here are some examples: It is adaptive to feel pain when your leg is injured, because then you spare the leg and avoid the cause of the injury in future. (This account must be coupled, of course, with a causal account of the internal mechanism for detecting tissue damage and for learning to avoid similar circumstances in the future.) How would the advantage be lost if the tissue damage were detected unconsciously, the sparing of the leg were triggered and maintained unconsciously, and the circumstances to avoid were learned and avoided unconsciously? In other words: identical internal mechanisms of detection, learning, avoidance, but no consciousness?

Another example: It is adaptive to pay conscious selective attention to the most important of the many stimuli impinging on an organism at any one time. (This must be paired with a causal account of the mechanism for attending selectively to input and for detecting and weighting salient information.) How would that adaptive advantage be lost if the input selection and salience-detection were all taking place unconsciously? What is the advantage of conscious fear over unconscious vigilance, danger-detection and avoidance? Of conscious recall and retrieval from memory over mere recall and retrieval? Conscious inference over unconscious inference? Conscious discrimination over unconscious? Conscious communication? Conscious discourse? Con-

scious cognition? And all these comparisons are to be made, remember, in the context of an internal mechanism that is generating it all: the behavior, learning, memory, and the consciousness accompanying it.

The point that I hope is being brought out by these examples is this: An adaptive explanation must be based on differential consequences for an organism's success in surviving and reproducing. This can even include success in the stock market, so the problem is not with the abstractness or abstruseness of the adaptive function in question, it is with the need for *differential consequences* (Catania & Harnad 1988). Adaptive consequences are functional consequences. A difference that makes no functional difference is not an adaptive difference. *The Blind Watchmaker* (Dawkins 1986) is no more a mind-reader than any of the rest of us are. He can be guided by an organism's capacity to detect and avoid tissue injury, but not by its capacity or incapacity to *feel* pain while so doing. The same is true for conscious attention vs. unconscious selectivity, conscious fear vs. unconscious danger avoidance, conscious vs. unconscious memory, inference, discrimination, communication, discourse, cognition.

So for every story purporting to explain the adaptive advantage of doing something consciously, look at the alleged adaptive advantage itself more closely and it will turn out to be a functional advantage (consisting, in the case of cognition, of a performance capacity and the causal mechanism that generates it); and that exact same functional advantage will turn out to remain intact if you simply *subtract* the consciousness from it (Harnad 1982, 1991, 2001).

Indeed, although the comparison may seem paradoxical (since we all know that we are in fact conscious), those who have tried to claim an evolutionary advantage for consciousness are not unlike those uncritical computer scientists who are ready to impute minds even to the current generation of toy computational models and robots (Harnad 1989): There is an interesting similarity between claiming that a thermostat has (rudimentary) consciousness and claiming that an organism's (real) consciousness has an adaptive function. In both cases, it is a mentalistic *interpretation* that is misleading us: In the case of the organism that really is conscious, the interpretation of the organism's state as conscious happens to be correct. But the imputation of functionality (over and above the adaptive function of its unconscious causal mechanism) is as gratuitous as the interpretation of the thermostat as having a consciousness at all, and for roughly the same reason: The conscious interpretation is not needed to explain the function.

Why do we have the conviction that consciousness must have survival value? Well, in part it must be because evolutionary theory is not in the habit of viewing as prominent and ubiquitous a biological trait as consciousness as just a causal dangler like an appendix, a “spandrel” (Gould 1994), or worse. A partial reply here might be that there are reasons for believing that the mind could be rather special among biological traits (it is surely not a coincidence that centuries of philosophy have been devoted to the mind/body problem, not the “blue-eye/brown-eye” problem, or even the “phenotype/genotype” problem). But I suspect that the real reason we are so adaptationistic about consciousness has to do with our experience with and intuitions about free will (Dennett 1984). We are convinced that if/when we do something consciously, it’s because we choose to do it, not because we are unconsciously impelled to do it by our neurophysiology (Libet 1985). So it’s natural to want to establish an adaptive value for that trait (free will) too.

Yet it seems clear that there is no room for an *independent* free will in a causal, functional explanation unless we are prepared to be dualists, positing mental forces on a par with physical ones, and thereby, I think, putting all of physics and its conservation laws at risk (Alcock 1987). I don’t think the mental lives of medium-sized objects making up the relatively minuscule biomass of one small planet in the universe warrant such a radical challenge to physics; so let us assume that our feeling of free will is caused by our brains, and that our brains are the real causes of what we do, and not our free wills.

This much explains why not many people are telling adaptive stories directly about free will: Because it leads to embarrassing problems with causality and physics. Yet I think our *feeling* of free will is still behind the motivation to find an adaptive story for consciousness, and I think the latter is wrong-headed for about the same reason: If it is clear why it is not a good idea to say that there is a selective advantage for an organism that can *will* its actions (as opposed to having its brain cause them for it), it should be almost as clear why it is not good to say that there is a selective advantage for an organism that really *sees blue* as opposed to merely detecting and responding to blue. We really see blue alright, but there’s no point trying to squeeze an adaptive advantage out of *that*, since we have no idea *how* we manage to see, detect, or respond to blue. And once we do understand the causal substrate of that, then *that causal substrate* and the functional capacities it confers on our bodies will be the basis of any adaptive advantages, not the consciousness of blue.

## Reverse engineering and Turing indistinguishability

How are we to arrive at a scientific understanding of that causal substrate? First, I think we have to acknowledge that, as Dennett (1994, 1995) has suggested, the behavioral and cognitive sciences and large parts of biology are not basic sciences in the sense of physics and chemistry, but branches of “reverse engineering.” Basic sciences study and explain the fundamental laws of nature. Forward engineering then applies those laws to designing and building useful things such as bridges, furnaces, and airplanes, with stipulated functional capacities. Reverse engineering, by contrast, inherits systems that have already been designed and built by the Blind Watchmaker with certain adaptive functional capacities, and its task is to study and explain the causal substrate of those capacities.

Clearly, what reverse engineering needs first is a methodology for finding that causal substrate: a set of empirical constraints that will reliably converge on them. The logician Alan Turing (1964) has provided the basis for such a methodology, although, as you will see, his original proposal needs considerable modification because it turns out to be just one level of a (“Turing-”) hierarchy of empirical constraints (Harnad 1994a, 2000).

According to Turing’s Test, a machine has a mind if its performance capacity (i.e., what it can do) is indistinguishable from that of a person with a mind. In the original version of the Turing Test (T2), the machine was removed from sight so no bias would be introduced by its appearance (the indistinguishability had to be in performance, not in appearance). Then (although this is not how Turing put it), the machine had to be able to correspond (by exchanging letters) with real people for a lifetime in such a way that it could not be distinguished from a real pen-pal. There are accordingly two dimensions to the Turing Test: The candidate (1) must have all the performance capacities of a real person and (2) its performance must be indistinguishable *from* that of a real person *to* (any) real person (for a lifetime — I add this to emphasize that short-term tricks were never the issue: the goal was to really generate the total capacity; Harnad 1992b).

T2 has been the subject of much discussion, most of it not pertinent here (Harnad 1989). What is pertinent is that the out-of-sight constraint, which was intended only to rule out irrelevant biases based on appearance, also inadvertently ruled out a lot of human performance: It ruled out all of our robotic capacity (Harnad 1995b), our capacity to discriminate, manipulate, and catego-

alize those very objects, properties, events and states of affairs that the symbols in our pen-pal correspondence are about (Harnad 1987, 1992a, Harnad et al. 1995). So although the symbolic level of performance to which T2 is restricted is a very important one (and although there are even reasons to think that T2 could not be successfully passed without drawing indirectly upon robotic capacity), it is clear that human performance capacity amounts to a lot more than what can be tested directly by T2. Let us call a test that calls for Turing-Indistinguishable symbolic *and* robotic capacity the Total Turing Test, or T3. A T3 robot would have to be able to live among us and interact with the people and objects in the world Turing-Indistinguishably from the way we do.

At this point people always ask: Well, *how* indistinguishably? What about the question of appearance? Would it have to be able to shave? A little common sense is needed here, keeping in clear sight the fact that this is not about tricks or arbitrary stipulations (Harnad 1994a, 1995a). The point of Turing Testing is to generate functional capacities. What one aims for is generic capacities. Just as a plane has to be able to fly, but doesn't have to look like or fly exactly like any *particular* DC-11 — it just has to have flying capacity Turing-indistinguishable from that of planes in general — so a T3 robot would only have to have our generic robotic capacities (to discriminate, manipulate, categorize, etc.), not their fine-tuning as they may occur in any particular individual.

But T3 is not the top of the Turing hierarchy either, for there *is* more that one could ask if one wanted to capture Turing-Indistinguishably every reverse engineering fact about us, for there are also the internal facts about the functions of our brains. A T4 candidate would be Turing indistinguishable from us not only in its symbolic and robotic capacities but also in its neuromolecular properties. And T4, need I point out, is as much as a scientist can ask, for the empirical story ends there.

So let's go for T4, you are no doubt straining to say. Why bother with T3 or T2 at all? Well, there are good reasons for aiming for something less than T4, if possible. For one thing, (1) we already know, in broad strokes, what our T3 capacity is. The T3 data are already in, so to speak, so there we can already get to work on the reverse engineering. Comparatively little is known so far about the brain's properties (apart from its T3 capacity, of course). Furthermore, it is not obvious that we should wait till all the brain data are in, or even that it would help to have them all, because (2) it is not at all clear which of the brain's properties are relevant to its T3 capacities. And interesting though they are in their own right, it is striking that (3) so far, T4 neuroscientific data have not yet

been helpful in providing functional clues about how to reverse-engineer T3 capacity. Turing also had a valid insight, I think, in implicitly reminding us that we are not mind-readers with one another either, and that (4) our intuitive judgments about other people's minds are based largely on Turing Indistinguishable performance (i.e., T2 and T3), not on anything we know or think we know about brain function.<sup>1</sup>

There is one further reason why T3 rather than T4 might be the right level of the Turing hierarchy for mind-modelling; it follows from our earlier discussion of the absence of selective advantages of consciousness: The Blind-Watcher is likewise not a mind-reader, and is hence also guided only by T3. Indeed, T4 exists in the service of T3. T4 is one way of generating T3; but if there were other ways, evolution would be blind to the differences between them, for they would be functionally — hence adaptively — indistinguishable (Harnad 2000).

### Undetermination of theories by data

Are there other ways to pass T3, apart from T4? To answer that we first have to consider the general problem of “scientific underdetermination.” In basic science, theories are underdetermined by data. Several rival theories in physics may account equally well for the same data. As long as the data that a particular theory accounts for are subtotal (just “toy” fragments of the whole empirical story — what I call “t1” in the T-hierarchy), the theory can be further calibrated by “scaling it up” to account for more and more data, tightening its empirical degrees of freedom while trimming excesses with Occam's razor. Only the fittest theories will scale all the way up to T5, the “Grand Unified Theory of Everything,” successfully accounting for all data, past, present and future; but it is not clear that there will be only one survivor at that level. All the “surviving” rival theories, being T5-indistinguishable, which is to say, completely indistinguishable empirically, will remain eternally underdetermined. The differences among them make no empirical difference; we will have no way of knowing which, if any, is the “right” theory of the way the world really is. Let us call this *ordinary* scientific underdetermination. It's an unresolvable level of uncertainty that even physicists have to live with, but it does not really cost them much, since it pertains to differences that do not make any palpable difference to anyone.

There is likewise underdetermination in the engineering range of the Turing hierarchy (T2 - T4). T2 is the level of symbolic, computational function, and here there are several forms of underdetermination: One corresponds to the various forms of computational equivalence, including Input/Output equivalence (also called Turing Equivalence) and Strong Equivalence (equivalence in every computational step) (Pylyshyn 1984). The other is the hardware-independence of computation itself: the fact that the same computer program can be physically implemented in countless radically different ways. This extreme form of underdetermination is both an advantage and a disadvantage. With it goes the full power of formal computation and the Church-Turing Thesis (Church 1936, Turing 1937) according to which *everything* can be simulated computationally. But it has some liabilities too, such as the symbol grounding problem (Harnad 1990, 1994b), because the meanings of symbols are not intrinsic to a symbol system; they are parasitic on the mind of an external interpreter. Hence, on pain of infinite regress, symbols and symbol manipulation cannot be a complete model for what is going on in the mind of the interpreter.

There is underdetermination at the T3 level too. Just as there is more than one way to transduce light (e.g., Limulus's ommatidia, the mammalian retina's rods and cones, and the photosensitive cell at your local bank; Fernald 1997) and more than one way to implement an airplane, so there may be more than one way to design a T3 robot. So there may well be T3-indistinguishable yet T4-distinguishable robots. The question is, will they all have a mind, or will only T4 robots have one? Note that the latter question concerns a form of underdetermination that is much more radical than any I have mentioned so far. For unlike T5 underdetermination in physics, or even T2 ungroundedness, T3 underdetermination in mind-modelling involves a second kind of difference, over and above ordinary empirical underdetermination, and that difference does make a palpable difference, but one that is palpable to only one individual, namely, the T3 candidate itself (Descartes' "Cogito"). This extra order of underdetermination is the (Cartesian) mark of the mind/body problem and it too is unresolvable; so I propose that we pass over it in silence, noting only that, scientifically speaking, apart from this extra order of uncertainty, the T3-indistinguishable candidates for the mind are on a par with T5-indistinguishable candidates for the Grand Unified Theory of Everything, in that in both cases there is no way we can be any the wiser about whether or not they capture reality, given that each of them can account for all the data (Harnad 2000).<sup>2</sup>

- t1 toy fragment of human total capacity
- T2 Total Indistinguishability in symbolic (pen pal) performance capacity
- T3 Total Indistinguishability in robotic (including symbolic) performance capacity
- T4 Total Indistinguishability in neural (including robotic) function
- T5 Total Physical Indistinguishability

**Figure 1**

Is T4 a way? In a sense it is, because it is certainly a tighter empirical approximation to ourselves than T3. But the extra order of underdetermination peculiar to the mind/body problem (the fact that, if you will, *empiricism* is no mind reader either!) applies to T4 as well. Only the T4 candidate itself can know whether or not it has a mind; and only the T3 candidates themselves can know whether or not we would have been wrong to deny them a mind for failing T4, having passed T3.

The T-hierarchy is a hierarchy of empirical constraints (Figure 1). Each successive level tightens the degrees of freedom on the kinds of candidates that can pass successfully. The lowest, “toy” level, t1, is as underconstrained as can be because it captures only subtotal fragments of our total performance capacity. There are countless ways to generate chess-playing skills, arithmetic skills, etc.; the level of underdetermination for arbitrary fragments of our Total capacity is far greater than that of ordinary scientific underdetermination. T2 is still underconstrained, despite the formal power of computation and the expressive and intuitive power of linguistic communication, because of the symbol grounding problem (Harnad 1990) and also because T2 too leaves out the rest of our performance capacities. T4 is, as I suggested, overconstrained, because not all aspects of brain function are necessarily relevant to T3 capacity, and it is T3 capacity that was selected by evolution. So it is T3, I would suggest, that is the right level in the T-hierarchy for mind-modelling.

I could be wrong about this, of course, and let me describe how: First, the question of “appearance” that we set aside earlier has an evolutionary side too. Much more basic than the selection of the mechanisms underlying performance capacity is the selection of morphological traits, both external ones that we can see and respond to (such as plumage or facial expression) and internal ones (such as the macromorphology and the micromorphology [the physiology and the biochemistry] of our organs, including our brain). The Blind Watchmaker may be blind to T3-indistinguishable differences underlying our performance capacity, but he is not blind to morphological differences, if they

make an adaptive difference. And then of course there is the question of “shape” in the evolutionary process itself: the shape of molecules, including the mechanism of heredity, and the causal role that that plays. And we must also consider the status of certain special and rather problematic “robotic” capacities, such as the capacity to reproduce. Morphological factors are certainly involved there, as they are involved in other basic robotic functions, such as eating and defecation; there might well prove to be an essential interdependency between cognitive and vegetative functions (Harnad 1993a).

But let us not forget the monumental constraints already exerted by T3 alone: A causal mechanism must be designed that can generate our full performance capacity. To suppose that this is not constraint enough is to suppose that there could be mindless T3 Zombies (Harnad 1995c), and that only the morphological constraints mentioned above could screen them out. But, as has been noted several times earlier, this would be a remarkable coincidence, because, even with “appearance” supplementing T3, indeed, even with the full force of T4, evolution is still not a mind-reader. It seems more plausible to me that T3 itself is the filter that excludes Zombies: that mindless mechanisms are not among the empirical possibilities, when it comes to T3-scale capacity.

In any case, even if I’m wrong, T3 seems to be a more realistic goal to aim for initially, because the constraints of T3 — the requirement that our model generate our full robotic performance capacity — are positive ones: Your model must be able to do all of this (T3). The “constraints” of T4, in contrast, are, so far, only negative ones: They amount to a handicap: “However you may manage to get the T3 job done, it must be done in this brainlike way, rather than any old way.” Yet at the same time, as I have suggested, no positive clue has yet come from T4 (neurobiological) research that has actually helped generate a t1 fragment of T3 capacity *that one could not have generated in countless other ways already*, without any T4 handicaps. So the optimal strategy for now seems to be a division of labor: Let mind-modelers do T3 modelling and let brain-modelers do T4.

Obviously, if T4 work unearths something that helps generate T3, then T3 researchers can help themselves to it; and of course if T4 research actually *attains* T4 first, then there is no need to continue with T3 at all, because T4 subsumes T3. But if T3 research should succeed in scaling up to a T3-passer first, we could then fine-tune it, if we liked, so it conforms more and more closely to T4 (just as we could calibrate it to include more of the fine-tuning of behavior mentioned earlier).

Or we could just stop right there, forget about the fine-tuning, and accord civil rights to the successful T3-passer. I, for one, would be prepared to do so, since, not being a mind-reader myself, I would really not feel that I had a more compelling basis for doubting that it feels pain than I do in the case of my natural fellow creatures. Never mind. The purpose of this excursion into the Turing hierarchy was to look for methodological and empirical constraints on the reverse engineering of the mind, and we have certainly found them; but whether your preference is for T3 or for T4, what the empirical agenda yields, once it is completed, is a causal mechanism: one that is capable of generating all of our T3 capacities (in a particular way, if you prefer T4). And as I have stressed repeatedly, neither T3 nor T4 can select directly for consciousness *per se*, for there is no Turing-distinguishable reason that anything that can be done consciously cannot be done unconsciously just as well, particularly since what does the causal work cannot be the consciousness itself but only the mechanism we have laboriously reverse-engineered till it scaled up to T3.

So, since, for all the T3 or the Blind-Watchmaker can determine, the candidate might as well be a Zombie, does it follow that those who have been stressing the biological determinism of behavior (Dawkins 1989; Barkow et al. 1992) are closer to the truth than those who stress cognition, consciousness and choice?

### **Are we driven by our Darwinian unconscious?**

Let's consider specific examples. The following kind of suggestion has been made (e.g., by Shields & Shields 1983 and Thornhill & Thornhill 1982; more recently by Baker 1996, in a curious juxtaposition of pornography and biopsychodynamic hermeneutics that, had the writing been better, would be reminiscent of Freud; and most recently by Miller 2001): For reasons revealed by game-theoretic and inclusive-fitness assumptions and calculations, there are circumstances in which it is to every man's biological advantage to rape. Our brains accordingly perform this calculation (unconsciously, of course) in given circumstances, and when its outcome reveals that it is optimal to do so, we rape. Fortunately, our brains are also sensitive to certain cues that inhibit the tendency to rape because of the probability of punishment and its adverse consequences for fitness.

This is also a result of an unconscious calculation. So we are rather like

Zombies being impelled to or inhibited from raping according to the push and pull of these unconscious reckonings. If we see a potential victim defenceless and unprotected, and there is no indication that we will ever be caught or anyone will ever know, we feel inclined to rape. If we instead see the potential victim flanked by a pair of burly brothers in front of a police station, we feel inclined to abstain. If the penalties for rape are severe and sure, we abstain; if not, we rape.

Similarly, there is an unconscious inclusive fitness calculator that assesses the advantages of mating with one's sibling of the opposite sex (van den Berghe 1983). Ordinarily these advantages are vastly outweighed by the disadvantages arising from the maladaptive effects of inbreeding. However, under certain circumstances, the advantages of mating with a sibling outweigh the disadvantages of inbreeding, for example, when great wealth and status are involved, and the only alternative would be to marry down (as in the case of the Pharaohs). To put it dramatically, according to the function of this unconscious biological calculator, as we approach the pinnacle of wealth and status, my sister ought to be looking better and better to me.

These explanations and these hypothetical mechanisms would make sense, I suggest, if we really were Zombies, pushed and pulled directly by unconscious, dedicated "proximal mechanisms" of this kind. But what I think one would find in a T3-scale candidate, *even a T3 Zombie*, would not be such unconscious, dedicated proximal mechanisms, but other, much more sophisticated, powerful and general cognitive mechanisms, most of them likewise unconscious, and likewise evolved, but having more to do with general social and communicative skills and multipurpose problem-solving and planning skills than with any of the specifics of the circumstances described. These evolved and then learned T3 capacities would have next to nothing to do with dedicated fitness calculations of the kind described above (with the exception, perhaps, of basic sexual interest in the opposite sex itself, and its inhibition toward those with whom one has had long and early contact, i.e., close kin).

The place to search for Darwinian factors is in the origin of our T3 capacity itself, not in its actual deployment in a given individual lifetime. And that search will not yield mechanisms such as rape-inhibition-cue-detectors or status-dependent-incest-cue-detectors, but general mechanisms of social learning and communication, language, and reasoning. The unconscious substrate of our actual behavior in specific circumstances will be explained, not by simplistic local Darwinian considerations (t1 "toy" adaptationism, shall we call

it?), but by the T3 causal mechanism eventually revealed by the reverse engineering of the mind. The determination of our behavior will be just as unconscious as biological determinism imagines it will be, but the actual constraints and proximal mechanisms will not be those dictated directly by Darwin but those dictated by Turing, his cognitive engineer (Cangelosi & Harnad 2002).<sup>3</sup>

What, then, is the role of the mind in all this unconscious, causally determined business? Or, to put it another way, why *aren't* we just Zombies? Concerns like these are symptomatic of the mind/body problem, and that, it seems to me, is going to beset us till the end of time — or at least till the end of conscious time. What is the mind/body problem? It's a problem we all have with squaring the mental with the physical, with seeing how a mental state, such as feeling melancholy, can be the same as a physical state, such as certain activities in brain monoamine systems (Harnad 1993b).

The old-style “solution” to the mind/body problem was simply to state that the physical state and the mental state were the same thing. And we can certainly *accept* that (indeed, it's surely somehow true), but what we can't do is *understand* how it's true, and that's the real mind/body problem. Moreover, the sense in which we do not understand how it's true that, say, feeling blue is really being low in certain monoamines, is, I suggest, very different from the kinds of puzzlement we've had with other counterintuitive scientific truths. For, as Nagel (1974, 1986) has pointed out (quite correctly, I think), the understanding of all other counterintuitive scientific truths except those pertaining to the mind/body problem has always required us to translate one set of appearances into a second set of appearances that, on first blush, differed from the first, but that, upon reflection, we could come to see as the same thing after all: Examples include coming to see water as  $H_2O$ , heat as mean kinetic energy, life as certain biomolecular properties, and so on.

The reason this substitution of one set of appearances for another was no problem (given sufficient evidence and a causal explanation) was that, although appearances changed, *appearance itself* was preserved in all previous cases of intuition-revision. We could come to see one kind of thing as another kind of thing, but we were still seeing (or picturing) it as *something*. But when we come to the mind/body problem, it is appearance itself that we are inquiring about: What are appearances? — for mental states, if you think about it, *are* appearances: they are what it feels like to perceive things. So when the answer is that appearances are really just, say, monoaminergic states, then that appearance-to-appearance revision mechanism (or “reduction” mechanism, if you

prefer) that has stood us in such good stead time and time again in scientific explanation fails us completely. For what precedent is there for substituting for a previous appearance (feeling), not a new (though counterintuitive) appearance (feeling), but no appearance at all (just physics)?

This, at least, is how Nagel evokes the lasting legacy of the mind/body problem. It's clearly more than just the problem of ordinary underdetermination, but it too is something we're going to have to live with. For whether your preference is for T3 or T4, it will always take a blind leap of faith to believe that the candidate has a mind. Turing Indistinguishability is the best we can ever do. Perhaps it's some consolation that the Blind Watchmaker could do no better.

## Notes

1. The work of some authors on "theory of mind" in animals (Premack & Woodruff 1978) and children (Gopnik 1993) and of some adult theorists of the mind when they adopt the "intentional stance" (i.e., when they interpret others as having beliefs and desires; Dennett 1983) can be interpreted as posing a problem for the claim that consciousness cannot have had an adaptive advantage of its own. "Theory of mind" used in this nonstandard way (it is not what philosophers mean by the phrase) corresponds in many respects to Turing-Testing: Even though we are not mind-readers, we can tell pretty well what (if anything) is going on in the minds of others (adults, children, animals): We can tell when others are happy, sad, angry, hungry, menacing, trying to help us, trying to deceive us, etc. To be able to do this definitely has adaptive advantages. So would this not give the Blind Watchmaker an indirect way of favouring those who have mental states? Is it not adaptive to be able to infer the mental states in others?

The problem is that the adaptive value of mind-reading (Turing Testing) depends entirely on how (1) the appearance and behaviour of others, (2) the effects of our own appearance and behaviour on others, and (3) what it feels like to be in various mental states, covary and cue us about things that matter to our (or our genes') survival and reproduction. We need to know when someone else threatens us with harm, or when our offspring need our help. We recognise the cues and can also equate them with our own feelings when we emit such cues. The detection of internal state correlates of external cues like this is undeniably adaptive. But what is the adaptive value of actually *feeling* something in detecting and using these correlates? Nothing needs to be *felt* to distinguish real affection from feigned affection, in oneself or in others. And real and feigned affection need not be based on a difference in feelings, or on any feelings at all.

Dennett is the one who has argued most persuasively for the necessity and the utility of adopting the intentional stance, both in order to live adaptively among one's fellow creatures and in order to reverse-engineer them in the laboratory. But let us not forget that

Dennett's first insight about this was based on how a chess-playing computer programme can only be understood if one assumes that it has beliefs and desires. Let us not entertain here the absurd possibility that a computer running a chess-playing programme really has beliefs and desires. It is not disputed that interpreting it as if it had beliefs and desires is useful. But then all that gives us is an adaptive rationale for the evolution of organisms that can *act as if* they had minds and as if they could read one another's minds; all that requires is a causal exchange of signals and cues. It provides no rationale for actually feeling while all that unconscious processing is going on.

2. We could decide to accept the theory that has the fewest parameters, but it is not clear that God used Occam's Razor in designing the universe. The Blind Watchmaker certainly seems to have been profligate in designing the biosphere, rarely designing the "optimal" system; which means that evolution leaves a lot of nonfunctional loose ends. It is not clear that one must duplicate every last one of them in reverse-engineering the mind.
3. Language, after all, has evolved to make the explicit symbolic mode (T2) dominate the implicit sensorimotor one (T3) in our species (Cangelosi & Harnad 2002).

## References

- Alcock, J. E. 1987. "Parapsychology: Science of the anomalous or search for the soul?" *Behavioral and Brain Sciences*.10: 553–643.
- Baker, R. 1996. *Sperm Wars: Infidelity, sexual conflict and other bedroom battles*. London: Fourth Estate, 1996.
- Barkow, J., Cosmides, L. and Tooby, J. 1992. (eds.) *The Adapted Mind: Evolutionary psychology and the generation of culture*. NY: Oxford University Press.
- Cangelosi, A. & Harnad, S. 2002. "The adaptive advantage of symbolic theft over sensorimotor toil." *Evolution of Communication* (in press).
- Catania, A.C. & Harnad, S. (eds.). 1988. *The Selection of Behavior: The Operant Behaviorism of BF Skinner: Comments and Consequences*. New York: Cambridge University Press.
- Church A. 1936. "An unsolvable problem of elementary theory." *American Journal of Mathematics*. 58: 345–63.
- Dawkins, Richard. 1989. *The selfish gene*. Oxford; New York: Oxford University Press
- Dawkins, Richard 1986. *The blind watchmaker*. NY: Norton.
- Dennett, D.C. 1983. "Intentional systems in cognitive ethology: The "Panglossian paradigm" defended." *Behavioral & Brain Sciences* 6: 343–390.
- Dennett, D.C. 1984. *Elbow room: the varieties of free will worth wanting*. Cambridge, Mass.: MIT Press
- Dennett, D.C. 1994. "Cognitive Science as Reverse Engineering: Several Meanings of 'Top Down' and 'Bottom Up'." In: Prawitz, D., & Westerstahl, D. (Eds.) *International Congress of Logic, Methodology and Philosophy of Science*. (9th: 1991) Dordrecht: Kluwer.
- Dennett, D.C. 1995. *Darwin's dangerous idea: Evolution and the meanings of life*. London; New York: Allen Lane.

- Fernald, R.D 1997. "The evolution of eyes". *Brain Behavior and Evolution* 50: 253–259
- Gopnik, A. 1993. "How we know our minds: The illusion of first-person knowledge of intentionality". *Behavioral & Brain Sciences* 16: 29–113.
- Gould, S.J. 1994. "The spandrels of San Marco and the panglossian paradigm: A critique of the adaptationist programme". In: Sober, E. (Ed.). *Conceptual issues in evolutionary biology..Second edition*. Cambridge MA: MIT Press Massachusetts, Pp. 73–90. [Reprinted from *Proceedings of the Royal Society B*, London 205: 581. 1979.]
- Harnad, S. 1982. "Consciousness: An afterthought". *Cognition and Brain Theory*. 5: 29–47.
- Harnad, S. 1987. (ed.) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.
- Harnad, S. 1989. "Minds, Machines and Searle". *Journal of Theoretical and Experimental Artificial Intelligence* 1: 5–25.
- Harnad, S. 1990. "The Symbol Grounding Problem". *Physica D* 42: 335–346.
- Harnad, S. 1991. "Other bodies, Other minds: A machine incarnation of an old philosophical problem". *Minds and Machines* 1: 43–54.
- Harnad, S. 1992a. Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds.) *Connectionism in Context*. Springer Verlag.
- Harnad, S. 1992b. "The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion". *SIGART Bulletin* 3(4). (October. 9–10).
- Harnad, S. 1993a. "Artificial Life: Synthetic Versus Virtual". *Artificial Life III. Proceedings, Santa Fe Institute Studies in the Sciences of Complexity. Volume XVI*.
- Harnad S. 1993b. Discussion (passim). In: Bock, G.R. & Marsh, J. (Eds.) *Experimental and Theoretical Studies of Consciousness*. CIBA Foundation Symposium 174. Chichester: Wiley
- Harnad, S. 1994a. "Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life". *Artificial Life* 1(3): 293–301.
- Harnad, S. 1994b. "Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't". Special Issue on "What Is Computation" *Minds and Machines* 4:379–390
- Harnad, S. 1995a. "Does the Mind Piggy-Back on Robotic and Symbolic Capacity?" In: H. Morowitz (ed.) *The Mind, the Brain, and Complex Adaptive Systems*. Santa Fe Institute Studies in the Sciences of Complexity. Volume XXII. P. 204–220.
- Harnad, S. 1995b. Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) *The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents*. New Haven: Lawrence Erlbaum. Pp. 277–286.
- Harnad, S. 1995c. "Why and How We Are Not Zombies". *Journal of Consciousness Studies* 1: 164–167.
- Harnad, S. 1996. The Origin of Words: A Psychophysical Hypothesis In Velichkovsky B & Rumbaugh, D. (Eds.) *Communicating Meaning: Evolution and Development of Language*. NJ: Erlbaum: pp 27–44.
- Harnad, S. 2000. "Minds, Machines and Turing". *Journal of Logic, Language and Information* 9(4):425–445.
- Harnad, S. 2001. "No easy way out". *The Sciences* 41(2):36–42.
- Harnad, S. Hanson, S.J. & Lubin, J. 1995. Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. In: V. Honavar & L. Uhr (eds.) *Symbol Processors*

- and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration*. Academic Press. pp. 191–206.
- Libet, B. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action". *Behavioral and Brain Sciences* 8: 529–566.
- Miller, G. 2000. *The Mating Mind*. Doubleday.
- Nagel, T. 1974. "What is it like to be a bat?" *Philosophical Review* 83: 435–451.
- Nagel, T. 1986. *The view from nowhere*. New York: Oxford University Press.
- Premack, D. & Woodruff, G. 1978. "Does the chimpanzee have a theory of mind?" *Behavioral & Brain Sciences* 1: 515–526.
- Pylyshyn, Z.W. 1984. *Computation and cognition*. Cambridge MA: MIT/Bradford
- Shields, W.M. & Shields, L.M. 1983. "Forcible rape: An evolutionary perspective". *Ethology & Sociobiology* 4: 115–136.
- Steklis, H.D. & Harnad, S. 1976. "From hand to mouth: Some critical stages in the evolution of language". In: Harnad, S., Steklis, H.D. & Lancaster, J. B. (eds.) *Origins and Evolution of Language and Speech*. Annals of the New York Academy of Sciences 280: 445–455.
- Thornhill, R & Thornhill, Nancy W. 1992. "The evolutionary psychology of men's coercive sexuality". *Behavioral and Brain Sciences* 15: 363–421.
- Turing, A.M. 1937. "On computable numbers". *Proceedings of the London Mathematical Society* 2 42: 230–65
- Turing, A.M. 1964. Computing machinery and intelligence. In: *Minds and machines*. A. Anderson (ed.), Engelwood Cliffs NJ: Prentice Hall.
- Van den Berghe, P.L. 1983. "Human inbreeding avoidance: Culture in nature". *Behavioral and Brain Sciences* 6: 91–123.



PART I

Natural consciousness



# Consciousness, adaptation and epiphenomenalism

Thomas Polger and Owen Flanagan  
University of Cincinnati/Duke University

## 1. Consciousness and adaptation

The question of the adaptive advantage of consciousness has been introduced into the debate among philosophers of mind as support for one or another view of the nature and causal efficacy of consciousness. If we can explain how and why we came to be conscious, then that will shed light on what sort of thing, or process, or property consciousness is. If consciousness has been selected for because it is fitness enhancing then we may rest assured that it is causally efficacious, and the epiphenomenalist suspicion becomes less worrisome (Flanagan 1992). And the odds seem good that some kinds of consciousness are adaptations.

For example, surely acute pain states are adaptive. You place your hand in a fire. The fire is hot. Your hand hurts. The pain causes you to remove your hand from the fire. Pain has certain effects relative to human bodies that figure in explanations of our overall capacity to avoid serious injury. *Prima facie*, pain in humans is an adaptation for, among other things, causing us to remove our hands from fire and other sources of injury. Generalizing from cases like pain, the standard view is that consciousness evolved because it conferred its bearers an adaptive advantage.

However it is no easy task to show that consciousness is an adaptation in the strict sense, that is, that it was originally selected for because it increased the fitness of its bearers (Fox Keller and Lloyd 1992). One reason is that consciousness is at once phenomenologically homogeneous and heterogeneous.

Considered at the coarsest grain, conscious states share the property of being experienced: all and only conscious mental states seem a certain way. Indeed only conscious mental states *seem* any way at all; without consciousness there is no subjective, phenomenological point of view.<sup>1</sup> This phenomenological unity of experience distinguishes conscious states from non-conscious states. Exam-

ined more closely, however, conscious mental states vary widely. Experiences of red differ from experiences of green, experiences of colour differ from auditory and olfactory experiences, and so forth. This heterogeneity or variety of conscious mental states has caused some philosophers to wonder whether there is any single phenomenon, *consciousness*, after all (Churchland 1983).

We recognize consciousness under its phenomenal descriptions; conscious states share the Nagel-property: there is something that it is like to have a conscious state (Nagel 1974). But we do not know how to describe it in non-phenomenological terms. Is phenomenal consciousness a single trait, or a host of related traits? That is, are conscious mental states realized in one way in the brain, or in many ways?

### 1.1 Consciousness: Unity and variety

There is the consciousness in the sensory modalities; there are emotions, moods, dreams, and conscious propositional attitude states; there are various kinds of neuroses and psychoses. All of these are kinds of conscious states. This is not a final description or taxonomy of consciousness. But one must start by picking out the phenomena to be explained and these are ways we point at the phenomena.

If we knew how consciousness was realized in the brain we could give a neural specification. There is important work going on to investigate whether there is some such neural property and what it might be that the states we call 'conscious' have in common.

One possibility is that all conscious mental states are realized by a single neural property, say 40Hz oscillations (Crick and Koch 1990), or recurrency or reentry (Churchland 1995; Edelman 1989). Supposing this were so, then it might be the case that consciousness arose when human brains settled on a certain oscillation frequency or on a certain functional architecture. Settling on the brain states that give rise to consciousness might have been an adaptation or it might have been an evolutionary accident. Even if consciousness was an adaptation, it would not follow that all manifestations of consciousness, paranoia or dreams, for example, were themselves things Mother Nature aimed to be experienced because they were fitness enhancing.

The abilities to walk and run are likely adaptations. Being able to walk and run enable us to be able to waltz, and tango, and tap dance, and pole vault. But Mother Nature did not give a hoot about these bonuses. The ability to tango is not an adaptation. Suppose that, considered from both the phenomenological

and neuroscientific points of view, consciousness is a general trait with a common underlying neural feature or set of features that was selected for; it would not follow that all the varieties of consciousness, all the manifestations of consciousness, are adaptations.<sup>2</sup>

Another possibility is that the underpinnings of consciousness are as various as the phenomenology. Perhaps, considered neurophysiologically, consciousness is a disunified phenomenon — an array of processes that are similar only in that they happen to be states that have phenomenal properties. It might then be the case that each conscious process, each kind of consciousness, independently came to be. Some kinds might be adaptations, others neutral free-riders; still others might be exaptations, traits that were not initially selected for but that were later co-opted for their adaptive advantage. There might then be no one answer to the question of the adaptive advantage of consciousness. But it might also be that despite the variety in their instantiation, conscious states were all independently selected for the same reasons; that is, that the having of phenomenal properties, however realized, always confers the same sort of advantage to its bearers. The question of the adaptive advantage of consciousness might then be like the question of the adaptive advantage of camouflage: no sensible person thinks there is a single trait, camouflage, that some creatures have. The ability to camouflage one's body can be achieved by any of many heterogeneous physiologies. Nevertheless, it seems that sensible things can be said about the adaptive advantage of camouflage in general.

The phenomena of consciousness may be realized by a single neurophysiological process that manifests itself in many ways; or the variegated phenomenology of consciousness may be realized in similarly diverse physiology. Whether the physiological realizations of consciousness turn out to be more or less heterogeneous than their phenomenological manifestations will surely be relevant to questions about the adaptive advantage of consciousness. But the relation between these questions runs in both directions: Evolution makes traits; it may be that we cannot determine the answers to questions about the homogeneity or heterogeneity of consciousness prior to answering questions about its evolutionary history. That is, whether phenomenologically distinguishable states will count as one biological trait or many can depend on whether they are the result of one selection process or many.

Such complications do not spell doom for the project of providing ideally complete adaptation explanations of consciousness. But they caution against the glib assumption that the explanatory project is easy. One should not assume that all varieties of consciousness will be found to have etiological

functions, and one cannot assume from the fact that a kind of consciousness is currently adaptive that it was selected for.

## 1.2 Evolution by natural selection

The following fable is a possible account of the evolution of consciousness: In a finite population of interbreeding organisms, random mutation caused a portion of the population to have some sort of conscious states (i.e., for those states, there is something that it is like for the organism to be in that state.) In each case, the new phenotypic trait (speaking generally, *consciousness*) was heritable. Sadly, a nearby volcano erupted. By chance, the eruption killed all and only the non-conscious organisms. The conscious organisms, however, survived and reproduced successfully, passing on the trait — consciousness. Consciousness evolved.

Evolution occurred because there was phenotypic variation, heritability, and differential reproduction. This is an evolutionary explanation. Does it show that consciousness evolved because it was an adaptation? No. Although *evolution* of consciousness occurred in this case, it was not evolution by *natural selection* but rather by *random drift*. Only by chance did the conscious organisms out-reproduce their non-conscious counterparts; it was not *because* they were conscious that they survived. Evolution by natural selection — adaptation — requires a further element: there must be a cause other than chance for the differential reproduction that leads to evolution (Brandon 1990: 6–9). There must be something about a trait that accounts for the relative advantage of its bearer in a selective environment.

In order to give an adaptationist explanation of consciousness, we need to specify what the adaptive advantage of the feature in question was for a particular type of organism in a particular selective environment. We need to know that the trait has an etiological function. Etiological notions of function are the most common way of thinking about functions among philosophers of biology these days. The idea behind the family of views is that the functions of a thing are those effects for which it was selected.<sup>3</sup> The details of the etiological notion are disputed, but it is helpful to see what one way of formulating it looks like:

It is the/a proper function of an item (*X*) of an organism (*O*) to do that which items of *X*'s type did to contribute to the inclusive fitness of *O*'s ancestors, and which caused the genotype, of which *X* is the phenotypic expression, to be selected by natural selection. (Neander 1991: 174)

Etiological functions are those that figure in explanations according to the theory of evolution by natural selection.<sup>4</sup> The etiological function of a trait is an effect that gave it an adaptive advantage. To claim that a trait has an etiological function is to claim that it is an adaptation (Amundson and Lauder 1994).

### 1.3 Dreams and other spandrels of the brain

There are many features of organisms that are not adaptations; human chins, for example (Gould and Lewontin 1978). Consciousness is a trait like any other; although it might be disappointing, it would hardly be surprising if some varieties of consciousness have no evolutionary function. Gould and Vrba call those traits of an organism that are not themselves adaptations but are byproducts of other traits that have been selected for 'spandrels' or, if they later come to be selected for, 'exaptations' (Gould and Vrba 1982). Such traits lack etiological function.

Dreams are a plausible candidate for a type of consciousness that lacks an etiological function. Dreams are simply the byproducts of brains doing the things that brains do during sleep (Flanagan 1992, 1995, 1996, 2000). Some brain activity that occurs during sleep is an adaptation. The phenomenal mentation that occurs, although it is an effect of those processes, is an evolutionary byproduct of those brain activities for which sleep was selected. Dreaming *qua* experience makes no difference to the inclusive genetic fitness of organisms that dream. The neurochemical processes going on in their brains while they are asleep, including those that cause dreams, *do* make a difference to inclusive genetic fitness; it is just that *dreams* make no difference.

In broad strokes the hypothesis goes like this. Sleeping has an elegant neurophysiological profile, exemplified by reliable changes in brain waves and in the release ratio of aminergic versus cholinergic neurochemicals. There is good evidence that what the brain is doing during different stages of sleep is implicated in cell repair, hormone adjustment, learning, and memory consolidation.

Dreaming during NREM sleep is rationally perseverative and relatively non-bizarre. A person might think that she did not sleep because she could not stop worrying about the exam tomorrow. In fact, she did sleep. NREM sleep is like being awake in many respects and it is easily confused with being awake. NREM mentation is what gets left over from a normal brain gone to sleep. If one were awake one would first worry about the exam and then study. Since the brain does not turn off one continues to worry, but, being asleep one

doesn't get up. The perseverative dream rut doesn't affect the brain's ability to get one into a hypometabolic state in which cell repair and hormone adjustment can take place.

If one sleeps eight hours, then during two of those hours, one's eyes are bolting around under the eyelids. This is REM sleep. Neurochemically the NREM to REM shift marks (roughly) the shift from labor devoted to cell reparation to labor devoted to memory consolidation and storage. The mechanisms required to turn off certain neurons and to turn on others cause waves that incidentally activate areas throughout the brain, especially in the visual areas. Some of these activations are experienced as "thoughts" and "sensations."

Suppose the conscious brain is *independently* prone to try to make sense of thoughts it has. If so, there is no surprise that it tries — and in part succeeds — to supply a coherent story line to the noise it generates while the system as a whole is doing what it is does during sleep.

If being an adaptation is having an etiological function, then we may call the denial that consciousness is an adaptation *etiological epiphenomenalism*. Dreams, according to the story above, are etiological epiphenomena. Dreams are the spandrels of sleep.

Etiological epiphenomenalism is an empirical claim. It claims that the presence of a certain type of consciousness has no adaptation explanation — there is no effect for which that type of consciousness was selected. One can be an etiological epiphenomenalist about specific types of consciousness, e.g., dreams, or one might be an etiological epiphenomenalist about consciousness generally. To say that dreams are etiological epiphenomena is to say that there is no effect of dreams for which they have been selected. Etiological epiphenomenalism about dreams does not draw into question the existence of dreams, or the causal role of dreams in, say, the project of self-knowledge. It just says that, as a matter of historical fact, having dreams is not a trait that was selected for by natural selection. Culture may come to select for dream interpretation. But it is unlikely that sleep activity itself, say, dopamine reuptake, is enhanced by dreaming things that can be interpreted as having certain significance rather than other, or by a population becoming virtuosos of dream interpreters.<sup>5</sup>

Defending the etiological epiphenomenalism of dreams does not commit one to any particular conclusion about other varieties of consciousness — least of all those that may be surreptitiously activated during dreaming. One advantage of adopting an approach that treats consciousness as an array of states that share the Nagel-property is that it allows space for the discovery that some sorts

are epiphenomenal, e.g., dreams, while other sorts may have etiological functions, e.g., visual perception.

## 2. Explaining the evolution of consciousness

Although it seems likely that some varieties of consciousness are adaptations, specifying what the adaptive advantage of a kind of consciousness *might be* is difficult. But it is child's play compared to finding the sort of evidence that would indicate that any such "how possibly" story reflects how some variety of consciousness *actually* gave an organism an adaptive advantage in a selective environment. This problem, the problem of establishing that a "how possibly" explanation is a "how actually" explanation, requires empirical data.<sup>6</sup> By specifying the adaptive advantage of a variety of consciousness, we give an *ecological* account of its relative adaptedness. But even if we can discover the adaptive advantage of some variety of consciousness, that is only one piece of an adaptationist explanation.

### 2.1 Ideal adaptation explanation

Robert Brandon (1990) formulates five elements for an ideally complete adaptation explanation:

1. evidence that selection has occurred,
2. an ecological explanation of relative adaptedness,
3. evidence that the traits in question are heritable,
4. information about population structure,
5. phylogenetic information about trait polarity.

These five elements figure in explanations in terms of evolution by natural selection.<sup>7</sup>

The second element of ideal adaptation explanation, the ecological explanation, is the one that most cognitive scientists and philosophers of mind focus on when discussing the evolution of some feature of mind — consciousness in the present case. Ecological explanations of relative adaptedness tell why some trait increased the fitness of its bearer in a particular selective environment. Such explanations describe the etiological function of that trait. But giving a plausible story that satisfies the demand for an ecological explanation of relative adaptedness is not by itself sufficient for giving an adaptation explanation. There are four other conditions that need to be satisfied.

Regarding the first element, if one is going to give a story about why consciousness was favored by natural selection, one needs some evidence that *selection* for consciousness has occurred. This is different from the demand for some evidence that the *evolution* of consciousness occurred. In the volcanic random drift story discussed above we have the evolution of consciousness but without selection for consciousness. Evolution by natural selection requires that the cross-generation change was due to some advantage conferred by the trait that was selected for.

Some sorts of evidence that would fit the bill would be fossil evidence, especially if such evidence involved fossils from competing groups of, say, hominids. It is sometimes thought that *Homo erectus* and *Homo sapiens* roamed the earth together. And it is widely thought that *Homo sapiens* were favored because they were more intelligent than these other hominids, allowing, for example, development of linguistic capacities. What might be the evidence that such selection occurred? Intelligence, it has been argued, is linked to encephalization, and language to specific cortical regions of larger hominid brains (Byrne 1995; Wills 1993; Nahmias 1997). According to this line of thought, the fossil evidence provides support for the idea that selection occurred because it shows increased ratio of brain size to body size in *Homo sapiens* compared to other hominids, as well as space for, e.g., Broca's and Wernicke's areas in the larger skull space of *Homo sapiens*.<sup>8</sup>

Similar difficulties arise for finding evidence of heritability, the third requirement for an ideally complete adaptation explanation. Since consciousness is fixed in our population, that is, we are *all* conscious whereas we are not all six feet tall, we can't observe selection for consciousness in the way we can for height. But there are variations in consciousness between persons which might give us the information we seek. Colour blindness may be a case where there is a heritable variation in the qualitative structure of conscious visual experience. Other congenital sensory deficits (blindness, deafness) also suggest the heritability of consciousness bearing traits.

The fourth element requires information about population structure. Information about the frequencies of different traits in the population is needed in order to determine whether selection is at work, rather than, e.g. random drift. In addition, some evolutionary models, for example, group selection models, refer directly to the frequency of a trait in a given population. The idea behind those models is that some traits are group traits, and are sensitive to population density.

The fifth element of an ideal adaptation explanation requires information

about trait polarity; that is, what evolved from what. We need evidence that non-conscious creatures evolved into conscious creatures, not vice versa (Brandon 1990: 165–174).

It should be quite clear that few if any ideal adaptation explanations can be given for any trait, much less for consciousness. The point of having an ideal model is not to satisfy it (though that would be nice) but to have a principled standard against which proposed explanations may be evaluated. Brandon's criteria for an ideal adaptation explanation is such a gauge.

Understanding and taking seriously what it would require to give an adaptationist account of consciousness makes it very clear why it's so hard. Such explanations are difficult to give for any trait. The task of giving an adaptation explanation for consciousness inherits those difficulties intrinsic to adaptationist explanation, and complicates them with all the philosophical and scientific problems attendant to consciousness. Nevertheless, if we are going to take consciousness seriously as a natural biological phenomenon then our explanations (adaptationist, mechanical, and otherwise) of consciousness will have to be measured by the same criteria applied to other biological phenomena (Polger and Flanagan 1999).

## 2.2 Necessity and natural selection

There is a confusion that arises in discussions of consciousness and necessity that we want to get clear about. Suppose there is some organism that performs function  $f$  by going into physical state  $p$ . Suppose further that  $p$  is a conscious state. That is, whatever the relationship between conscious states and physical states turns out to be (identity, supervenience, etc.),  $p$  has that relationship to conscious state  $c$ . Now, we can ask several sorts of questions about  $p$ . One kind of question is: Is it logically, metaphysically, or nomically necessary that any system in state  $p$  is thereby in conscious state  $c$ ? Another question is: Is it logically, metaphysically, or nomically necessary that  $f$  be accomplished consciously? These are metaphysical questions. A third question is: Is it logically, metaphysically, or nomically necessary that evolution produce  $p$ ? This is a historical question. Whatever one thinks the answer to the first two metaphysical questions is, the answer to questions of the third sort is: no.

It is a consequence of taking seriously the idea that consciousness is a natural phenomenon that we must treat it like any other naturally occurring trait of a living organism. The presumption is that those traits that are currently adaptive were formed by adaptation; the burden of proof is on the objector to

show otherwise. As Brandon (1990: vii) puts it, natural selection is “the only general and scientifically legitimate theory of adaptation.” Thus, given its apparent adaptedness, the null hypothesis must be that consciousness was selected for in the process of evolution by natural selection. And selection produces contingencies — traits that did not have to be. Even if there are certain configurations of matter (or matter\*, depending on your favorite theory) that are necessarily conscious states (*qua* the first sort of necessity distinguished above), and even if they are necessary for some capacity of the organism (*qua* the second sort of necessity), those configurations did not have to be realized. The answer to the question, “Why did consciousness come to be?” cannot be, “Evolution selected it because consciousness is necessary for learning and plasticity?”

This is not because consciousness doesn’t give us learning and plasticity — maybe it does. It is because answers like, “Because consciousness is necessary for *x*” are of the wrong form. “*T* is necessary for *x*” is not the form of a proper adaptation explanation of *any* trait. If *T* is necessary for *x* then one does not need to appeal to evolutionary theory to explain the presence of *x* in *T*. Adaptation explanations do not explain why organisms have mass (presumably a necessary property of organisms) although they may explain why some organisms have the particular mass they do. Even if consciousness were in some metaphysical sense necessary for a capacity of an organism, that would not explain why it came to be. To think so would be to construe evolution as unduly forward-looking, teleological in a strong way not acceptable to most contemporary philosophers.<sup>9</sup>

This exposes the fundamental flaw in every proposal for an evolutionary explanation of consciousness that we have seen. We maintain that no credible adaptationist account of consciousness has been given (Flanagan and Polger 1995). We’re not just lamenting that there are no *complete* explanations of consciousness out there — there aren’t any that even approximate full adaptationist explanation. They are all explanations in terms of one or another cognitive function that consciousness is alleged to be, in some sense, necessary for.

### 2.3 Inessentialism and the adaptation question

Daniel Dennett, responding to our claim that no plausible adaptationist account of consciousness has been given, argues that it is a mistake to ask what the adaptive advantage of consciousness might be:

The question of adaptive advantage, however, is ill-posed in the first place. If consciousness is... not a single wonderful separable thing ('experiential sensitivity') but a huge complex of many different information capacities that individually arise for a wide variety of reasons, there is no reason to suppose that 'it' is something that stands in needs of its own separable status as fitness enhancing. It is not a separate organ or a separate medium or a separate talent.

To see the fallacy, consider the parallel question about what the adaptive advantage of *health* is. Consider 'health inessentialism': for any activity *b*, performed in any domain *d*, even if *we* need to be healthy to engage in it (e.g., pole vaulting, swimming the English Channel, climbing Mount Everest), it could in principle be engaged in by something that wasn't healthy at all. So what is health for? Such a mystery! (Dennett 1995: 324–325)

Dennett's parody misses its mark; "health inessentialism" (HI) is unmotivated. Whereas no-one holds that a human being — or an organism remotely like us — could climb Mount Everest without being healthy, many philosophers believe it is possible that there are creatures that could do all the things that we do, but without being conscious. This is the thesis of *conscious inessentialism*.

Conscious inessentialism (CI) is the view that "for any intelligent activity *i*, performed in any cognitive domain *d*, even if *we* do *i* with conscious accompaniments, *i* can in principle be done without these conscious accompaniments" (Flanagan 1992: 5).<sup>10</sup> CI became plausible in the wake of work in artificial intelligence. A system could be very smart but lack experience altogether. Here the contrast case involves a biological system such as a human being and an inorganic computer. There are in fact computers that play world-class chess and are not conscious; therefore, it is possible to play world-class chess without being conscious. Consciousness is not essential to the ability to play chess.

HI might seem plausible in a similar way. A bulldozer might be able to get up Mount Everest, but it is not healthy. Roughly speaking, therefore, one could get up Everest without being healthy. Dennett encourages us to think that just as consciousness is not essential for chess playing, so health is not essential for mountain climbing.

If CI and HI seem equally plausible *prima facie*, important differences reveal themselves upon reflection. CI is meant to apply to both the contrastive cases where organic creatures and inorganic ones are involved *and* where only biological organisms are involved. If we focus on organisms, the claim is that we can imagine experientially blank creatures relevantly like us behaving intelligently across most every domain, i.e., they will pass the toughest Turing-test and more.<sup>11</sup> They will do all the things that pre-theoretically we think

require conscious intelligence, but which upon reflection seem only to require intelligence.

In contrast, the concept of “health” doesn’t even apply to bulldozers. It’s true that we cannot imagine non-healthy organisms that are relevantly like us, i.e., with cardiovascular and muscle systems like ours, who can climb Mount Everest with weak muscles, lungs, and hearts. But HI doesn’t even apply to radically different systems.

One reason HI sounds absurd is this disanalogy. HI fails not because it construes health as a “single wonderful separable thing.” Rather, the very notion of *health* doesn’t apply to non-organisms.<sup>12</sup> Furthermore, *health* is conceptually intertwined with fitness, and thus with the very notion of adaptation, in a way that consciousness is not. Asking about the adaptive advantage of health is tantamount to asking about the adaptive advantage of fitness. Such a mystery, indeed!

Dennett’s health inessentialism parody does not succeed. For the reasons given, it would not succeed even if one thought that consciousness “were a single, separable, wonderful thing.” But Dennett knows better than to suggest that we think that! A major theme in our work in philosophy of mind is the idea that there are multiple kinds of consciousness when specified from the phenomenological point of view and probably the neurophysiological point of view as well (Flanagan, 1985, 1992, 2000; Flanagan and Polger 1995). We treat consciousness as a superordinate, rather than as a middle-level or subordinate, category. Consciousness is to conscious vision, and conscious vision to perception of red, as vehicle is to car, and car to Mustang. Consciousness is a kind that admits of complex taxonomization.

As we argued above, that some trait has many subvariants, that its name covers a “huge complex” of traits even within a species, does not render foolish the question of its adaptive advantage. It just means that there has to be an answer for each instantiation — possibly not the same answer.

Sometimes the fact that a trait admits of variations helps explain the trait’s existence, as is the case of butterfly wing patterns. There are various wing patterns within individual butterfly species; but this is not by itself sufficient reason to think that it is a mistake to seek an adaptationist explanation of butterfly wing patterns in general. Moreover, the evidence about the frequency, morphology, and similarity of variants, far from undermining an adaptationist explanation, provide crucial clues for theses about wing pattern evolution and development. Recognizing that differing wing patterns are related is crucial to understanding their evolution. To look for independent

evidence for the evolution of each individual wing pattern could be to fail to recognize an important connection. Indeed, in this case it turns out that there is a common developmental mechanism that accounts for the phenotypic variation in butterfly wing patterns (Nijhout 1990). But this discovery *followed* recognition of the adaptive advantage of not only particular wing patterns, but of variation in wing pattern. The presence of phenotypic variation is not itself an obstacle to adaptationist query. Although the variation in butterfly wing patterns turns out to result from a single underlying process, the mere fact of variation does not render senseless the question of the adaptive advantage of wing pattern unless one thinks that the question presupposes a unified answer. But that cannot be the case, since the answer (or answers) to the adaptation question may figure in deciding whether some feature is to be counted as one trait or many.

Asking if consciousness is an adaptation does not depend on knowing ahead of time whether the phenomenal varieties of consciousness will turn out to be neurophysiologically homogeneous or heterogeneous phenomena. Dennett claims that various sorts of consciousness “individually arise for a wide variety of reasons.” Maybe yes, and maybe no. Our approach is neutral on this question. When we ask about the adaptive advantage of consciousness to an organism, we are asking for some of those reasons to be specified for some type of conscious state — whether the mechanisms of consciousness turn out to be one or many. We are asking what etiological functions, if any, the various kinds of phenomenal consciousness have in, say, *Homo sapiens*. It may be that consciousness was selected as a single phenomenological and brain trait — the product of recurrency or reentry (Churchland 1995; Edelman 1989). If this was so, consciousness as specified at the superordinate level would have an etiological function, but some, possibly many of its lower-level types, or varieties, might not have etiological functions.<sup>13</sup> Some kinds of consciousness might be etiological epiphenomena.

### 3. Varieties of epiphenomenalism

The etiological variety of epiphenomenalism is not the only one, and it is not the one that most often figures in the philosophy of mind. We now turn our attention to three ways in which consciousness might be thought of as epiphenomenal:

- i. Etiological epiphenomenalism: consciousness depends on the physical and has physical effects, but those effects are not adaptations.
- ii. Causal-role epiphenomenalism: consciousness depends on the physical and itself has physical effects, but those effects are not “mechanistic functions” they play no important causal role in the organismic system.
- iii. Strict metaphysical epiphenomenalism: consciousness depends on the physical, but cannot have physical effects, period.<sup>14</sup>

We’ve seen what etiological epiphenomenalism involves. What about the other varieties of epiphenomenalism?

### 3.1 Causal role epiphenomenalism

Physical systems have causal powers, and causal effects. Some effects of a system, but usually not all, are functions of the system. Robert Cummins has given a formalization of this notion of function:

$x$  functions as a  $\phi$  in  $s$  (or: the function of  $x$  in  $s$  is to  $\phi$ ) relative to an analytical account  $A$  of  $s$ ’s capacity to  $\psi$  just in case  $x$  is capable of  $\phi$ -ing in  $s$  and  $A$  appropriately and adequately accounts for  $s$ ’s capacity to  $\psi$  by, in part, appealing to the capacity of  $x$  to  $\phi$  in  $s$ . (Cummins 1975: 762)

Functions are the effects of something that play a causal role in an explanation of an overall capacity of a containing system. Capacities of a containing system are explained in terms of capacities of components of the system in Cummins’ “functional analysis” model of explanation. On this view, functions are those capacities that are appealed to in such an explanation. Cummins-style functions are always ascribed “against the background of a containing system” (Cummins 1975: 763). Amundson and Lauder (1994) follow Neander (1991) in calling Cummins-style functions *causal role* functions.

Epiphenomenalism of the second sort denies that consciousness plays a causal role function in the explanation of human capacities, so we will call it *causal role epiphenomenalism*. Causal role epiphenomenalism is the sort that applies to the noise that an automobile engine makes: engine noise is a physical effect of a physical system, but it plays no role in a mechanistic explanation of how that system operates. It plays no explanatory role because it is not a mechanism in the functioning of the engine as an engine. Of course it may well play a role for *some* system — it might, say, function as a mechanism (or part of a mechanism) for getting an infant to fall asleep. Consider that the Harley-Davidson motorcycle company has filed to trademark the noise that its motor-

cycles make precisely because other companies are designing their cycles to mimic the distinctive Harley sound. It seems that engine noise can help sell a car or motorcycle. Still, although the noise may play a causal role in the motorcycle-selling system, it plays no causal role in the operation of the vehicle.<sup>15</sup>

### 3.2 Strict metaphysical epiphenomenalism

*Strict metaphysical epiphenomenalism* is the sort of epiphenomenalism traditionally discussed by philosophers of mind, and historically associated with property dualism (the idea that mental properties are non-physical properties of physical states.) Strict metaphysical epiphenomenalism, as the name suggests, is a metaphysical doctrine. It is a claim about a kind of state and about the properties (in particular, the lack of causal properties) of that kind of state. Strict metaphysical epiphenomenalism, as Paul Churchland puts it, is the view that mental phenomena “are entirely impotent with respect to causal effects on the physical world” (1990: 11).

When philosophers want to explain strict metaphysical epiphenomenalism, they almost invariably do so by analogy to common cases of causal role epiphenomenalism. They say things like: strict metaphysical epiphenomenalism is the view that mental states are like the sound that the whistle on a steam train makes, or like the thumping sounds that hearts make as they pump blood. Thumping plays no causal role function in hearts; and thumping of hearts plays no causal role function in explanations of human biological capacities.

Strict metaphysical epiphenomenalism need not claim that it is *logically* impossible for non-physical entities to have causal efficacy in the physical world. Rather, given the physical laws of our universe it is *nomologically* or *naturally* impossible. In contrast (i) and (ii) are claims about the actual evolutionary history of kinds of organisms, and about their actual organismic mechanisms, respectively. Strict metaphysical epiphenomenalism is a claim not about what consciousness does or does not *actually* do, but rather about what consciousness *can* and *cannot* do. It is a claim about what is naturally possible — that it is nomologically impossible that consciousness have any physical effects. That is what makes strict metaphysical epiphenomenalism the spooky sort of epiphenomenalism. It is the strange idea that there is something that is itself caused but which can have no effects at all.

#### 4. Consequences of epiphenomenalism

Although we have raised worries about providing an account of the etiological function of consciousness, we have established nothing about the causal powers of consciousness or even about the possibility of an account that assigns etiological functions to some kinds of consciousness. We have not argued for “epiphenomenalism” as it is usually discussed in the philosophy of mind.

Indeed, even if consciousness is a spandrel, an accidental feature of we mutants who were lucky enough to have been missed by the lava from a volcanic eruption, nothing has been said, let alone established, about its causal role or even its current adaptedness. That some varieties of consciousness are not adaptations tells us precious little about the nature and causal efficacy of consciousness in general. The bridge of the nose is not, as Dr. Pangloss thought, an adaptation for supporting spectacles; but that fact makes it no less causally able to do so.

This point is simple but often missed. No sensible person thinks that the abilities to do calculus or quantum physics are the direct result of selection pressures, although they may well be the by-products of selection for traits needed to get around in the world. Abilities to do calculus or quantum physics are, like dreams, good candidates for etiological epiphenomena. The trouble arises because debates over the function of consciousness have often confused several notions of function, and thus several kinds of epiphenomenalism.

As research on the nature, evolution, and adaptedness of consciousness proceeds, different types of epiphenomenalism will need to be kept apart. One reason is that an entity can lack some kinds of function under a certain description but possess one or more functions under other descriptions. It can be an etiological epiphenomenon without being epiphenomenal in the causal role sense, much less in the strict metaphysical sense. Moreover, some organisms have etiological functions that they never, or no longer, perform. “The tree is dead. I keep it because I can climb it in autumn and clear the gutters. The kids love it because of the woodpeckers. When they go to college and my knees give out, I’ll hire a gutter cleaner who has a ladder. Then I’ll cut the tree down and use it for firewood.”

In a case such as the dead tree, its branches no longer perform some or all of their etiological functions. But functions come in different kinds. Functional status is interest relative; for each effect designated as a function, there may be a variety of effects that are less interesting, considered side-effects for some purpose. The kind of function one assigns primacy to will depend on one’s

interests and purposes. Perhaps distributing foliage to allow increased exposure to sunlight is one function of tree branches, and one of a kind that is important to explaining why the branches evolved. Other functions, ladder-to-roof, woodpecker-attractor, etc., are secondary to the project of explaining how tree branches came to be, but are useful for other purposes. That the branches of a dead tree no longer perform their etiological function says nothing about the other sorts of functions they can serve.

Epiphenomenalism is also interest relative. 'Epiphenomenalism' is just the philosopher's epithet for lowly functional status. If the branches never had an evolutionary function, if they were etiological epiphenomena, one would not conclude that they had no causal powers at all. So too, even if consciousness in general were a spandrel, this would do nothing to defeat our beliefs about the nature and causal powers of consciousness.

## Acknowledgments

We would like to thank audiences in Tucson, Arizona and Atlanta, Georgia for their helpful questions. Special thanks to commentator Todd Grantham. Valerie Hardcastle provided some useful suggestions. Robert Brandon and Carla Fehr helped us work out our ideas about functions and epiphenomenalism, and commented on an early draft of what has become, among other things, this essay. Francis Crick and Christof Koch pressed especially hard their intuitions about the unity and function of consciousness. We appreciate the help of our friends and critics; it is doubtful that we have satisfied all of their concerns.

## Notes

1. See Graham and Horgan (this volume) on the question of grain.
2. Francis Crick has encouraged us to consider whether consciousness is realized as a unified phenomenon (40 hertz oscillations.) He regards it is obvious that information need to be bound together in order for action to be generated, and that consciousness supervenes on this binding; so both the unity of consciousness and its function are intuitively clear to Crick. We remain unconvinced; and we certainly disagree with Crick's claim that the answer is obvious. But whereas we believe that conscious states are phenomenologically diverse, our intent is to remain neutral on the issue of the unity or disunity of the subvenient base(s) of consciousness. We further maintain that the fact, if it is a fact, of the neurophysiological unity of consciousness does not entail that there is *one* answer to the adaptation question with respect to all the manifestations of consciousness.
3. The etiological account is customarily traced to Wright (1973) but has received its most widely discussed treatments from Millikan (1989) and Neander (1991). Variations of the

etiological account, broadly construed, have been endorsed by philosophers of biology (e.g., Brandon 1990; Kitcher 1993; Godfrey-Smith 1994; Sober 1985) as well as philosophers of mind (see especially Millikan 1993 and Lycan 1987, 1996).

4. Amundson and Lauder (1994) call this formulation the *selected effect* account of function to distinguish it from more broadly construed etiological accounts. For example, Wright formulates under a more general account of function that also accommodates artifacts. Millikan (1989) calls her version “proper function.” Some philosophers have argued that etiological functions can be subsumed under causal role functions (e.g., Griffiths 1993; Davies 2000). We shall use the narrow, evolutionary, notion of etiological functions.

5. We could put the main point this way: The effects of dreams are not functions relative to the nervous system, as dopamine reuptake is. Rather they are effects relative to the whole-person system. If those individuals who engage in dream interpretation are able to improve their self-understanding, and if this aids them in carrying on with their lives, and this leads to differential reproductive success (and is heritable) then it could be that there *are* selective pressures for dreaming, and for recalling and interpreting dreams. Though dreaming may not have been selected for in the past, it may come to be selected for in the future. Dreams could come to have an etiological function. Unlikely perhaps; but possible.

6. The notions of “how possibly” and “how actually” explanations are explicated in detail in Brandon (1990) §5.3.

7. These five elements are listed in the order that Brandon presents them, which is not to indicate any relative importance. He argues that the relative importance of the various elements will vary on a case by case basis. For example, Brandon (in conversation) has suggested that trait polarity is not an issue in the case of consciousness.

8. Some researchers worry that even if we can show that selection occurred for general intelligence and language, nothing has been established about consciousness. There are, one might point out, no well-established background theories about what sorts of evidence might be signs of consciousness as there are for intelligence and language. Perhaps that is an accurate assessment of the present state of the science of consciousness, but it is likely to change as attention is focussed on the evolutionary questions concerning consciousness. One researcher (Fink 1996) suggests that evidence of consciousness may be found in the complexity of pharynx, the “energy intake hub.” The argument is based on the idea that the brain structures required to support consciousness are energy hungry and thus require particularly large air intake. This proposal, however speculative, illustrates that the search for evidence of consciousness need not be restricted to the relative size of fossilized skulls. Our point is not that encephalization or pharynx data is the answer, but merely that the evolutionary project in consciousness studies is not in principle obstructed by the lack of soft tissue evidence, much less the by impossibility of well fossilized phenomenal experiences.

9. Todd Grantham pressed us to distinguish this objection from a related but weaker point based on the thesis of conscious inessentialism (see §2.3 of this essay.)

10. Note that this formulation makes a claim about ‘activity’ not ‘acts’ as defined by action theory. ‘Activities’ in the sense relevant to conscious inessentialism do not involve conscious intentions to act. See Flanagan (1992: 129–30). The thesis of conscious inessentialism

is just that — a thesis. It is an attempt to articulate the intuition that lies behind the consensus that the issues of intelligence and consciousness can, in principle, be pried apart. Flanagan (1992), in fact, denies that *Homo sapiens* could do what they do without consciousness. It is a matter of *a posteriori* necessity that certain actions we perform, e.g., lying, require conscious motivation. On this view, a non-conscious system might mislead but it cannot lie.

11. See Flanagan and Polger 1995, and other discussions of so-called zombies.

12. Some philosopher would argue that consciousness, like health, is a concept that does not apply to non-organisms. Dennett is not one of those philosophers; so we shall leave this objection aside.

13. An example of superordinate traits that could be adaptive without particular subordinate types also being adaptive would human eye colour. It seems at least plausible that colouration of the iris could be an adaptation without any particular iris pigment being selected for over others. Thus, eye colour could be an adaptation, but not blue eye colour, *per se*. We have no idea whether this is true of human eye colour, but it illustrates the possibility.

14. There appears to be at least one instance of an interesting variation on type (iii) in the literature (Chalmers 1996):

- iii\*. *Limited-exception metaphysical epiphenomenalism*: consciousness is an effect of the physical, but it cannot have physical effects *except* that it causes us to report having conscious experience.

We call this the “limited exception” version of metaphysical epiphenomenalism; it stipulates an exception in order to explain the obvious fact that we often report being or having been conscious. We are going to ignore (iii\*), or at least assume that it must be assimilated into (ii) or (iii), because we cannot see that (iii\*) is a sustainable position. The conservation laws do not admit of limited exceptions.

15. Notice that the claim is merely that engine noise does not in fact play such a role, not that it could not. There might be some sort of engine or vehicle in which the engine noise does play a causal role. An example from household appliances and the ever tricky business of microwaving popcorn illustrates the possibility: Some microwave ovens have sensors that detect the sound of popcorn popping and, based on the decreasing frequency of pops, automatically stop the process to prevent burning the popcorn. The sound of popping corn plays a causal role function for those ovens, though not in the popping of individual kernels of corn.

## References

- Amundson, R. and Lauder, G. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and philosophy* 9: 443–469.
- Brandon, R. 1990. *Adaptation and environment*. Princeton: Princeton University Press.
- Byrne, R. 1995. *The thinking ape: Evolutionary origins of intelligence*. New York: Oxford.
- Chalmers, D. 1996. *The conscious mind*. New York: Oxford.

- Churchland, P. S. 1983. Consciousness: The transmutation of a concept. *Pacific philosophical quarterly*, 64: 80–95.
- Churchland, P. M. 1988. *Matter and consciousness*, revised edition. Cambridge, MA: MIT Press.
- Churchland, P. M. 1995. *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA: MIT Press.
- Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the neurosciences*, 2: 263–275.
- Cummins, R. 1975. Functional analysis. *The Journal of Philosophy* LXXII, 20: 741–765.
- Davies, P. 2000. The nature of natural norms: Why selected functions are systemic capacity functions. No. 34, 1: 85–107.
- Dennett, D. 1995. The unimagined preposterousness of zombies. *Journal of consciousness studies*, 2, 4: 322–26.
- Edelman, G. 1989. *The Remembered present: A biological theory of consciousness*. New York: Basic Books.
- Fink, B. 1996. The evolution of conscious behavior: An energy analysis. Presentation at Towards a science of consciousness II, Tucson, Arizona, April 1996.
- Fox Keller, E. and Lloyd, E. (eds.) 1992. *Keywords in evolutionary biology*. Cambridge, MA: Harvard University Press.
- Flanagan, O. 1985. Consciousness, naturalism, and Nagel. *Journal of mind and behavior*, 6, 3: 373–390.
- Flanagan, O. 1992. *Consciousness reconsidered*. Cambridge, MA: MIT Press.
- Flanagan, O. 1995. Deconstructing dreams: The spandrels of sleep. *Journal of philosophy*: 5–27. Reprinted with revisions in Flanagan 1996: 32–52.
- Flanagan, O. 1996. *Self expressions: Mind, morals, and the meaning of life*. New York: Oxford University Press.
- Flanagan, O. 2000. *Dreaming souls: Sleep, dreams and the evolution of the conscious mind*. New York: Oxford University Press.
- Flanagan, O. and Polger, T. 1995. Zombies and the function of consciousness. *Journal of consciousness studies*, 2, 4: 313–321.
- Godfrey-Smith, P. 1994. A Modern history theory of functions. *Noûs* 28,3: 344–362.
- Gould, S. J. and Lewontin, R. 1978. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist program. *Proceedings of the Royal Society, London* 205: 581–598.
- Gould, S. J. and Vrba, E. 1982. Exaptation: A missing term in the science of form. *Paleobiology*, 3: 115–151.
- Griffiths, P. 1993. Functional analysis and proper functions. *The British journal of philosophy of science* 44: 409–422.
- Kitcher, P. 1993. Function and design. In P. A. French, T. E. Uehling, Jr., H. K. Wettstein (Eds.), *Midwest studies in philosophy XVIII*. Notre Dame, Ind.: University of Notre Dame Press, 379–397.
- Lycan, W. 1987. *Consciousness*. Cambridge, MA: MIT Press.
- Lycan, W. 1996. *Consciousness and experience*. Cambridge, MA: MIT Press.

- Millikan, R. 1989. In defense of proper functions. *Philosophy of science* 56: 288–302. Reprinted in Millikan 1993: 13–29.
- Millikan, R. 1993. *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical review*, LXXXIII: 435–450.
- Nahmias, E. 1997. Why our brains got so big: Reciprocal altruism deception, and theory of mind. Presented at the Southern Society for Philosophy and Psychology, Atlanta, Georgia, April 1997.
- Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.
- Nijhout, H. 1990. A Comprehensive model for colour pattern formation in butterflies. *Proceedings of the Royal Society of London B* 239: 81–113.
- Polger, T. and Flanagan, O. 1999. Natural answers to natural questions. In V. Hardcastle (Ed.), *Where biology meets psychology* (Cambridge, MA: MIT Press).
- Sober, E. 1985. Panglossian functionalism and the philosophy of mind. *Synthese* 64: 165–193.
- Wills, C. 1993. *The runaway brain*. New York: Basic Books.
- Wright, L. 1973. Functions. *Philosophical review* 82: 139–168.



# The functions of consciousness

David Cole

University of Minnesota, Duluth

Animals move. Accordingly, animals need a system to control movement — an animal soul. The function of the nervous system is to make the movement of animals appropriate to their surroundings. The various activities of the nervous system in general conduce to this function. So at first blush, it is reasonable to hold that the overarching function of consciousness, an activity of the nervous system, is to increase the appropriateness of behavior, to be adaptive. This should be our guiding principle in looking at the various aspects of consciousness: as awareness, as self-awareness, and as states of subjective representation, or qualia.

It may be that the principle shall fail, and that consciousness, in at least one of its forms, such as qualia, will be found not to be adaptive and not to have a natural function. But if that aspect of consciousness is valued by humans — if humans are willing to modify their behavior in order to alter that aspect of consciousness, then that aspect expresses itself and can be selected for. So it seems likely that at least all valued aspects of consciousness will have a function.

## The functions of awareness

If one doubts that consciousness per se has a biological function, one may perform a simple (thought) experiment. Situate oneself in some moderately dangerous situation (rooftop, raft approaching falls, pedestrian on a busy highway) and self-administer an agent that will render one unconscious — say sodium pentathol. Impaired consciousness will diminish one's ability to survive. Complete lack of consciousness will greatly impair one's prospects. The specific liabilities resulting from lack of consciousness will vary widely — failure to maintain balance on rooftop, to navigate raft to shore, to avoid oncoming traffic, to find food, shelter, mate. So consciousness is important for

survival in myriad ways, reflecting the tremendous range of obstacles, hazards, and opportunities that our environments afford us.

And so it seems clear that consciousness *per se* has many functions. However, consciousness, like cognition generally, and also perception, is not a single phenomenon. Philosophers have pointed out that there is more than one aspect to consciousness (or, alternatively, more than one sense of “consciousness”). In the ordinary sense, consciousness is being awake, aware. This state admits of degrees, and being fully conscious is being alert. Thus “conscious” functions more or less as a generic perceptual state (and indeed Leibniz uses “perceive” for being conscious.) David Rosenthal calls this general awareness form of consciousness “creature consciousness”; Ned Block and Gulven Guzeldere also distinguish this form of consciousness from other forms. Such consciousness can, but need not, have an object. One can be said to be conscious of a bird, or one can just be conscious. Animals generally are capable of awareness in this sense, that is, they sense objects and events in their vicinity, and states of their surroundings.

### Metaconsciousness

On the other hand, as Rosenthal points out, in addition to consciousness as wakeful awareness of things around one, there are conscious psychological states, as when I have a conscious belief rather than an unconscious one. When I have a conscious belief, I am aware that I have the belief. A mental state is conscious in this sense if one is aware of the state, that is, one is aware of one’s having the state. A psychological state then is a conscious state if it is the object of higher order consciousness, or metaconsciousness. How many states are conscious is a matter for speculation, and the current mood is that the proportion of mental states that are conscious is far fewer than was thought by Descartes and many of his successors. For example, Sir William Hamilton held that “In the simplest act of perception I am conscious of myself as the perceiving subject and of the external reality as the object perceived” (cited in W. R. Sorley 1920/1965: 245) While we may be rightly wary of Hamilton’s sanguine assurance that such awareness is universal in the “simplest act” of perception — relying on vision to avoid the leftside door jamb as one heads out, e.g. — nevertheless it is the case that we *can* be aware of ourselves as the perceiving subject.

It is much more problematic to ascribe this form of consciousness to animals generally, as there is no evidence that they have representations of their own first-order psychological states (see, e.g., Carruthers 1996). Below I will discuss the functions of this form of consciousness, paying particular attention to two cases that are of interest both in assessing the function of metaconsciousness, and in accounting for qualia. Blindsight constitutes a form of sensing unaccompanied by awareness that one is sensing. And David Armstrong's example of a truck driver, who drives "automatically" while thinking about something else, constitutes a more mundane example of awareness (creature consciousness) without self-awareness, that is, without consciousness of the state of awareness.

The finding of many functions for consciousness that we began with applies to the various forms of creature consciousness, but not without additional argument to second order consciousness. While there is obvious survival value to awareness of objects and events in one's surroundings, it is not nearly so obvious that there are like advantages to awareness of one's own internal psychological states, including first order states of awareness. Is there a function for second order awareness? There are reasons for thinking higher-order consciousness has several functions.

Metaconsciousness, though just one of the aspects of consciousness, may itself have several functions. Intelligent social animals have special epistemic needs. It is important to know what someone else believes. Explicit belief ascription requires the ability to use representations of the form of "X believes that p." But this form is productive, and will accept not only second and third person subjects, but also first person. Thus the required ability, representation of the beliefs of others, carries with it the capacity to represent one's own beliefs. As one can come to know the beliefs of others, so one can come to explicitly know one's own beliefs. Furthermore, social interaction will likely require one to compare one's beliefs with those of others, and to provide justification for one's beliefs and one's interpretations of perceptual data. With a capacity for metaconsciousness, one can also explicitly represent one's own goals, and consider means of fulfilling them. These plans can be communicated, promoted, evaluated, improved. They can be coordinated with others, and the help of others can be enlisted. Thus second order representations of beliefs, goals and desires serve important functions. They operate by making the first order beliefs and desires conscious, available for scrutiny by oneself and others.

Thus there are several ways in which attributions of mental states have an important and wide-ranging value for intelligent social creatures. It is valuable to be able to construct a model of minds generally. He knows this, believes that, can see this, is listening to us but can't make out our words. Being able to represent what other creatures know, believe, and perceive, is extremely useful, whether those creatures are of one's own kind or not, and whether they are predator, prey, competitor, cooperator, offspring or mate. The straightforward extension of the ability to represent mental states generally to represent one's own states allows comparisons — he believes that, I believe this; it seems that way to him, this way to me. This may allow us to predict how his behavior will diverge from one's own. More importantly, it will allow one to seek causes for the differences — why do our beliefs differ? Can he see what I see? What does he know that I don't? Critical self-evaluation becomes possible.

Indeed at least one level of representation higher than this may also be useful, though perhaps not as useful as second order metaconsciousness. Surely there will be times at which it may be useful for me to represent the way I represent him. "What do you think of his idea?" makes us self-conscious of our own representings of others. And we may wonder what she will think of us for representing him as we do. In general, metaconsciousness allows us to apply norms to our own cognitive states and processes (cf. Nozick 1994). It also allows us to become aware of causes and effects of our cognitive states. Once I am aware of how I think about a certain thing, I can ask what makes me think so. I can also recognize and track the causes of my feelings — why do I like turtles? And effects may become apparent as well — I can ask what effect my dislike of fundamentalist Christians is having on my behavior, and how this might be affecting my children, my friends, or my students. This may permit planning — given my fear of heights, how can I clean the gutters without climbing a ladder?

So here are three central functions of metaconsciousness:

1. metaconsciousness enables us to apply norms to our cognitive and affective states, and
2. makes it possible for us to do autopsychanalysis, discovering the causes of our psychological states, and
3. allows us to plan in such away as to maximize satisfactions over dissatisfactions.

In addition to awareness of ones doxastic and telic states, there also is awareness of one's perceptual states (and one's mental imagery). One can be aware of

being appeared to in such and such a way. In the doxastic cases discussed above, both the metarepresentation and its object are propositional. In the latter case, it is not so clear what form the metarepresentation takes, and there are advocates of propositional second order representations (e.g. Rosenthal 1990) and also advocates of imagistic or perceptual second order representations (e.g. Lycan 1996). In the end, the difference may not be as clear cut as it appears at first glance, for there are arguments (e.g. Pinker 1999) that images per se are too ambiguous for thought, and so must involve propositional labels (Pylyshyn and Fodor have also been critics of images as vehicles of thought). And others have argued that perceptual accounts of HOT collapse into propositional accounts (e.g. Guzeldere 1995).

Here my concern will be confined to the possible *functions* of Higher Order Representation (HOR) of perceptual states, no matter what form that HOR might take, imagistic or propositional. What might be the function of HOR of perceptual states? Many of the apparent functions of second order propositional states, such as of beliefs, depend upon the represented psychological states being propositional — justifications, communication, means/end analysis. On the face of it, these functions cannot apply to explicit awareness of non-propositional sensory states.

Yet there are several possibilities for how awareness of non-propositional states might have a function. For one thing, recognizing that something is novel in one's experience requires representation of one's experiencing. On the other hand, one may find it very useful to consider where one has seen, smelled, or heard such and such before. Here again, representations of one's perceiving are required.

An important lesson in a child's life is that auditors cannot always see what the child sees. This requires the ability to represent individuals as having certain perceptual experiences. Later in life, this ability to represent the extent and character of the representations of others will be useful in a variety of situations.

In addition, representations of perceptual states allow one to describe them, to talk about them, and to use linguistically framed rules of interpretation to analyze them. Not all objects can be identified immediately, or by low-level modules (see Kosslyn and Koenig for a discussion of levels of analysis in the visual and other systems). Advanced hypothesis testing, often involving shared norms and data from others, requires explicit representation of the data to be explained, consideration of how this or that aspect of the data might be explained on this or that hypothesis, closeness of fit, what one might

counterfactually expect the data to be on each hypothesis, and so forth. A key function of higher order representation of sensory data then appears to be very naturally understood as an ability permitting sophisticated hypothesis testing. This in turn allows for more sophisticated perception than automatic categorization and interpretation by modules with no or limited access to general information about the world.

## Qualia

The most difficult aspect of consciousness for which to discover a function are qualia. As near as I can tell, qualia are what would more ordinarily be called sensations, and akin to what philosophers used to call sense data. Qualia are the subjective, experiential, side of perception, and these subjective states can also occur in hallucination, dreams, and (vivid) imagination as well. Qualia presumably therefore play a key role in “what it is like” to be conscious, as opposed to merely sensing and responding appropriately to stimuli. The latter is something many machines do, from motion-detecting light switches, to robotic vision systems. But presumably in these systems, there is no subjective experience, no sensations, no qualia. Even though a robot may have propositional representations of its environment, updated on the basis of sensors, including cameras, intuitively one can account for all that robots do without attributing sensations or qualia to them. Qualia, then, are imagistic rather than propositional, representational (*qua* imagistic), and subjective.

There appears to be a consensus, among some philosophers at any rate, that qualia have no function, or at least that the specific character of any given qualia has no distinctive function that might not be performed just as well by quite different qualia. Two considerations prominently support such conclusions: the possibilities of inverted qualia, on the one hand, and of completely missing qualia on the other.

## The inverted spectrum

Formerly an oddity that haunted behaviorism and provided a puzzle about our knowledge of other minds, the inverted spectrum possibility has emerged as a major challenge to functionalism (joining forces in this role with Searle’s Chinese Room argument). According to the inverted spectrum hypothesis,

which goes back at least to Locke, the possibility is raised that two individuals might behave in exactly the same way yet have different qualia. The traditional example is of an individual who has a subjective color spectrum that is inverted with regard to that had by other individuals. Where most have the qualia standardly produced by red objects, this individual has the qualia standardly produced in others by blue objects. And when presented with a blue object, this individual experiences qualia that most persons experience only when presented with red objects. And so forth — the Invert's color spectrum is the inverse of normal, there are systematic inter-subjective differences in qualia. But the Invert can make all the color discriminations that a normal person can make — unlike those handicapped by color blindness. And since he learned color terms by being presented with objects and color samples in the shared physical world, his use of language and his other color related behavior agrees with normal individuals. Finally, his color proximity judgments will be the same as a normal person's. Thus this individual has an in principle undetectable condition — his nonstandard qualia cause no functional difference. And so it seems that the particular qualia he has — and by extension that any of us have — have no distinctive function. Perhaps it is important that the qualia we have when we are presented with red objects differ from those that we have when we are presented with blue objects, in order that we may distinguish them, but just what qualia we have does not seem to matter. Qualia are epi-functional.

But even this last limited concession to the role of qualia has been called into question. It appears that someone who had no qualia at all could behave exactly as those who do. A Zombie, in this context, would be a person who had no qualia, yet behaved as though he did. Such a person might have all the behavioral, dispositional, and functional characteristics of ordinary humans, yet be missing all qualia. I'll discuss this possibility, and the related problem of blindsight, below, after an examination of a recent influential form of the inverted spectrum argument.

## Inverted earth

An interesting recent elaboration of, and variation on, the inverted spectrum possibility is presented in Ned Block's (1990) "Inverted Earth". Inverted Earth is *just* like Earth, except that the colors of everything are all changed around — the color spectrum is inverted, or complementary colors on the color wheel are swapped, in just the way that appearances would be changed if one were

wearing spectrum inverting spectacles on earth. Grass is red, the sky is yellow, etc. In addition, on Inverted Earth the color vocabulary is also inverted — they call their yellow sky “blue”, their bright red grass “green”, and so forth.

Now suppose mad scientists render you unconscious, implant color-inverting lenses in your eyes, change your body pigment so that it will look normal to you upon awakening, and then transport you to Inverted Earth. When you wake on Inverted Earth, you notice NO difference. As Block says, “‘What it’s like’ for you to interact with the world and with other people does not change at all.” (683 in B F G). “So once 50 years have passed [during which time the “causal groundings” — the reference — of your color terms shift to those standard on inverted earth], you and your earlier state at home would exemplify...a case of functional and intentional inversion together with same qualitative contents — the converse of the inverted spectrum case. This is enough to refute the functionalist theory of qualitative content and at the same time to establish the intentional/qualitative distinction” (ibid). (Block presents a second version of the scenario involving identical twins separated at birth in order to address possible concerns about the change of intentional properties.)

### Criticism of inverted spectrum arguments

There have been various criticisms of inverted spectrum arguments, including the Inverted Earth version. Some concern the meaning of color terms, some deny the possibility of inverting the actual color spectrum experienced by normal humans while preserving all judgments of color similarity. These may be important, but, as Block suggests, there may be ways of meeting them, insofar as they turn on peculiarities of actual human color perception or physiology, by considering individuals that are like humans, except that they have a simpler color spectrum, or who are physiologically such that their spectrum can be inverted.

But I wish to press what I take to be a more basic objection. The Inverted Earth scenario, like any thought experiment, makes several assumptions. One of these, I believe, is questionable, and amounts to rejecting functionalism at the outset. Block supposes that the inhabitants of Twin Earth have different qualia than Earthlings, but that these have the same functional role as the corresponding qualia of Earthlings. For example, Twin Earthlings look at the sky, have a sensation like those an Earthling has when looking at something yellow, but then go on to call it Blue, and say that it is similar to the color of a

mountain lake, fountain pen inks, the background of the field of stars on the American flag, Paul Newman's eyes, and so forth. OK. But in addition, if functional identity is preserved, they must have all the same affective reactions. They must find this color as relaxing and soothing as an Earthling finds blue, they must judge that a deep blue color (that is, subjectively yellow) is darkish, shades easily off into black, is cool, and so forth. Now if the qualia they are experiencing and which prompt them to say "deep blue" are exactly the same as a saturated yellow produces in a normal Earthling, is it really plausible that this quite different qualitative experience could be to them just as an experience of saturated blue is to us? Or, if the experience of the qualia is just the same in *all* its associations and effects, is it plausible to believe the Twin Earthlings really have different qualia?

Block is assuming, as near as I can tell, that light of certain wavelengths can continue to always produce the same qualia in beings with human physiology, no matter how the functional roles of those qualia are changed. But that is just what is at issue! Furthermore, it does not seem to be a plausible assumption, when considering the affective aspects of color. Apparently even infants find different colors to have different affect, and to cause mood changes. Some colors have been found to be more agitating than others. These are important parts of the qualitative experience of color. It is entirely possible then that beings who are just like us, except raised on inverted earth, will not have the same functional roles for color qualia as we do.

Furthermore, it is possible that if say they were physiologically modified such that their qualia have exactly the same functional roles, in all their richness, as do the corresponding (inverted) qualia of Earthlings, the Twin Earthlings qualia would *thereby* be changed to become identical to those of Earthlings. At least some of the available empirical evidence suggests that when qualia are altered using artifices like inverting goggles, and then behavior gradually adapts to the changed sensations, the qualitative feel is changed by the adaptation, by the change in functional role. (For more on this possibility, linking it to real world experiments with inverting goggles by I. Kohler, G. Stratton and others, see Cole 1990, "Functionalism and Inverted Spectra".)

## Luminance inversion

Colors are the favorite qualia in philosophical arguments about qualia, functionalism, and consciousness. I suspect that this is because the functional roles

of different colors differ in very subtle ways, so subtle that they are easily overlooked, and so inversion arguments become tempting. But color is just one aspect of experience, even of visual experience. In (most) visual experience, there are a variety of colors, and we know that some, but not all, can be arrayed in a continuum, a spectrum (e.g. browns are not on the spectrum, and many of a box of 24 or more crayons may fail to be spectral colors).

Luminance or brightness is another property of visual images. Accordingly, we can imagine a person with inverted luminance: what looks bright and white to us appears dark and black to him, and vice versa. Now try to suppose that there is no functional difference between the experience of a person with normal vision, and one who has inverted luminance. A scene that appears bright to us will appear dark and murky to him — yet he will presumably report that things are well illuminated and distinct. How can this be possible? Where we report a scene as bright and cheerful, he will experience darkness, yet report the same. I do not know if this is flat out impossible, but surely there is reason to suspect that he cannot be having the very same qualitative experience that I have, when I survey a dark, murky, gloomy scene, if such qualia in him cause reports that the scene is bright, clear, distinct and cheery.

### **Pitch inversion**

Vision is not the only modality that might have inversions. Musical sounds have an experienced loudness and pitch. Taking pitch first, we can then suppose that there is an individual who has inverted pitch, relative to normal humans. What you and I hear as thundering bass, the pitch invert (call him “PI”) hears as a high pitched siren. What you and I hear as the high pitched screaming alarm of a smoke detector, PI hears as fog horn. Except, of course, PI’s use of language is not overtly any different than ours — when he has the subjective experience of high pitched squeaking, he calls it “thundering bass”, and when he experiences a low boom, he remarks on the shrillness. We all agree that tubas are capable of much lower notes than are piccolos.

Let us suppose then that Pitch Invert’s auditory qualia resulting from any sounds other than precisely mid-auditory-range are different from ours, but that his inverted qualia have exactly the same functional roles as ours. Is this possible? It must be if qualia are not determined by functional role. What evidence is there that qualia are not determined by functional role? Why, thought experiments involving twin planets and inverts.

However reflection suggests that there cannot be Pitch Inverts who are qualitatively distinct but functionally identical to ourselves. The first blush plausibility of the possibility of the inversion fades as we start to consider the richness of the experience. For one thing, on the face of it, PI will have quite different experiences of timbre. Timbre is a very important part of auditory experience; timbre is determined by harmonics. Since harmonics are multiples of a fundamental frequency, high pitched tones can have fewer audible harmonics than can low pitched tones. Toward the limits of the upper limit of auditory perception, sounds can have no audible harmonics at all. However, low frequency sounds can have very rich harmonic composition. We can discern a variety of timbres based on the same fundamental, and can experience the shift in timbre as this harmonic composition changes. For the invert, however, there can be no experience of harmonic composition for such notes. He cannot experience the change from sine to square to saw tooth waveform of a 100Hz note — these all involve the experience of harmonics.

Other acoustic phenomena are frequency dependent — the experience of beats is an example. Beats are produced when two simultaneous tones differ slightly in frequency. The number of beats per second is the (absolute) difference of the frequency of the two tones. Low frequency notes then will produce slow beats as the frequency is changed by a given percent, high frequency tones will produce rapid beating. Normal humans can discern up to about 7 distinct beats per second. It is not clear how invert's experience could be like a normal person's. As low frequency notes shift, the beats "emerge" as the tones change audibly in pitch. But high frequency notes will have beats emerge even when there is no audible difference in pitch — a feature exploited in tuning instruments, by relying on beats to get more accurate tuning than could be accomplished by relying on perceived pitch differences alone. The beats are determined by the absolute difference in frequency, but, as is typical of much perception, pitch perception is logarithmic. The beats will emerge from the shifting tones in a different way for a normal hearer and for a pitch invert. (Note: Another phenomenon tied to the absolute frequency of sound is the Doppler effect, but it is not clear to me how this might affect perception in an invert.)

More importantly, auditory experience has a kinesthetic aspect. High tones are associated with tightenings — of vocal cords, increased lung pressure, diaphragm tension, and so forth. Low notes are low in the throat, produced by relaxings, larger air flow, and so forth. These associations most likely permeate all experience of pitch, not just that of tones the perceiver is producing or contemplating producing. The vast richness of these myriad connections and

associations is an important part of what gives experience its character. The conscious brain is an unbelievably vast network, with connections across multiple areas, sensory and motor. Part of the experience of a sound may be shaped by subliminal experience of what is required, what it is like, to produce similar sounds. That experience will not be the same for PI. While it might be possible to attempt to avoid this objection by supposing the kinesthetic qualia to be inverted as well, the point is that qualia are not raw and simple, they are embedded in a vast network of connections and associations.

### **Inverted loudness**

Inverted subjective loudness presents similar problems. Although apparently a logical possibility, a subject who experienced low intensity sounds as very loud, and loud sounds as we experience whispers, seems hardly conceivable, past the initial bald statement of the logical possibility of such inversion. Silence, for the loudness invert, would be deafeningly loud, yet the subject, if functionally identical to us, would report that it was restful, an excellent opportunity for thinking things through, and perhaps even so quiet one could hear a pin drop. The supposition that the character of qualia is independent of functional role seems to make no sense here. And why should this not be a general feature of qualia, whether this dependence on functional role is immediately apparent or not?

### **Radical synesthesia and transesthesia**

Synesthesia is a condition where a perceiving subject experiences qualia typically associated with another sense modality. For example, heard sounds might be accompanied by an experience of flashing lights. Let us then imagine an extreme case: suppose that there could be a victim of synesthesia who had visual stimuli *instead of* auditory stimuli. For example, the synesthete might locate sound sources in a visual field, and different sounds might be experienced as varying colors and luminance.

This possibility of what I will call “radical synesthesia” is on a continuum of proposed subjective qualia differences, from two persons who differ slightly in how they experience the taste of a wine, at one end, to the alleged possibility of absent qualia, at the other. Let us suppose that the functional role of the visual

qualia in the hypothesized radical synesthete are exactly the same as those of auditory qualia in a normal human.

My purpose in raising the possibility, of course, is to cast doubt upon its reality, and thereby, upon its presuppositions. The radical synesthete will to be a real possibility if functional accounts of qualia are rejected (we can suppose that the synesthete differs in some non-functional way from normal humans). But I suggest that supposing that a radical synesthete is a real possibility is radically unattractive. Besides the immediate epistemic questions — how do you know that your loved ones are not radical synesthetes? how do you know that *you* are not such a synesthete? — there are the more interesting questions of how it could make sense that there be such divergent qualia that have no effect on (or are unaffected by) functional role.

Let us consider the two forms that radical synesthesia might take. First, as described above, the radical synesthete has normal visual experiences; just his auditory qualia are non-standard, and are similar to his (and our) visual qualia. But then surely such a synesthete will differ functionally from the rest of us — he will find audition to be very like vision, indeed, to be subjectively just a form of vision. Indeed, it is not clear how he could see and hear at the same time, since he would then be experiencing two sets of visual qualia in a (single? somehow doubled?) visual field. Not only is the very experience of such a synesthete difficult to understand, even if it were possible, his judgments about his experience would be non-standard — sounds would be experienced as visual, and so say loud sounds might be bright, high pitched sounds violet, and so forth. Even if such a person can successfully identify pitch in agreement with his normal peers, his judgments about the character of his experience would be very non-standard. This should not be a great surprise — synesthesia in a less radical form is a real condition, and has just such symptoms.

This form of radical synesthesia then would differ functionally from normal experience. Let us consider a second more philosophically interesting possible form of synesthesia then — a modality transposing radical synesthete, with all qualia swapped between two modalities, a condition we might call “transesthesia”. The vision-audition transesthete will “see” sounds and “hear” sights, that is, experience (only) visual qualia when hearing, and experience (only) auditory qualia when seeing. Let us suppose that the qualia have the functional roles *normally* associated with the sensory modality (and thus the visual qualia function as auditory qualia do in normal subjects). The entire qualia spectra are transposed across these two sensory modalities.

The transesthete's visual qualia for sound will be located in a visual qualia field. Unlike the ordinary visual field, this will somehow have to extend behind the head — we hear sounds behind us. Multiple distinct sounds can come from a single location — as when listening to a monaural recording of a symphony concert or crowd noises. Somehow this will be experienced by the transesthete as colored shapes all at the same point.

The transesthete's qualia for vision will be sounds. Distinct timbres might represent different colors; loudness might represent brightness, and so forth. The transesthete will somehow have the functional equivalent of a foveal region. Looking at a complex brightly colored scene will presumably be loud and cacophonous — yet will not distract the transesthete any more than gazing out over a busy plaza or a sunlit rippling wooded stream distracts a normal perceiver. What will the unison of octaves represent? Chords? The experience of subjective octaves and chords will have to produce no discernible difference between the transesthete and the normal experiencer — how is that possible? What will it be like to have auditory experience with the spatial resolution of vision (such that one can read letters by sound, or follow a circuit diagram, or knit, or see the subtle smile of the Mona Lisa)?

Considering the alleged possibility of a transesthete raises many questions about the connection between experience and judgment and behavior. These cannot be fully explored here — but I think it is clear that it is reasonable to suspect that a transesthete may in fact be impossible. But then the character of qualia is linked to their functional role, a role that ultimately leads to overt behavior, including linguistic report. And one way of accounting for this link is to suppose that functional role determines the character of qualia.

### **Inverted hedonic spectrum**

As a final inversion possibility, consider someone who experiences pleasure when injured, and pain under circumstances where others experience pleasure. However, the sensations produced by injury, though very pleasant, have the functional roles that pains have in normal individuals. They are aversive, and so a primary effect is to make the subject do what is required to avoid them. He moans and cries, whimpers and cringes, when pleasures are great. The hedonic invert regards the possible occurrence of pleasures with dread. However, he seeks and looks forward to pains.

The impossibility of hedonic inversion is apparent. States that are located in the body image and are strongly aversive, are pains and *cannot* be pleasures. We can be sure that others have pains because they have such aversive states. Furthermore, the senselessness of the claim that there could be intrinsically attractive pains or aversive pleasures is not epistemic — I cannot conceive in my own case having such “pleasures” or “pains”. The problem here has to do with the functional essence of these qualitative states.

### **Absent qualia: Zombies, truck drivers and blindsight**

Zombies are hypothetical individuals who behave as normal persons but lack qualia. Of course, while it is possible to stipulatively describe such cases, that alone does not make them real possibilities, any more than does the possibility of producing multiple screenplays filled with time travel show that time travel is possible. If qualia have important functional roles to play, if qualia are defined by their functional roles, then Zombies will not be functionally equivalent to normal persons.

Of course, a Zombie would have to assert that he or she has qualia, if he or she were to be behaviorally normal. In this respect, at least, a Zombie differs from an interesting real world case, that of blindsight. Blindsight victims are relatively rare — they deny that they can see. But when asked to ‘guess’ what is in front of them, they can reply with a respectable success rate.

Nevertheless, blindsight patients exhibit some serious deficiencies. Most notably they cannot initiate action based upon their blindsight. Since they are not aware that they see, these patients cannot use visual stimuli to plan their actions. This is a major problem, with clear survival implications. In addition, while some blindsight victims can make gross discriminations of line orientation, such as the difference between a projected circle and a cross, it seems that blindsight victims are generally incapable of making fine discriminations based on appearance. This is also an important deficiency. Very often the devil is in the details. The difference in detail is the difference between edible and poisonous, the face of friend or foe, mine and his, and so forth virtually without end. So in many real world situations, blindsight victims are effectively blind, even though they can answer at above chance levels when asked to guess where things are located in their visual field. (It would be interesting to know how blindsight sufferers respond to truly threatening events and objects in what

would be their visual field were they normal — to lions, and tigers, and bears. In general, where initiative is required, they behave as one who is blind.)

There are other real cases which approximate to Zombiehood. One of these is the experience of the long distance driver (David Armstrong's truck driver example) where the driver does not experience qualia yet manages to respond appropriately to the demands of driving. Sleepwalking is another real case of behavior like waking behavior but apparently without qualia. However, there are well known deficiencies in both cases. Sleepwalkers (who may or may not experience qualia) behave inappropriately. The truck driver can respond to driving conditions appropriately as long as those conditions are normal and mundane. But there are severe limits to what the truck driver can do in responding to conditions that require planning or subtle discriminations. A pattern begins to emerge: consideration of real world cases of missing qualia suggests that qualia have a function in challenging perceptual tasks.

Indeed it is here, in the sharp contrast between normal seers and both blindsight victims and more ordinary approximations to Zombiehood, such as automatic driving, that having sensory states that one is aware of *as* sensory states — qualia — may be seen to have a function. When the interpretation of sensory information is unproblematic, we can act reflexively, without thinking, without awareness of having sensory representations. We can be zombies, for the routine activities that make minimal demands. But the portrayal of zombies in the movies is no accident — they are incapable of subtleties, savoring, long term planning. When difficult discrimination tasks present themselves, we need to treat our sensory states as objects to be interpreted. Here is the state, the visual image — now, what is going on? That dark patch over there might be an animal in the shadows. That green is not the green of mold... or is it? Haven't I seen that color before, been affected just this way? And so we make objects of our sensory states in order to interpret and explain them. Metaconsciousness then allows sensory data to become qualia. To deal with (represent and use) the imagistic data from the senses *as* states of oneself is to turn the partially interpreted data into qualia, *something it is like for the system* to represent. Furthermore, turned into the objects of experience, these data have a highly dynamic nature. This is not inessential to the occurrence of qualia: if attention does not shift or there is otherwise no change, qualia fade away. Having qualia is a high-level dynamic process of interpretation of data from the senses.

## A theory of qualia

Consciousness in humans is a large scale phenomenon involving the activity of a massive neural network. Qualia are non-propositional representational states. Far from being the raw data input for processing and interpretation, qualia are highly cooked, reflecting activity in a massively interconnected network. In general, any physical or very low-level functional element of that network is identical to any other; elements differ in how they are connected to other elements, in position, rather than in any intrinsic qualities. The internal representational states of a network differ from one another in the patterns of activation, which are very complex and include large amounts of feedback, or “reentrant” processing (Edelman and Tononi 2000). The different subjective ways things are must reflect different patterns of activation.

On my view, the occurrence of a subjective way things are for a system requires two main components. The first is metaconsciousness — the system must be aware of its being aware, must represent its representing. This makes the states be *for* that system. In the absence of metaconsciousness, a system may represent, but that representing will not be sufficient to allow for the internal states to be appreciated by the system as states of itself. Thus what is it like to be say, a bat, turns in part on whether a bat is on autopilot reflexes, or whether its internal representational states can be appreciated as such by itself.

Metaconsciousness is a necessary condition for a system to have qualia. Metaconsciousness is the cure for Zombiehood. But higher order awareness does not itself determine the specific character of those qualitative states. The character of qualia is determined by the occurrent relational properties of the first order states. Since these relational properties are causal, they are functional properties. What makes an experience of pitch high rather than low are the myriad connections with kinesthetic states, motor states, affect, and possibilities of discrimination, including the absence of harmonics and so the absence of subtle distinctions of timbre.

The typical *external* acoustic sources of sounds of high pitch, although they characteristically cause the qualia, are not what give the auditory qualia their unique character. That externalist component of classical functionalism should be discarded. The external causes of qualia are not essential determinants of their character. Very real pains need not have injury as their causes; white things can look blue, and so forth. Strong externalist views of the subjective content of experience, such as Dretske (1995) are, on my view, mistaken.

What something represents, and how it represents it, are two quite different things. The former depends on connections with the world, the latter not. A bat and I may both represent the same tree, but how this representing is *for us*, may differ considerably. The function of these states is, broadly construed, the same for each system — it represents the tree, and helps us navigate around. But the local function, the functional role within the enormous webs that are our neural nets, will not generally be the same.

This account of qualia is relational, on two fronts. The nonpropositional representational states of a system will not amount to qualia in the absence of availability to metaconsciousness. This distinguishes a video camcorder, which has imagistic information, from a conscious subject with qualia. It also distinguishes a normal seer from a blindsight victim. Secondly, the character or “feel” of qualitative states is determined by their relations to other first order representational states of the system. These relations are current causal relations, activation in a network, not the static unique position of a point in a network space. Having qualia is a process, one that involves large scale activity. As William James points out in his wonderful classic phenomenology of consciousness (James 1910), consciousness is always dynamic, and each object of consciousness is surrounded by a “fringe”. As James points out, there is a *feel* to blue as well as to cold. One implication, I believe, is that the character of qualia will not be the same in very dissimilar systems, no matter what the connections to external stimuli and behavior. How something is to a system will depend on large scale organizational and so functional features of the system.

At this point, there appear to me to be no viable alternatives to such a network account of the character of qualia. Supposing consciousness is irreducibly mysterious (McGinn, Nagel) or requires positing new basic features of the universe (Chalmers) or depends on quantum gravitational effects (Penrose) seem to me to be wildly premature and, even then, profoundly unpromising. It is understandable that a functional understanding of what determines the precise character of qualia seems difficult, especially in the philosophically favorite case of color — we don’t understand color well, and the color qualia are the result of the functioning of an incredibly large and complex system. Color is arguably the most difficult case, the most cursed of the qualia. But what could an alternative to a functional account look like? What will an explanation be like in terms of quantum gravitational effects of exactly why deep blue looks deep blue? Surely such non-accounts are on the face of it not serious rivals to functional accounts of pain, pitch, proprioception, and other qualia.

## Conclusion

Thought experiments yield conclusions that are the products of assumptions and presuppositions. This is clearly true in the case of inverted spectrum arguments — they *presuppose* that qualia are independent of functional role. The limited real world experience that we have with altering perception and qualia, as with inverting goggles, suggests that as subjects adapt functionally to altered qualia, the qualia *change*. Qualia are not immutable with functional change. Qualia depend in particular upon the uses we make of data from the senses.

Consciousness has many forms and functions. Inverted spectrum and absent qualia arguments do not succeed in showing that qualia cannot be explained functionally. Far from showing that states of consciousness, including qualia, have no function or functional role, it appears that a reasonable understanding of qualia, as nonpropositional relational discriminatory states that are available to metaconsciousness, is defensible on its own grounds and compatible with our current understanding of the functioning brain as an enormous recurrent network. Consciousness evolved and has several functions, and the character of conscious states depends upon those natural functions.

## References

- Armstrong, D. 1981. *The Nature of Mind*. Cornell University Press.
- Block, N. 1990. "Inverted Earth" in *Philosophical Perspectives*, Vol 4. J. Tomberlin (ed) Ridgeview Publishing; reprinted in Block, Flanagan and Guzeldere 1997.
- BFG: Block, N., O. Flanagan and G. Guzeldere (eds) 1997. *The Nature of Consciousness: Philosophical Debates*. MIT Press.
- Carruthers, P. 1996. *Language, thought and Consciousness*. Cambridge University Press.
- Chalmers, D. 1997. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Cole, D. 1990. "Functionalism and Inverted Spectra" *Synthese* 82, 207–220.
- Dretske, F. 1995. *Naturalizing the Mind*. Bradford Books, MIT Press.
- Edelman, G. and G. Tononi. 2000. *A Universe of Consciousness*. Basic Books.
- Guzeldere, G. 1995. "Is consciousness the perception of what passes in one's own mind?" In T. Metzinger (ed). *Conscious Experience*. Schoeningh-Verlag; reprinted in Block, Flanagan and Guzeldere 1997.
- James, W. 1910. "The Stream of Consciousness." Chapter XI of *Psychology*, Henry Holt and Company; reprinted in Block, Flanagan and Guzeldere 1997.
- Kosslyn, S. and O. Koenig 1995. *Wet Mind: the New Cognitive Neuroscience*. The Free Press, Simon and Schuster.

- Lycan, W. 1996. *Consciousness and Experience*. MIT Press.
- McGinn, C. 1991. *The Problem of Consciousness*. Blackwell.
- Nagel, T. 1974. "What is it like to be a bat?" *Philosophical Review* 83:4, 435–450.
- Nozick, R. 1994. *The Nature of Rationality*, Princeton University Press.
- Penrose, R. 1996. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.
- Pinker, S. 1999. *How the Mind Works*. W. W. Norton and Company.
- Rosenthal, D. 1990, 1997 "A Theory of Consciousness" Zif Technical Report, Bielfeld Germany, adapted in Block, Flanagan and Guzeldere 1997.
- Sorley, W. R. 1920/1965. *A History of British philosophy to 1900*. Cambridge: CUP.
- Weiskrantz, L. 1986 *Blindsight: A Case Study and Implications*. Oxford University Press.

# Sensations and grain processes\*

George Graham and Terry Horgan

University of Alabama at Birmingham / University of Memphis

## Introduction

This chapter celebrates an anniversary, or near anniversary. As we write it is just more than 40 years since U. T. Place's "Is consciousness a brain process?" appeared in the *British Journal of Psychology*, and just less than 40 since J. J. C. Smart's "Sensations and brain processes" appeared, in its first version, in *The Philosophical Review* (Place 1962/1956; Smart 1962/1959). These two papers arguably *founded* contemporary philosophy of mind. They defined its central preoccupation (the ontology of consciousness), introduced its regnant ontology (materialism/physicalism), offered its initial logical techniques (e.g. appeals to the concepts of identity and event) as well as empirical reliances (on neuroscience), and, finally, offered its most seminal sectarian doctrine (central state materialism). No history of philosophy of mind can afford to neglect them. (See for discussion Macdonald 1989.) It is to be near both the letter of Smart's paper and the spirit of his and Place's concern for how best to develop a philosophical understanding of consciousness that we entitle our own paper 'Sensations and grain processes'. (That is not a typo, as will be seen momentarily.)

In the years since Smart's and Place's contributions, the landscape of philosophy of mind has changed in many ways. The empirical reliances of very recent philosophy of mind have expanded to include the cognitive sciences (not just neuroscience); central state materialism alternately has been displaced by causal role and functional specification theories of mind; fresh logical techniques have been introduced (e.g. the concept of supervenience); and it is less clear than it was in the 1950s whether materialism is to be preferred to some more ecumenical ontology (such as naturalism).

This paper is about the current status of the philosophy of consciousness (which we take to be phenomenal consciousness, for purposes of the paper; of which more momentarily); not so much the case for materialism (which we take to be, for the most part, complicated by considerations which we shall adduce below) as what the philosophical program for doing the philosophy of

the conscious mind is and where it can, and most importantly can't, rely on cognitive science. There is quite a lot of ground to cover in a short space. So let's begin by demarcating the subject matter and outlining the paper to follow.

## Subject and outline

Phenomenal consciousness is the “what it’s like” aspect of our mental lives. It is especially salient in bodily sensations and in certain perceptual experiences. Jackson (1982) gives as examples of phenomenal states (so-called *qualia*) “the hurtfulness of pains, the itchiness of itches, pangs of jealousy,... the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise, or seeing the sky” (p. 127). But phenomenal consciousness is more extensive in scope than this. As Bieri (1995) remarks, using the term “sensing” for mental states that have a phenomenal aspect, “Sensing comprises a variety of things: sensory experiences like seeing colours and hearing sounds; bodily sensations like lust and pain; emotions like fear and hatred; moods like melancholy and serenity; and finally, desires, drives and needs, i.e., our experienced will. All these states are not only *there*; rather, *it is like something* to be in them” (p. 47). The specific *thisness* of such states — “It is like *this*,” as said or thought by someone while undergoing the state and attending to it from the subjective, first-person, point of view — is its phenomenal character.

Phenomenal consciousness strikes many philosophers as puzzling or mysterious, more so than other aspects of mentality (including other aspects that are often characterized using the language of consciousness, such as attention or second-order awareness of one’s first-order mental states). Specific philosophical puzzles arise about its ontological or metaphysical status, given our overall scientific worldview; about its causal efficacy, if any; about how it is to be explained in naturalistic terms; and about why it should have emerged at all, in the course of evolution.

In Section 1 of this paper we describe an interdisciplinary research program in cognitive science that bears directly on the goal of achieving a scientific understanding of phenomenal consciousness: identifying what we call the ‘causal grain’ of phenomenal states, at both the neurophysical and the functional-representational levels of description. Scientific progress on this research program is foreseeable, and is directly relevant to philosophical issues about phenomenal consciousness; moreover, at present it is hard even to imagine other relevant kinds of scientific progress. In Section 2 we briefly elaborate the

three main philosophical puzzles about phenomenal consciousness that have made it seem especially problematic to many philosophers — puzzles involving ontological status, causal role, and explainability. In Section 3, the heart of the paper, we argue that as far as one can now tell, from our current epistemic existential situation, even if the causal grain of phenomenal consciousness were to become fully understood within cognitive science, various theoretical options concerning qualia that are presently live theoretical options in philosophical discussion would all still *remain* live theoretical options. We conclude, in Section 4, with some methodological remarks about the status of competing philosophical approaches to phenomenal consciousness, in light of our argument in Section 3. With respect to evolutionary approaches specifically, there turn out to be certain important, tractable-looking, questions about the evolution of states with the distinctive causal grain associated with qualia — questions that can be posed and addressed independently of the competing philosophical positions about phenomenal consciousness.

## 1. The causal grain of phenomenal consciousness

Here is a large-scale, long-term, but potentially empirically tractable project for cognitive science: to identify the specific causal role or roles associated with phenomenal states — to identify what, in some sense, those states do. We use ‘associated with’ and ‘in some sense’ in a deliberately neutral way, in order not to prejudge the issue whether these states themselves, as opposed to other states that co-occur with them (e.g., representational states, or neurophysical states), have causal efficacy; that is one of the philosophical issues to be discussed below. Our usage also is not meant to prejudge certain identity questions, e.g. whether states of phenomenal consciousness are literally identical with “associated” representational or neurophysical states.

Let us call this scientific program the *grain project*, since it involves investigating the causal roles associated with phenomenal consciousness at several levels of detail or resolution. One level is common-sense psychology (so-called folk psychology). This is the surface or coarse grain, so to speak; it includes various familiar platitudes about the phenomenal aspects of experience — for instance, that the hurtfulness of pain is unpleasant, that pink looks more like red than it does like green, that the smell of rotten eggs is disgusting, and so forth. Folk psychology need not be completely *correct* in what it says about qualia, but its homely platitudes are innocent until proven guilty.

Below the surface level of folk psychology are two finer-grained levels of causal detail. One involves the specific functional-representational roles, in the human cognitive system, that are associated with phenomenal states. (The platitudes of folk psychology do say something about these roles, of course, but presumably there is more to say.) Identifying these roles is a task for empirical research and theory, and is likely to occur in the context of working out a more general theoretical account of human cognitive processing.

The other, yet more detailed, fine-grained level involves the specific neural states and processes that directly subserve phenomenal consciousness. Here again, identifying these states is a task for empirical research and theory, and doing so is likely to occur in the context of working out a more general account of how various kinds of mental states are physically subserved in the brain.<sup>1</sup>

Relevant work on the neural correlates of consciousness proceeds apace, with traditional empirical tools and newly available technology: neural-anatomical studies of cell types and connectivity patterns, recordings from individual neurons, deficit and lesion studies in which special neural structures are destroyed, and most recently neuroimaging techniques.

At the level of cognitive architecture, there is certainly much theoretical work that purports to address itself to the functional-representational roles associated with states of consciousness. However, it seems to us that David Chalmers (1995, 1996) is right in maintaining that much of this work really addresses various other aspects of mentality that are often described in the language of consciousness — e.g., attentional processes, the ability to access and/or report one's own mental states, the deliberate control of one's own behavior, etc. — rather than directly addressing phenomenal consciousness itself. Concerning cognitive roles associated with phenomenal consciousness in particular, it appears to us that perhaps the most pertinent discussions are to be found in the recent writings of philosophers like Dretske (1995), Lycan (1996), and Tye (1995). (More on this in Section 3.1.)

Ideally, a completed account of the causal grain within cognitive science would integrate the three levels of detail. Also, it would itself be integrated into a more general and well articulated general account of mental processing, addressing a variety of mental phenomena and their interconnections with the phenomenal aspects of mentality. In research on color vision, just to take one example, one approach is to assume that functional-representational grain level descriptions can be developed through behavioral studies of color mixing and matching behavior and then subsequently connected with underlying neural processes. One of the more heralded findings of modern color science is

that properties of retinal receptors can explain many of the facts of color mixing and matching. (For discussion see Clark 1998.) There also has been considerable discussion within color science of whether both the functional-representational and neural levels in color perception are modular in organization and design: whether color vision is functionally and neuronally independent of perception of shape, motion, and other properties of a visual stimulus. (For discussion see Davidoff 1991 and Livingstone and Hubel 1987.) We would expect that on the basis of research on color vision, and on numerous other types of phenomenal experience, information gathered from the brain and functional-representational levels will be mutually informative. Just as information drawn from functional descriptions will inform descriptions of the neural level, information drawn from the brain will guide development of functional-representational models of phenomenal consciousness and perhaps even help to inform models at the folk psychological surface level. (For related discussion see Flanagan 1996 and Zawidzki and Bechtel in press.)

The grain project is a rich program of inquiry within cognitive science. It is empirically and theoretically tractable, and progress on this project will certainly enhance our understanding of phenomenal consciousness. But enhancement of understanding is one thing; resolution of deep philosophical problems is quite another.

## 2. Three philosophical problems

We begin with two familiar philosophical thought experiments, as groundwork for the discussion below: inverted qualia and absent qualia. It seems to make sense — not to be logically or conceptually contradictory — to imagine a “possible world” that is just like the actual world in all physical respects (including being governed by all the same physical laws that prevail in the actual world) but in which qualia are differently instantiated than they are in the actual world. On “inverted qualia” versions of this thought experiment, one imagines that phenomenal properties are instantiated in a way that is somehow inverted relative to their actual-world instantiations: for instance, the qualitative aspects of color-experiences are systematically inverted. (What it’s like for someone to see red is what it’s like for us to see green; and so forth.) On “absent qualia” versions, one imagines that phenomenal properties are not instantiated at all. The creatures in the given world who are duplicates of humans and other sentient actual-world creatures are *zombies*, in the sense that

the states they undergo lack phenomenal content altogether; although these zombies do instantiate those aspects of mentality that are characterizable in functional-representational terms, there isn't anything at all that it's like, for them, to undergo such states.

The assumption that there seem to be possible worlds just like the actual world but containing inverted or absent qualia raises the question of just what sort of possibility we are considering when we say that apparently there could be inverted and absent qualia. In answering this question it will be useful to introduce a concept that is frequently employed in recent metaphysics and philosophy of mind: the notion of *supervenience*. Supervenience is an ontological determination-relation between facts or properties at different levels of description: the lower-level facts and properties determine the facts and properties supervenient upon them, in the sense that there cannot be a difference at the higher level without some underlying difference at the lower level. Two types of supervenience are important to distinguish: (1) *logical* (or *conceptual*) supervenience, which says that it would be logically impossible for the higher-level facts and properties to differ without some underlying lower-level difference; and (2) *nomological* (or *natural*) supervenience, which says that it would be contrary to certain laws of nature for there to be one kind of difference without the other. (For further discussion see Kim 1990, Horgan 1993, Chalmers 1995 Chapter 2.)

In saying that there apparently could be inverted and absent qualia in worlds otherwise just like the actual world we are saying that even if inverted or absent qualia were contrary to laws of nature, it seems *logically* (or *conceptually*) possible for there to be inverted or absent qualia. Phenomenal properties of experience appear not to logically supervene on those aspects of mentality that are characterizable in functional-representational or neural terms. They seem to constitute — as we shall suggest momentarily — a kind of *metaphysical residue* that provides the basis for three philosophical puzzles or problems concerning phenomenal consciousness.

### 2.1. The problem of ontological status

The apparent logical possibility of physical-duplicate worlds in which qualia are differently instantiated, or are absent altogether, creates a problem about their ontological status in relation to our overall scientific worldview. For, with the exception of phenomenal properties, it is plausible that the other properties posited in special sciences and in common sense are reductively explainable —

and that reductive explanation rests on logical supervenience, typically involving the functional analyzability of the supervenient properties. As Chalmers (1995) observes:

A reductive explanation of a phenomenon need not require a *reduction* of that phenomenon, at least in some senses of that ambiguous term. In a certain sense, phenomena that can be realized in many different physical bases — learning, for example — might not be reducible in that we cannot *identify* learning with any lower-level phenomenon, but this multiple realizability does not stand in the way of reductively *explaining* any instance of learning in terms of lower-level phenomena. ... In general, a reductive explanation of a phenomenon is accompanied by some rough-and-ready *analysis* of the phenomenon in question, whether implicit or explicit. The notion of reproduction can be roughly analyzed in terms of the ability of an organism to produce another organism in a certain sort of way. It follows that once we have explained the processes by which an organism produces another organism, we have explained that instance of reproduction. ... The possibility of this kind of analysis undergirds the possibility of reductive explanation in general. Without such an analysis, there would be no explanatory bridge from the lower-level physical facts to the phenomenon in question. With such an analysis in hand, all we need do is show how certain lower-level physical mechanisms allow the analysis to be satisfied, and an explanation will result. ... For the most interesting phenomena, including phenomena such as reproduction and learning, the relevant notions can be analyzed *functionally*. ... It follows that once we have explained how those functions are performed, then we have explained the phenomena in question. ... The epistemology of reductive explanation meets the metaphysics of supervenience in a straightforward way. A natural phenomenon is reductively explainable in terms of some low-level properties precisely when it is logically supervenient on those properties. ... What is most important is that if logical supervenience fails. ... then *any* kind of reductive explanation fails, even if we are generous about what counts as explanation. (pp. 43–50)

Although reductive explanation is not by any means *a priori* (since the lower-level facts and principles invoked are straightforwardly empirical), there is an *aspect* of the overall explanation that is relatively *a priori*, viz., the appeal to what is essential about the higher-level supervening property — typically (as in Chalmers' examples of learning and reproduction) its definitive functional role. Reductive explanation is a matter of logical supervenience.

Now, the problem about qualia is that they seem not to be logically supervenient on physical properties and facts, and thus not to be reductively explainable. This is the moral of the apparent logical intelligibility of inverted-qualia and absent-qualia thought-experiments. Thus, they are a metaphysical anomaly or residue, vis-a-vis our overall scientific account of the world.

By contrast, it does not make conceptual sense to suppose, for example, that there is a physical-duplicate world in which the creatures who are our physical counterparts lack genes. The *concept* of gene rules this out, given that there are physical constituents of human cells (*viz.*, components of DNA) that play the gene-role. Likewise, it does not make conceptual sense to suppose that there is a physical-duplicate world in which the stuff that is the physical counterpart of this stuff in my glass fails to be liquid. Again, the *concept* of liquidity rules this out, given that the macro-dispositions that constitute liquidity are the result of the operative inter-molecular physical forces. On reflection, this kind of point looks to be highly generalizable. It is plausible that for most all of our higher-level concepts, there will be logical supervenience at work: in principle, the instantiation of higher-level properties will be reductively explainable on the basis of the facts and laws of natural science (arguably ultimately those of basic physics), together with facts about the nature of higher-level concepts and of the meanings of terms expressing those concepts. (For further articulation and elaboration of this point, see Chalmers 1995 Section 2.5.) But phenomenal consciousness is metaphysically puzzling or mysterious because it does not seem to be logically supervenient on physical facts and properties, and it therefore it does not appear to be reductively explainable.

## 2.2. The problem of causal efficacy

If qualia are not logically supervenient on underlying physical facts and properties, then serious doubts arise whether qualia play any causal role in generating behavior. After all, in a physical-duplicate world in which a person's physical counterpart has inverted qualia, or lacks qualia altogether, that counterpart-person still behaves (by hypothesis) exactly as the person in the actual world behaves, despite having different qualia, or none at all. This being so, it appears *prima facie* that the phenomenal aspects of one's mental life play no genuine causal role at all with respect to one's behavior; rather, the real causal work seems to be done by properties that we share in common with our physical counterparts in these inverted-qualia and absent-qualia physical-duplicate worlds. (For further elaboration of this line of reasoning, see Horgan 1987.)

## 2.3. The problem of explaining phenomenal consciousness

This problem is closely connected to the other two. A thoroughly satisfying explanation of phenomenal consciousness would be a reductive one in which

(i) certain lower-level phenomena would be explained, (ii) qualia themselves would then be explained as logically supervenient on these lower-level phenomena, and (iii) qualia would thereby be shown to be causally efficacious. But since the prospects for this kind of explanation are thrown into doubt by the apparent conceivability of inverted-qualia and absent-qualia scenarios, we find ourselves faced with certain recalcitrant looking ‘why’-questions about phenomenal consciousness. For instance: even if one could explain, in broadly evolutionary terms, why there would have emerged states with the functional-representational roles associated with phenomenal consciousness, the question would remain why these states have the specific phenomenal character they do, rather than other kinds of phenomenal character (e.g., inverted in certain respects relative to the actual world), or none at all. Moreover, trying to answer such questions by appeal to some distinctive survival/reproductive advantage allegedly accruing to the qualia we actually experience — an advantage in virtue of which these states supposedly have emerged in evolution under pressures of natural selection — seems to be thwarted by the same considerations that throw into doubt the causal efficacy of qualia: in a possible world that is physically just like our own but different with respect to how qualia are instantiated, the same neural states emerge under natural selection as have emerged in the actual world, with the same functional-representational roles; and yet these states have different phenomenal content in those worlds, or none at all. In short, selective advantage seems to attach only to the functional-representational roles *associated* with qualia, rather than with phenomenal content itself.

So there seem to be residual, intractable, explanatory mysteries about phenomenal consciousness. Why aren’t we all zombies? And, given that we are not, why does our experience have the *particular* phenomenal character it does? And how could phenomenal character make any causal difference to behavior? These explanatory puzzles inevitably arise if inverted-qualia and absent-qualia physical-duplicate worlds are logically possible, as they seem to be.

### 3. Three persistently live theoretical positions

We will discuss three approaches to phenomenal consciousness, each of which has some serious currency in the contemporary philosophical literature. In each case we will consider both the attractions or theoretical benefits and the problems or theoretical costs in the given view, and we will explain why, as best

one can now tell, both benefits and costs would remain largely intact even if one had available a complete and detailed account of the causal grain of phenomenal consciousness. The upshot will be that each position apparently would remain an epistemically live theoretical option concerning phenomenal consciousness, even if the grain project were to be successfully completed.

By ‘epistemically live theoretical option’ we mean an option that someone who is well informed of all relevant information, and who carefully weighs the comparative theoretical advantages and disadvantages of the given position in relation to others, could rationally judge to be true, or probably true. This does not mean that all rational persons with full relevant information *would* make this judgment, however. On the contrary, individual rational persons, all possessing the same relevant information and all judging rationally, could diverge as to which of the theoretically live options (if any) they believe — because they weigh the comparative pros and cons differently from one another. Thus, the competing positions would be live theoretical options *for a community of rational inquirers*, even if different members of that community have incompatible — but respectively rational — beliefs about which position is true and which positions are false.

Of course, some individual members might find themselves unsure what to believe, because for them no single position sufficiently outweighs all the others in terms of theoretical benefits and costs. This is essentially the state of mind of each co-author of the present paper, with respect to the various competing theoretical positions we will now consider. There are no rules, beyond a certain point, about how theoretical benefits and costs should be weighed. There is room for both individual judgment and suspension of judgment.

### 3.1. Phenomenal states as functional-representational states

A currently popular approach in philosophy of mind is to maintain that phenomenal states are functional-representational states of a certain distinctive kind — that phenomenal character is just a specific kind of representational content. Such views construe mental representation itself in broadly functional terms, as a matter of the typical causal role — possibly an environmentally situated role — that the states play in a creature’s cognitive economy.

For concreteness we will focus on the specific version of this approach developed by Michael Tye (1995). (The points we will make are generalizable to other versions, as we will explain below.). Tye says:

Phenomenal content, I maintain, is content that is appropriately poised for use by the cognitive system, content that is abstract and nonconceptual. I call this the PANIC theory of phenomenal character: phenomenal character is one and the same as Poised Abstract Nonconceptual Intentional Content. ... The claim that the contents relevant to phenomenal character must be *poised* is to be understood as requiring that these contents attach to the (fundamentally) maplike output representations of the relevant sensory modules and stand ready and in position to make a direct impact on the belief/desire system. ... The claim that the contents relevant to phenomenal character must be *abstract* is to be understood as demanding that no particular concrete objects enter into these contents. ... Since different concrete objects can look or feel exactly alike phenomenally, one can be substituted for the other without any phenomenal change. ... The claim that the contents relevant to phenomenal character must be *nonconceptual* is to be understood as saying that the general features entering into these contents need not be ones for which their subjects possess matching concepts. ... Consider. ... color. ... We have names for only a few of the colors we can discriminate, and we also have no stored representations in memory for most colors either. There is simply not enough room. (pp. 137–39)

Tye, like many others in current philosophy of mind, construes mental intentionality — the ‘I’ in ‘PANIC’ — as a matter of causal covariation between representing state and item represented. “The key idea,” he says, “is that representation is a matter of causal covariation or correlation (*tracking*, as I shall often call it) under optimal conditions” (p. 101). Concerning the intentionality of phenomenal content, he says the following. (Red<sub>29</sub> is a specific, fine-grained, shade of red.)

Which features involved in bodily and environmental states are elements of phenomenal consciousness? There is no a priori answer. Empirical research is necessary. ... They are the features our sensory states track in optimal conditions. ... I conjecture that for perceptual experience, [these] will include properties like being an edge, being a corner, being square, being red<sub>29</sub>. (pp. 137–41)

If Tye’s account of phenomenal content as a species of intentional content (viz., PANIC) is correct, then the three philosophical puzzles discussed in Section 2 get cleanly resolved. There is no special problem of ontological status for phenomenal states; since these are just a species of functional-representational states, occurrences of such states are reductively explainable. Likewise, there is no special problem of causal efficacy, since it is not metaphysically possible for there to be world that is physically just like ours but in which qualia are inverted or absent. The problem of explaining phenomenal consciousness becomes tractable, being just the problem of explaining why and how there

came to be states with this specific kind of functional-representational role — i.e., states with poised, abstract, nonconceptual, intentional content.

It is important to appreciate, however, that the functional-representational roles associated with phenomenal states could conform with Tye's PANIC story even if phenomenal content itself — the 'what it's like' aspect of experience — is something over and above (i.e. is non-identical with) the associated representational content. Tye's position is that phenomenal content is literally *identical* with the specific kind of intentional content he describes (viz., the kind that is abstract, nonconceptual, maplike, and poised to influence the belief-desire system). Tye could be mistaken about this identity claim even if he is right about the associated functional-representational role.

What bearing does the grain project have on Tye's position? Well, successful completion of this project would tell us specifically, and in detail, just what sort of functional-representational roles are associated with phenomenal states, and just which neural states subserve phenomenal states. The resulting account might not uncover functional-representational roles that meet Tye's generic PANIC characterization; and if not, then his position would thereby get decisively disconfirmed. Suppose, however, that the associated functional-representational roles for qualia did turn out to conform to the PANIC account, with the completed full grain story filling in the empirical details — by telling us, for instance, just which bodily and/or environmental features get represented when qualia are instantiated, and just how qualia are physically realized in the brain. Although such an outcome would be consistent with Tye's position, it would not decisively confirm it; for, he nonetheless might be wrong in claiming that phenomenal content is *identical* to intentional content that is poised, abstract, and nonconceptual (even though phenomenal content is *associated* with such intentional content).

Furthermore, as far as one can now tell, the kinds of considerations that currently count as reasons to doubt Tye's proposed reduction of phenomenal content to a species of representational content would still be operative, and would still count against this reductive account, even if a successful completion of the grain project were to vindicate the PANIC approach to functional-representational role. The strong intuition would still persist that there is more to phenomenal states than their associated functional-representational role — viz., their specifically *phenomenal* character, what it's like to be in them. When one imagines an inverted-qualia world or an absent-qualia world that is physically just like the actual world, the world imagined is one in which there are indeed states instantiated that have all the relevant functional-representational

features that are uncovered in the grain story (PANIC states, we are now supposing); but the problem is that what they are like is different (inverted qualia), or there's nothing at all that they are like (absent qualia). Adding further details to the PANIC story, via a suitable completion of the grain project, would not change things, as far as the apparent intelligibility of inverted-qualia and absent-qualia scenarios is concerned. So their intelligibility would still tell against Tye's reductivist position, no less than it does now.

Tye does not ignore the idea that there is "something that it's like" to undergo phenomenal states, and the idea that "knowing what it's like" is something essentially subjective. He does attempt to accommodate these aspects of phenomenal experience. He says:

I call the concepts relevant to knowing the phenomenal character of any state 'phenomenal concepts.' Phenomenal concepts are the concepts that are utilized when a person introspects his phenomenal state and forms a conception of what it is like for him at that time. These concepts, in my view, are of two sorts. Some of them are indexical; others are predicative. Suppose, for example, I am having a visual experience of  $\text{red}_{29}$ . I have no concept  $\text{red}_{29}$ . So, how do I conceptualize my experience when I introspect it? I bring to bear the phenomenal concepts *shade of red* [a predicative phenomenal concept] and *this* [an indexical phenomenal concept]. Intuitively, possessing the phenomenal concept [*shade of*] *red* requires that one have experienced red and that one have acquired the ability to tell, in the appropriate circumstances, which things are red directly on the basis of one's experiences. ... What about the phenomenal concept *this*? Possessing this concept is a matter of having available a way of singling out, or mentally pointing to, particular features that are represented in sensory experiences *while* they are present in the experiences, without thereby describing those features (in foro interno). ... What one has [in having the indexical concept] ... is a way of singling out or discriminating the feature for as long as one attends to it in one's experience (and perhaps for a very short time afterward). (pp. 167–68)

In essence, Tye's approach treats 'knowing what it's like' as involving certain cognitive *abilities*: in the case of knowing what it's like to see red, (i) the ability to classify things as red directly on the basis of one's experiences (and without collateral information), and (ii) the ability to indexically pick out, in thought, a shade of red that is currently being represented PANIC-wise in one's experience (e.g.,  $\text{red}_{29}$ ).

For those who find themselves thinking that the PANIC account leaves out the genuinely phenomenal aspects experience, however, the trouble is that these kinds of cognitive abilities seem likewise to leave them out. Residue remains. Like PANIC states themselves, the kinds of cognitive abilities featured

in Tye's account of knowing what it's like would be instantiated by duplicates of ourselves in an absent-qualia world. These zombie-duplicates would possess the ability to classify red things as red just by looking, and also the ability to attend to, and indexically pick out, features that are currently mentally represented PANIC-wise (e.g., the current presence of red<sub>29</sub>). But there would not be anything that it's like when they undergo PANIC states, or when they exercise these cognitive abilities during the occurrence of PANIC states; hence there would be no *knowing* what it is like, for them.

To underscore the point, consider the famous thought experiment in Jackson (1982): Mary, a brilliant neuroscientist who knows all about how human color-vision works at the functional-representational and the neural levels of description (she knows the full story about the causal grain of color-qualia), but has spent her life in a black-and-white room (better: with lenses on her corneas that filter out color) without ever having had a color experience herself. Plausibly, she is burning with curiosity to know what color experiences are like. But surely what she is dying to acquire is not the mere *ability* to determine what color things are directly by looking, or the mere ability to indexically pick out color-features when they are visually presented to her. After all, she already understands how these abilities operate, and she is already adept (we may suppose) at using her scientific instruments as an aid to direct perception as a way to identify both the coarse-grained and the fine-grained colors of things. Rather, she is burning with curiosity because she wants to know what it's like to experience red. The abilities in question are a pale monochromatic ghost of genuine knowing what it's like, just as functional-representational PANIC states are a pale ghost of genuine phenomenal content. Or so it seems, anyway, given the apparent conceivability of a possible world that is physically just like ours but in which qualia are not instantiated.<sup>2</sup>

Now, these persistent and recalcitrant intuitions are ones that Tye is committed to explaining away. He makes a valiant effort, in a final chapter entitled "Can you really imagine what you think you can?" And his answer, as one may guess, is no. Tye charges: We cannot imagine possible worlds otherwise just like the actual world but with absent or inverted qualia. We merely think we can imagine such worlds. Such worlds are not conceptually possible.

We lack the space here to examine his argument in detail, but for us the upshot is this: Although Tye perhaps succeeds in explaining why our absent-qualia *counterparts* in a zombie-duplicate world cannot really imagine what they think they can (that is, given zombie concepts of the phenomenal they cannot really imagine a world physically just like their own in which certain

putative 'what it's like' aspects of mentality, which they think they themselves experience, are either systematically inverted or are altogether absent), he does not succeed in showing that *we ourselves* (given our non-zombie concepts) cannot imagine a world that is physically just like our own world but in which qualia are either systematically inverted or are altogether absent. His account seems correct for the phenomenally deprived, but incorrect for ourselves.

The points we have made about Tye's position apparently are applicable, *mutatis mutandis*, to any version of the generic view that phenomenal content is identical to some kind of functional-representational property; variants include Dretske (1995) and Lycan (1996). For each version, successful completion of the grain project might either decisively disconfirm the account, or else yield results that are consistent with the account and also fill in further detail. But even if things should go the latter way, this would not undercut the apparent conceivability of inverted-qualia and absent qualia-scenarios — and thus would leave intact the reasons for thinking that any attempt to explain away this conceivability as illusory would apply at best to zombie-duplicates who lack qualia, rather than to ourselves. So the apparent imaginability of inverted-qualia and absent-qualia scenarios would provide reason to claim that phenomenal-content properties are distinct from these associated functional-representational properties, and hence that the given account is mistaken in asserting that they are identical. Denying that phenomenal properties are identical to functional-representational properties would remain an epistemically live theoretical option.

On the other hand, there also would be significant theoretical advantages in the claim that phenomenal properties are identical to whatever functional-representational properties turn out to be associated with them under a successful completion of the grain project. The truth of this identity hypothesis would render nonproblematic the ontological status of phenomenal states, since they would be logically supervenient on facts and properties at the physical level of description.<sup>3</sup> There would be no special problem of causal relevance for phenomenal states, since conceptual supervenience would preclude the genuine metaphysical possibility of a world that is physically just like ours but different with respect to how phenomenal properties are instantiated. And the problem of explaining phenomenal consciousness would reduce to the tractable problem of explaining why and how there came to be states with the relevant functional-representational role, in creatures like ourselves. These theoretical advantages of the hypothesis that phenomenal properties are identical to functional-representational properties would be sufficiently powerful

that this hypothesis would be an epistemically live option — as would be Tye's contention that when it comes to putative inverted-qualia and absent-qualia scenarios, you cannot really imagine what you believe you can.<sup>4</sup>

### 3.2. Phenomenal states as ontologically *Sui Generis*

Another influential current approach to phenomenal consciousness treats qualia as ontologically fundamental features of the world, over and above those that figure in the basic laws of physics. This view, probably implicit in Jackson (1982), is explicitly articulated and defended by Chalmers (1996). Briefly, the position goes as follows. It really is logically possible (just as it seems to be) for there to be a world that is physically just like our world but in which qualia-instantiations are systematically inverted in some respect relative to the actual world. Likewise, it really is logically possible (just as it seems to be) for there to be a world that is physically just like ours but in which qualia are not instantiated at all. These are genuine conceptual possibilities because phenomenal properties simply are not logically supervenient on physical properties and facts. On the other hand, these logical possibilities are not nomological possibilities; they are contrary to the fundamental laws of nature. This is because a complete inventory of nature's fundamental properties and laws would include not only the theoretically basic physics-level properties (mass, charm, spin — whatever they turn out to be) and the theoretically basic physics-level laws governing these properties and their interrelations — but also phenomenal properties themselves, together with certain theoretically basic laws of inter-level supervenience for qualia. (Thus, inverted-qualia and absent-qualia worlds are nomologically impossible, even though they are consistent with the fundamental laws of physics, because they are not consistent with the full set of fundamental laws of nature.) Our scientific worldview needs to be expanded beyond materialism: phenomenal properties are no less metaphysically basic than the fundamental properties of physics, and the inter-level supervenience laws governing phenomenal properties are no less metaphysically basic than the fundamental laws of physics.<sup>5</sup> This expanded worldview is a version of scientific naturalism, but it is a denial of scientific materialism.

What should be said about the three philosophical problems mentioned in Section 2, given this position? The problem of the ontological status of phenomenal states receives a very dramatic answer: phenomenal properties have the distinctive, and highly elite, status of being among the metaphysically rock-bottom, basic and ultimate, properties instantiable in the natural world. (Ma-

terialists, who think that only the most basic properties posited in theoretical physics have this status, are simply wrong.)

The problem of causal efficacy is vexed and difficult, given this metaphysical account of qualia. For, even though qualia are *nomologically* linked to associated physical and functional-representational properties, there are counter-nomological possible worlds (i) that are just like our actual world in all physical respects (including being governed by all the same *physical* laws), but (ii) in which qualia are differently distributed or are absent altogether. This being so, there is *prima facie* reason to think that qualia as such do no *real* causal work at all — that the real work is done by the functional-representational properties, as realized by neurophysical properties. Chalmers discusses this issue at some length, suggesting that there may be an appropriate way of construing causal efficacy that will apply to qualia and will contravene the initial impression that they can do no real causal work if they are merely nomically — and not logically — supervenient on physical facts and properties. But in the end there may not be any viable such account of causal efficacy, in which case this ontological treatment of qualia would render them epiphenomenal.

Concerning the problem of explanation, the approach yields a mixed verdict. On the one hand, it allows certain explanatory questions concerning phenomenal consciousness to receive scientific answers — provided that the explanatory resources used to provide the answers include the relevant inter-level supervenience laws. For example, in answer to a question like ‘Why did phenomenal consciousness arise, in the course of evolution?’, an overall scientific answer would have two components: first: an explanation of the emergence, in evolution, of the relevant functional-representational states, as neurophysically realized; and second, an appeal to the applicable inter-level laws telling us that phenomenal properties supervene on those functional-representational states, so realized neurophysically. Also, particular instantiations of phenomenal properties can be explained by reference to relevant lower-level properties together with inter-level supervenience laws. On the other hand, certain explanatory questions about phenomenal consciousness simply have no answer, on this view. For example, if one focuses on the general supervenience laws governing qualia, and one asks why *they* obtain, the response has to be that these laws simply have no explanation. They are explanatory bedrock in science, in just the same way that the fundamental laws of physics are explanatory bedrock. The inter-level supervenience laws themselves must be accepted (as the British emergentists used to say) “with natural piety” (see McLaughlin 1992).

The principal — and significant — theoretical advantage of this position is that it thoroughly respects the data served up by first-person phenomenal experience. For many of us, the full richness of our phenomenal experience produces a persistent and deep-seated conviction that there is more to phenomenal content than its associated functional-representational role and/or its specific mode of physical realization; hence the apparent intelligibility of absent-qualia and inverted-qualia scenarios. On Chalmers' view, this conviction is correct, rather than being treated as a troublesome mistake that must be somehow explained away.

But this attractive feature brings significant theoretical costs in its wake. Positing fundamental properties outside the domain of physics, and additional fundamental laws of nature beyond those of physics, seriously complicates our overall scientific conception of the world and our conception of ourselves as denizens of the natural order; it is far simpler to suppose that physics alone is the realm of all fundamental properties and laws, and that all higher-order properties and facts are logically supervenient on physics-level properties and facts. Two additional considerations seriously exacerbate this loss of theoretical simplicity. First, as far as one can tell, phenomenal consciousness is unique — or close to unique — as a phenomenon that allegedly requires a departure from materialistic metaphysics. (Chalmers himself defends this claim in section 2.5 of his book, entitled “Almost everything is logically supervenient on the physical.”) Second, with the exception of putative inter-level supervenience laws involving qualia, the other fundamental laws of nature evidently apply directly to the fundamental physical constituents of things — whereas phenomenal consciousness is evidently instantiated only in complex, highly evolved, sentient organisms.<sup>6</sup> The epistemic force of such simplicity-driven considerations was well expressed 40 years ago in Smart (1962/1959), the paper whose title ours echoes. Smart wrote:

The suggestion I wish to resist is ... that to say ‘I have a yellowish-orange after-image’ is to report something irreducibly psychical. Why do I wish to resist this suggestion? Mainly because of Occam's razor. It seems to me that science is increasingly giving us a viewpoint whereby organisms are able to be seen as physicochemical mechanisms: it seems that even the behavior of man himself will one day be explicable in mechanistic terms. There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. ... I just cannot believe that this can be so. That everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together — roughly, biology is to physics as radio-engineering is to electromagnetism)

except the occurrence of sensations seems to me to be frankly unbelievable. Such sensations would be 'nomological danglers'. ... It is not often realized how odd would be the laws whereby these nomological danglers would dangle. ... Certainly we are pretty sure in the future to come across new ultimate laws of a novel type, but I expect them to relate simple constituents: for example, whatever ultimate particles are then in vogue. I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles). ... Such ultimate laws would be like nothing so far known in science. They would have a queer 'smell' to them. I am just unable to believe in the nomological danglers themselves, and in the laws whereby they would dangle. (Smart 1962 pp. 161–62)

In addition, there is also the very real possibility that under an adequate account of causal efficacy, a view like Chalmers' would end up having to embrace epiphenomenalism concerning qualia. Since the hypothesis that the what-it's-like aspects of our experience have no genuine causal influence on our behavior is grossly offensive to common sense, the threat of epiphenomenalism is yet another consideration on the debit side of the cost-benefit ledger.

Suppose that the grain project were to be successfully completed, fully specifying what functional-representational roles are associated with phenomenal states, and how phenomenal states are physically subserved. As far as one can now tell, a view like Chalmers' would still retain both the theoretical advantages and the theoretical disadvantages just canvassed; further details about causal grain would not change things significantly. This approach would remain a live theoretical option.

### 3.3. Phenomenal states as explainable in a cognitively inaccessible way

A third kind of position, perhaps implicit in Nagel (1974) and explicitly defended in McGinn (1991), is one whose attractions come to light when one asks whether there might be a way to combine the theoretical advantages of each of the views just considered, while avoiding the specific theoretical disadvantages of each. There will be *some* kind of theoretical cost, to be sure, but a different kind.

The leading ideas are as follows. On one hand, the view asserts that qualia really do supervene logically on lower-order properties — and thus that these supervenience relations are explainable in a materialistically acceptable way, rather than being fundamental laws of nature. Thus, the three problems about consciousness mentioned in Section 2 receive solutions like those they receive under functional-representationalist approaches like Tye's.

On the other hand, the view asserts that although materialistic explanations of inter-level supervenience relations involving qualia exist *in principle*, nevertheless human beings, because of our innate cognitive limitations, are “cognitively closed” to these explanations; we are constitutionally incapable of formulating and articulating them. There is something about the nature of phenomenal properties that we cannot fully grasp — something that makes them logically supervenient on the physical, even though we humans cannot see why or how this relation of logical supervenience obtains. Thus, the view also exhibits the principal advantage of Chalmers’ view: it takes seriously, and accommodates, the hard-to-shake intuitive thought that no amount of theoretical information that we humans could acquire about physics, and/or neural organization, and/or the functional-representational roles of mental states, could possibly provide an explanation of why it should be that when certain neural states are instantiated in humans, they are correlated with the specific phenomenal qualities we know about from our own experience — rather than with different phenomenal qualities (inverted qualia), or with none at all (absent qualia).

On this view, it is indeed the case that there are no possible worlds — not even counter-nomological worlds — that are physically just like our actual world but in which phenomenal properties are instantiated differently (or are not instantiated at all). But we humans cannot grasp *why* this is not possible. Because of our innate cognitive limitations, we cannot understand what it is about phenomenal properties, and/or what it is about the underlying physical facts and properties, in virtue of which there actually lurks some logical-conceptual contradiction in the idea that the world could have been physically just the same as it is but different with respect to how qualia are instantiated. Physical-duplicate worlds with inverted or absent qualia do indeed *seem* logically possible. But on the view in question, this appearance is deceiving, being a by-product of our cognitive closure to the reasons for their impossibility.<sup>7</sup>

The attractions of the cognitive closure approach are significant: in essence, it combines the principal theoretical benefits of each of the earlier two approaches we have discussed in this section, while avoiding the principal theoretical disadvantages of each. But needless to say, these virtues come at an enormously high theoretical price: the hypothesis of human cognitive closure, vis-a-vis the putative materialistic explanation of physical-phenomenal supervenience relations. Given the human race’s spectacular history of success in scientifically explaining other aspects of the natural world, many people are bound to find it hard — or impossible — to believe that phenomenal con-

consciousness is inherently beyond the conceptual bounds of human scientific understanding.

Suppose, once again, that the grain project were to be successfully completed, fully specifying what functional-representational roles are associated with phenomenal states, and how phenomenal states are physically subserved. As far as one can now tell, a view like McGinn's would still retain both the theoretical advantages and the theoretical disadvantages just canvassed; further details about causal grain would not change things significantly. This approach would remain a live theoretical option.

#### 4. Concluding methodological postscript

To say that approaches to phenomenal consciousness like those we have considered would all remain live epistemic possibilities even if the grain project were successfully completed is not necessarily to say that there could never be any way to determine which approach — one of the three, or some other view entirely — is correct. In science, epistemic appraisal of competing hypotheses and theories is a matter of overall cost-benefit evaluation — what philosophers call “wide reflective equilibrium.” The same is true, we would maintain, for questions of the sort that arise in philosophy — although often these questions prove more recalcitrant, more open to the possibility that reasonable people will be able to disagree because they make different overall cost-benefit assessments. In principle, some particular philosophical account of consciousness — perhaps some variant of those we have considered, or perhaps some quite different view — might eventually turn out to have theoretical benefits so powerful, and theoretical costs so minimal, in comparison to the benefits and costs of competing accounts, that a strong consensus would emerge among rational, well-informed, people that this account is very probably true.

On the other hand, there is certainly no guarantee that rational convergence on these issues would eventually occur, even given sufficient theoretical ingenuity and empirical information. Indeed, the preceding discussion suggests that strongly divergent views about the nature of qualia would remain live theoretical options even in the ideal limit of theoretical inquiry. This seems to us a plausible conjecture.

Although in some ways it is disappointing to realize that carrying out the grain project would not resolve the philosophical problems about phenomenal consciousness, in other ways it is liberating. This very realization could cause

those working in cognitive science to be happy with less: just seek to understand the causal grain of phenomenal consciousness, and settle for that. Doing so would advance our knowledge significantly, even though it would leave the philosophical puzzles about phenomenal consciousness still open. It seems like an attractive, empirically tractable, and non-tendentious perspective from which to pursue the grain project, neither dismissive of cognitive science, nor of conceptual intuitions, nor of metaphysics; but, on the other side, not too metaphysically wistful. Evolutionary approaches fit naturally into this scaled-down program, since there are at least two important kinds of evolution-related question to ask about states with the distinctive causal grain associated with qualia. First, why did such states emerge on earth, in the evolution of humans and other terrestrial creatures? Second, with respect to evolutionary landscapes in general (including non-local ones for planets in distant solar systems, for A-life environments, etc.), do states with the relevant causal grain occur upon many of the fitness-peaks that reach an altitude where mentality resides?

The aspirations of Place and Smart were at least partly right. Our science of the conscious mind can get us at least somewhere — if not all the way to ontology. Anniversaries are if anything humbling.<sup>8</sup>

## Notes

\* This paper is entirely collaborative; order of authorship is alphabetical.

1. To say that there are three levels of causal grain is of course compatible with further stratification into sub-levels. For instance, it might prove useful and important to subdivide the functional-representational level of causal grain for phenomenal states into two components, one involving aspects of causal role that are near the sensory periphery and are largely modular, and the other involving aspects that are within central processing and are susceptible to causal influence from other kinds of mental states. Our three-level typology also does not preclude other ways — some perhaps fully or partially orthogonal to ours — of characterizing cognitive/neural systems in terms of levels of description or levels of analysis.
2. After the present paper was written we expanded this paragraph's line of thought into a full-length paper: Graham and Horgan (2000).
3. This logical supervenience would obtain because phenomenal properties are functional-representational properties, on such a view. Likewise, the mental state-types involved in "knowing what it's like" are also held to be functional-representational properties. But it should be noted that these claims about the nature of the relevant mental properties are compatible with saying that phenomenal *concepts* are not functional-representational concepts. Tye's account of phenomenal concepts illustrates the point. Although the state

constituting the exercising of such a concept is a functional-representational state, on his view, to possess and wield such a phenomenal concept is not to think about one's present experiential state *as* a functional-representational state (e.g., as a PANIC state), but rather involves the exercise of certain cognitive abilities of a perceptual-recognitional kind and/or an indexical kind.

4. One approach to qualia we will not discuss in the text is the view that phenomenal properties are identical to certain neurophysical properties. On one version of the psychophysical type-type identity theory (Lewis 1966, 1980, 1994), mental *concepts* are functional concepts, but the properties they express are neurophysical properties rather than multiply realizable functional-representational properties. (On this view, mental-state names are "nonrigid designators" that refer to different physical properties relative to different creature-kinds.) Much of what we have said in Section 3.1 about the theoretical benefits and costs of views that identify qualia with functional-representational properties is also applicable, *mutatis mutandis*, to a treatment of qualia and qualia-concepts like Lewis's. (For an argument that even Lewis's version of the type-type psychophysical identity theory does not sufficiently accommodate multiple realizability of mental properties, see Horgan in press.)

5. Chalmers also discusses a variant of the position, involving "proto-phenomenal" properties that might be instantiated very widely in nature, even at the level of subatomic phenomena. In the text we ignore this variant, for simplicity of exposition; much of what we say will be applicable to it too.

6. A panpsychist version of Chalmers' approach, which attributes proto-phenomenal properties to the fundamental subatomic constituents of matter, would avoid this problem — but only at the cost of departing much more radically from materialist metaphysics, and in a way that currently lacks any clear theoretical motivation.

7. In principle, an approach like McGinn's could be wedded to any of three different views about the ontology of qualia: (i) they are functional-representational properties; (ii) they are neurophysical properties; or (iii) they are properties of an ontologically distinct kind. But on any variant, the view holds that there is something about their nature that is not fully graspable by humans.

8. We thank John Tienison for helpful comments.

## References

- Bieri, Peter. 1995. Why is consciousness puzzling? In T. Metzinger (Ed.), *Conscious Experience* (45-60). Paderborn: Shonigh.
- Chalmers, David. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 200-219.
- Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford.
- Clark, Austen. 1998. Perception: Color. In W. Bechtel & G. Graham (Eds.), *Companion to Cognitive Science* (282-88). Oxford: Basil Blackwell.

- Davidoff, Jules. 1991. *Cognition Through Color*. Cambridge, MA: MIT.
- Dretske, Fred. 1995. *Naturalizing the Mind*. Cambridge MA: MIT.
- Flanagan, Owen. 1996. Is a science of the conscious mind possible? In his *Self Expressions: Mind, Morals, and the Meaning of Life* (12-31). New York: Oxford.
- Graham, George & Horgan, Terence. 2000. Mary Mary, quite contrary." *Philosophical Studies*, 99, 59–87.
- Horgan, Terence. 1987. Supervenient qualia. *Philosophical Review*, 94, 491-520.
- Horgan, Terence. 1993. From supervenience to superdupervenience: meeting the demands of a material world. *Mind*, 102, 555-86.
- Horgan, Terence. in press. Multiple reference, multiple realization, and the reduction of mind. In G. Preyer, F. Siebelt, and A. Ulfig (Eds.), *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*. Lanham, MD: Rowman and Littlefield.
- Jackson, Frank. 1982. Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127-36.
- Kim, Jaegwon. 1990. Supervenience as a philosophical concept. *Metaphilosophy*, 21, 1-27.
- Lewis, David. 1966. An argument for the identity theory. *Journal of Philosophy* 63, 17-25.
- Lewis, David. 1980. Mad pain and Martian pain. In N. Block (Ed.) *Readings in the Philosophy of Psychology. Volume 1* (216-22). Cambridge MA: Harvard.
- Lewis, David. 1994. Lewis, David: reduction of mind. In S. Guttenplan (ed) *A Companion to the Philosophy of Mind* (412-31). Oxford & Cambridge MA: Blackwell.
- Livingstone, Margaret & Hubel, David. 1987. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7, 3416-3468.
- Lycan, William. 1996. *Consciousness and Experience*. Cambridge MA: MIT.
- Macdonald, Cynthia. 1989. *Mind-Body Identity Theories*. London: Routledge.
- McGinn, Colin. 1991. *The Problem of Consciousness: Essays Towards a Resolution*. Oxford and Cambridge MA: Blackwell.
- McLaughlin, Brian. 1992. The rise and fall of British emergentism. In A. Beckermann, H. Flohr, & J. Kim (Eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: Walter de Gruyter (49-93).
- Nagel, Thomas. 1974. What is it like to be a bat? *Philosophical Review*, 83, 435-50.
- Place, Ullin. 1956. Is consciousness a brain process? In V. C. Chapell (Ed.), *The Philosophy of Mind*. Englewood Cliffs: Prentice-Hall (101-109). Reprinted from *British Journal of Psychology* (1956).
- Smart, J. J. C. 1962. Sensations and brain processes. In V. C. Chappell (Ed.), *The Philosophy of Mind*. Englewood Cliffs: Prentice-Hall (160-72). Reprinted from *The Philosophical Review* (1959).
- Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge MA: MIT.
- Zawidzki, Tadeuz & Bechtel, William. in press. Gall's legacy revisited: decomposition and localization in cognitive neuroscience. In C. Erneling & D. Johnson (Eds.), *Mind as a Scientific Object: Between Brain and Culture*. New York: Oxford.



## PART II

### Special adaptations



# Evolution, consciousness and the language of thought

James W. Garson  
University of Houston

The language of thought is an attractive thesis to account for the propositional nature of higher order consciousness. But is the requirement that symbolic processing be the product of evolution a reason against that view? This chapter examines research in genetic programming to try to determine whether evolution is compatible with classical faith in a language of thought. We conclude that potential problems with evolvability should concern classicists, and we point to experiments which will help resolve the issue.

## 1. Consciousness and the symbolic paradigm

It is widely remarked that our concept of consciousness is composed of several contrasting notions. Lloyd's taxonomy of this variety (1989: 179–185) divides it in four categories: which he refers to as sensation, perceptual awareness, reflection, and introspection. The last three abilities all involve propositional contents of some form or another. For example, according to Lloyd's usage, perceptual awareness entails seeing (hearing, smelling, tasting) *that* some proposition holds. I can have a *sensation* of a burnt pot roast, but I am not *perceptually aware* of it until I perceive *that* the pot roast is burnt. Similarly, reflection includes having memories *that* the pot roast was burnt, and thoughts *that* I should pay more attention to the oven in the future. Finally, introspection involves an awareness *that* I am having other conscious states such as remembering that the pot roast was burnt, or worrying about what to tell the guests.

The recognition that much of consciousness is propositional would seem to support a classical account of consciousness, where symbolic processing is an essential ingredient. If conscious states include propositional contents, then it would seem that these states should somehow contain symbolic elements (similar to sentences) that represent those contents. On that theory, the infor-

mation processing that makes consciousness possible would have to include operations on symbols that pick out features of the body, the world, and even other conscious states. So consciousness would require a language of thought, a system of representation that resembles spoken language.

Even if its propositional nature were to be laid aside, there are other motives for adopting a symbolic processing model of consciousness. From its inception, cognitive science has been dominated by the classical paradigm, which is deeply committed to the principle that all mental processing can be modeled on digital computation. Although the classical paradigm has fallen out of favor as an account of low level sensory processing, classicists have continued to argue that higher cognitive abilities such as reasoning, representing the world, communicating with others, and planning our actions are essentially symbolic (Fodor and Pylyshyn 1988). Since conscious life involves the exercise of abilities such as these, classicists will favor a symbolic processing account of consciousness.

Should this convince us that consciousness depends on a language of thought? I think not. There is a growing movement (especially among connectionists) that would deny the need for symbolic processing even at the highest cognitive levels (Clark 1990, 1993). Although classicists have, for example, registered strong doubts about whether connectionist systems are capable of the representation and processing found in human language, connectionists have risen to the challenge. The work of Jeff Elman (1991) is especially interesting because it offers a model of how the brain's language processor might operate without depending on the standard classical trappings.

So it is far from clear that the presence of "propositional" features of consciousness entail a symbolic processing model. This paper does not attempt to engage the controversy between classicists and connectionists directly. Instead, it will focus on a different line of attack against the symbolic account — an attack based on the evolution of software. To a first approximation the guiding intuition is this. The classical view, it would seem, requires that the brain contain complex programs that could not possibly be the result of natural selection. If consciousness depends on a brand of symbolic processing that evolution cannot produce, then how did it arise?

The thought that evolution poses a difficulty for a symbolic theory of consciousness is, so far, crudely formulated. The idea will need considerable refinement if a convincing case is to be made. To provide a framework for exploring the needed improvements, let us express the intuition in the form of a naive argument against classicism.

The Naive Argument:

1. Complex symbolic software can't evolve.
2. The classical theory of consciousness requires that the brain contain complex symbolic software.
3. This software is an evolutionary product.

So the classical theory of consciousness is wrong.

Although the Naive Argument is valid, there are serious problems with each of its premises. By discussing these problems, and revising the argument, I hope to provide a better picture of exactly how evolution might challenge the symbolic processing paradigm. In the end, I will argue that classicists should be concerned about evolvability. However, there are unresolved empirical issues at stake, so instead of trying to measure the depth of these difficulties *a priori*, I will offer suggestions for lines of experimentation that could help resolve the matter. The value of this paper consists not so much in arguing a case against classicism but in locating the tests in genetic programming research that could help evaluate the symbolic approach to consciousness.

## 2. Symbols and their interpreters

Before discussing the premises of the Naive Argument, it will be helpful to lay an important idea on the table. The classical thesis that the brain contains symbolic structures that effect its behavior presupposes the presence of a mechanism whose function it is to “read off” those symbols and place them in their computational roles. To see why this is so, consider a computer controlled by a program written in machine code. The codes are strings of 1s and 0s that represent the various basic instructions that the computer can execute. Machine code installed in a computer's memory would have no computational consequences at all if it weren't for the existence of circuitry (the CPU and memory bus) whose job it is to repeatedly fetch a line of machine code and to convert it to the corresponding machine events. If it were not for roles this circuitry creates for the codes, it is hard to see why we should count them as machine instructions in the first place. (They are, after all, nothing but sequences of 1s and 0s.)

Now consider a program written in some higher level language such as LISP. The story is more complicated, but the main idea is the same. For a LISP program to play a computational role in a computer, there must be a mecha-

nism that accesses lines of LISP code and converts each command into the appropriate sequence of computer events. Without a mechanism of this kind, LISP programs are computationally inert. The same idea applies to data as well as programs. Instead of a program, consider now a set of symbolic representations that form the data structure for a computer system, for example, the data on registration in classes at a university. Again, these data would not play the appropriate roles in storing information about the various classes if it weren't for the procedures that link these data to the right computational outcomes. Again, we need routines whose purpose it is to access the data and to endow them with causal roles appropriate for their use *as* data. The same moral applies to the language of thought. The very notion that the brain contains symbols of this language is parasitic on the idea that the brain contains procedures that treat those symbols in the appropriate ways.

In this paper, I will call the mechanism responsible for placing symbols in their computational roles *the interpreter* of those symbols. For a system that runs machine code, the interpreter is the (roughly) the CPU. For LISP, the interpreter includes the CPU and whatever software is responsible for converting LISP to machine instructions. (So on the usage of this paper, an interpreter is more than a translator from LISP to machine language. It includes as well the circuitry that converts machine language into action.) In the case of a data structure, the interpreter is the program or set of programs that read off the data structures and insure that they have the right computational effects.

The importance of interpretive mechanisms in the classical paradigm is easily overlooked; we have a tendency to presume that symbolic structures wear their identities on their own sleeves. But the symbol '*'*' would serve for zero in any system that endowed it with the role of a numeral that satisfies the appropriate arithmetic laws, for example:  $x+(=x$  and  $x*(=.$  So interpretive mechanisms are essential for individuating mere marks into symbolic categories. It follows that the classical hypothesis consists of a pair of complementary proposals. Symbolic processing presupposes symbols together with their interpreter. If classicists are right and the brain contains "sentences" of a language of thought that enter into its computations, it must also contain an interpreter that converts those sentences into the appropriate brain events.

One might worry that this characterization of the classical thesis takes too literal an interpretation of what it means for a brain to engage in symbolic computation. Is it really true that a symbolic processor, by its very nature, presupposes a data-structure — interpreter divide? Isn't there room for a less robust version of the classical point of view, for example one where the brain

does its cognitive work without the literal existence symbols on one side, and an interpreter on the other? Wouldn't it suffice if the brain behaves *as if* it contained symbols and their interpreter, even if the symbols were never explicitly "read off" by any interpretive procedures?

But this reconstruction of classicism goes too far. A system that acts *as if* it contains symbolic processing mechanisms needn't be symbolic processor at all. Classicism is not the theory that the brain presents the *appearance* of being a symbolic processor, for that view is compatible with an instrumentalist theory of symbolic processing, one which contends that thinking of the brain in those terms is merely useful or convenient. Presuming that classicists are realists with respect to the presence and causal role of symbolic structures in the brain, they must be committed to the idea that the brain actually contains the interpreters that put symbols in play.

Not only ought classicists take the presence of interpreters seriously, there is ample evidence that they actually do. One widely adopted benchmark for classicism found in discussions of connectionism is functional discreteness, the idea that there must be a fact of the matter as to which data are playing a role in the processing at any given time (Ramsey, Stich and Garon 1991: 206). This presupposes the existence of an interpreter which fetches different data at different times. Furthermore, classicists typically consider information about processing time to be relevant evidence for choosing between theories about the brain's architecture. Anyone who takes this point of view assumes that the brain contains an interpretive mechanism for which it makes sense to time the interval during which it reads off a line of symbolic code. There may be great latitude in the classical position concerning exactly what structures and mechanisms support symbolic processing, but even a minimal realist must admit that there *are* such structures and mechanisms, and that some explanation of how they came to be in the brain is required. So any classicist who takes the theory of evolution seriously should be prepared to account for natural processes that lead to the creation of both symbols *and their interpreter* in the brain.

### 3. Can complex software evolve?

It is time to examine the premises of the Naive Argument. Consider first premise (1), which says that complex software cannot be the product of natural selection. Let us first develop the simple intuition that motivates this idea, and then see how well it holds up. Digital software is notoriously fragile.

Randomly changing a single symbol in a program is likely to destroy its function. When the changed code is processed, the typical result is a crash that brings the whole system to its knees. The fact that very small changes in software generally lead to very large decreases in its “fitness” would suggest that useful software cannot be the product of the process of evolution. The fitness landscape for symbolic programs appears entirely too rugged. The space of possible programs is immensely large, and programs that are fit for a given function have neighbors in the space that are invariably useless for the same function. A program update that provides even a marginal improvement in fitness usually requires *coordinated* changes at many different points in the code. So a path through the space linking a series of programs with increasing fitness for a given task traverses long expanses of low-fitness desert. Evolution of fitter programs by mutation and selection only works if a significant proportion of the progeny that can be produced by random changes are more fit than their parents. If the probability is infinitesimal that a small number of random changes in a marginally adequate program produce a fitter version, then even the fittest parents are highly unlikely to ever produce progeny fitter than they are. The result is that the average fitness of the population stalls at a very low value because a random walk in the space of a program’s potential children so rarely encounters anything useful. The high salaries of good programmers is a testament to the idea that improving on good programs requires intelligence that is difficult to achieve by random search.

This intuition is sound as far as it goes. However, the point is not all that relevant because natural selection has tools more powerful than mutation to exploit. The argument that the fitness landscape for programs subject to mutation is poor presumes that single alleles (gene types) code for single symbols in the program. However, like other phenotypic traits, it is possible that many symbols are coded for together by a single gene (pleiotropy), or that many genes are responsible together for the presence of a single symbol (polygeny). Even if there is a one-one correspondence between allele and symbol, we need not assume that all combinations of alleles is equally likely to be chosen during the reproductive process. By skewing the distribution towards viable combinations, the reproductive mechanism can enrich the fitness landscape.

We now have clear empirical evidence that programs can be made to evolve, provided an appropriate model of reproduction is used. Koza (1993) has robust success in simulating the evolution of LISP programs that solve tasks as diverse as curve fitting, foraging for resources, and forecasting. It is instructive to examine how the problem of the rugged landscape for digital

programs is overcome in this work. Koza's model of reproduction depends primarily on crossover (though some mutation is used as well). This radically improves the nature and topology of the fitness landscape. In crossover, new progeny do not result from random mutation of single symbols, but rather from swapping well-formed portions of the parent's programs. (The LISP programming language in which Koza's programs are written is nicely adapted to this requirement, since programs are tree-structures, the components (sub-trees) of which are always well-formed expressions of LISP.) This allows the shuffling of functionally viable subroutines of various sizes so that larger program modules can be generated and shared in future generations. The fitness landscape is enriched by clearing it of ill-formed programs, and by making subroutines rather than individual symbols the targets of modification in future generations. The resulting landscape provides a much more promising environment for the discovery and maintenance of useful innovation.

Koza's work is an existence proof that crossover (with some mutation) *can* create useful programs that solve interesting problems. So premise (1) of the Naive Argument is clearly false as written. However, we must ask whether Koza's models overcome the problem of ruggedness in ways that would be faithful to the conditions to be faced by an evolving brain. I can think of five reasons for concern about this.

### 3.1. The average fitness may stall

Inspection of the fitness curves for most of Koza's experiments shows a wide gap between the fitness of the best program (to which he draws our attention), and the average fitness of all programs in a given generation. Koza terminates his experiments shortly after the best program in his populations meets a predetermined measure of success. While performance of the best program found is impressive, the average performance of the population at the time the experiments were concluded is not very good. Although curves of average fitness in these experiments rise through time, none of the experiments were carried out long enough to bring the average performance of the population to anything near an acceptable level. Some of these curves, in fact, show signs of stalling (Koza 1993: 142, 168, 301, 455, 541). It may be that the fitness landscape here is still too rugged for good evolvability of *the population*.

To provide a convincing case, we need to see models where average fitness at least approaches modest standards for success. Koza's use of the best individual in a population rather than the average fitness is not a reliable way to

measure evolvability. If occasional winners don't "breed true" because they are immersed in a sea of losers, and so are forced to cross with poor mates, their gains will be lost in the next generation. In a large population, a few new winners may emerge by luck to take the place of the winners in the old generation that failed to breed true. But such maintenance of an tiny fit "elite" in a population whose average fitness remains low and constant would offer cold comfort to a classicist challenged to explain how symbolic programs that support consciousness manage to evolve. It is not as if consciousness is reserved for an elite in the human population. What we need to model is the evolution of a population where the functionality of the programs is uniformly adequate. So further research is needed to determine whether crossover is capable of creating populations of programs whose average fitness meets reasonable standards for success.

### 3.2. The primitive command set must be selected by the experimenter

Koza's programs are constructed from a primitive set of commands that is selected *by the experimenter* on the basis of his intuitions about what will work to solve a given task. As Koza notes (1993: 90) selection of the right set of primitive commands is important, since the omission of some commands blocks success. There is no reason to believe that the problem of command selection can be solved by using a large pool of basic commands because the inclusion of extraneous commands for a given task typically degrades performance (Koza 1993: 583ff.) It is crucial then, that realistic models of the evolution of programs locate evolutionary processes that can select a good command set appropriate to a class of problems. I know of no research that directly addresses this important question.<sup>1</sup>

I urge that this is not a trivial matter, at least not if Koza's experiments are to carry weight in supporting the symbolic approach to consciousness. In the case of the language of thought, the program command set corresponds to the primitive concepts from which all other expressions of the language are constructed. But crafting a good set of concepts for a viable interaction with the world just is to solve a fundamental epistemological problem. The development of models where the command set spontaneously adjusts to the nature of the task being learned would go a long way towards calming the worry that the success of Koza's models is partly dependent on the intelligence of the experimenter in selecting the command set.

### 3.3. The results may not apply to large programs

The tasks Koza investigates are relatively simple ones, and the programs that evolve are significantly more complex than solutions arrived at by human programmers. I wonder whether similar success can be achieved for more complex tasks. Could the method produce a program as complex in control structure as (say) an operating system or a word processor? One worry is that as the size of the evolving programs increases, the number of near neighbors in the fitness landscape grows very quickly. If the ratio of viable programs among the neighbors does not grow as fast, we can expect evolution to stall as program size increases. I do not have any evidence that this must happen, however some researchers have reported problems in trying to “scale up” their results to more complex programs. To be sure, reproductive mechanisms (such as crossover) may insure that fitness of neighbors of large programs is sufficiently enriched. On the other hand, size may pose problems for evolvability. So I suggest that genetic programmers focus explicitly on experiments that investigate rates of increase of average fitness as a function of the size of the programs in the population.

### 3.4. Language interpreters are needed

A related problem is that Koza’s models simply presume the existence of mechanisms that endow primitive commands with their causal powers. As we have already argued, a LISP interpreter is needed to bring Koza’s programs “to life”. So a classical evolutionary story must explain how interpreters came to evolve *along with* the programs they interpret. Even if the fitness landscape for the development of LISP code to solve various problems is smooth enough, we have no evidence that I know of concerning the fitness landscape for the creation of interpreters. The most difficult problem is to explain how evolution could modify an interpreter so that a new syntactic innovation in the programming language can emerge. I suggest that work on this question should be a high priority in genetic programming research. I will return to this question in Section 7.3.

### 3.5. Protection from program errors may block evolution of interpreters

One of the secrets of Koza’s success is that the reproductive process used never generates programs that fail due to syntactic or semantical errors. The basic

commands are carefully written to insure that the output of any subprogram will be legal input to any other. (This is relatively easy in languages with such a simple syntax as LISP, but even here pains must be taken to make sure that all code is protected from such potential errors as division by zero (Koza 1993: 760).) There is a price to be paid for this control. If the reproductive process always respects the syntax of the language in which the programs are written, then this syntax, and the interpretive mechanism that reads it, cannot evolve. On the other hand, if mutations to the syntax are allowed, then the fitness landscape becomes more rugged, and the evolution of programs for a *given* syntax is threatened. More research is needed to understand whether a viable tradeoff can be forged that would allow both the successful evolution of programs and the evolution of their interpreters. Research is also needed on the evolvability of mechanisms that enforce well-formedness and modularity in the generation of evolving programs. That highly motivated and intelligent programmers have serious problems in maintaining such standards during program revision is one reason to believe that doing so may be a serious obstacle for evolution as well. If Mother Nature has already solved the problem, discovery of her methods would lead to a revolution in computer science. Faced with such a daunting task, evolution may very well have finessed the problem by exploiting non-symbolic mechanisms for the creation of consciousness.

I do not pretend to have shown that problems 3.1–3.5 are insoluble. My conclusion is only that Koza's work raises new questions that should be matters of further research. What we have so far are opening moves in what looks to be a very long and interesting project: to show that complex symbolic programs and their interpreters can truly evolve.

#### 4. Does classicism require that the brain contain complex software?

It is time to turn to premise (2) of the Naive Argument. In what sense is the classical account of consciousness actually committed to the idea that the brain contains complex software? Is it a sense in which obstacles to the evolution of programs would even be relevant reasons for rejecting classicism? Classicists are realists when it comes to the presence of symbolic structures in the brain. But one may well wonder whether this realism entails the presence of *evolvable programs* in the brain.

Suppose a classicist has written what she claims to be a program modeling the brain's cognitive activity. What does her claim entail about the contents of

the brain? A literal classicist would require that each line of her program be installed in the brain's memory, and that the brain execute the program by loading each line of the code into the interpreter where code is converted into action. Granted, some classicists *do* consider programs to be literally installed in the brain in this way, but the classical approach need not be saddled with such a strong view. It is surely not essential to the symbolic processing account of cognition that processing be entirely controlled by code stored in memory. As Fodor and Pylyshyn (1988: 60) note, it is impossible for programs to do all the work in any case. At some point in the process, an interpreter must retrieve the code and convert it into action. The assumption that the interpreter must also be controlled by stored code soon lands us in an infinite regress. So the classical account of cognition, by its very nature, recognizes the need for mechanisms in the brain that are not controlled by symbolic structures. Classicists need not be bound to any particular theory about how to draw the line between symbolic structures on one hand and their interpreter on the other. In particular, they need not draw the line according to the way it is drawn in the computer systems used to develop their cognitive models. After all, the use of stored programs to *model* cognition in classical AI research is merely a matter of convenience. It is a lot easier to write and revise programs than to build and modify special purpose brain-surrogate machinery. Accidents related to the computer machinery and languages presently available and convenient for AI research need not saddle the classicist with implausible views about the nature of *brain* processing. A *liberal* classicist may take the view that her AI programs are (for the most part) not explicitly represented in the brain, but instead describe brain processing. Nevertheless, if she is a classicist at all, she assumes that at least some symbolic structures play a causal role in the brain.

What structures should these be? A traditional way to identify where the classicist's realism extends is to draw the distinction between software structures that control processing and those devoted to explicitly representing the problem to be solved. Consider a classical chess playing program designed to model a human chess master. On a liberal construal of what is going on in the master's head, the processes of move evaluation and move selection might be treated as brain *procedures* that are not symbolically represented in the master's memory. But symbolic data stored in memory would still play a crucial role in representing (say) positions of pieces on the board, and future states of play.

Unfortunately it is not easy to give a clear account of the distinction between the data and the procedural portions for a given task. The nature and even usefulness of the data-procedure distinction has been in dispute, and in

artificial intelligence languages such as LISP, the distinction is systematically undermined. But liberal classicism need not be bound by any particular view about where to draw the line between symbolically represented data and the rest of cognitive processing. What matters is that a line is drawn somewhere. On one side we find data with syntactic structure written in the language of thought, and on the other, an interpreter for that language, i.e. a processing mechanism that is responsive to that symbolic structures of that language and endows them with causal consequences. We conclude, then, that premise (2) is essentially correct, at least on a liberal construal of the term “software”. The software at issue is whatever symbolic structures the classicists brings to the table to explain consciousness.

## 5. Must the brain’s software evolve?

Premise (3) of the Naive Argument claims that the complex software which classicists presume exists in the brain must be the product of evolution. Given the liberal classicist’s account of what the brain contains, this premise will appear quite implausible. The “software” at issue is not a program at all, it is a language of thought *data structure* that represents our conscious propositional contents. This information is primarily the product of sensation and learning, and not part of our genetic inheritance. So research on genetic programming concerning obstacles to the evolution of *programs* would seem entirely irrelevant to an evaluation of classicism.

However, this tactic for ducking the issue of evolvability takes too narrow a view. There are two separate considerations that bring evolvability back into the picture. Although the data that underlies conscious experience may not be the product of evolution, the interpretive mechanisms that structure and read that data so as to convert them to action presumably are. The classicist owes us some account of how an interpreter for a language of thought could evolve. If that is unlikely, classicism is in trouble. But an account of how interpreters were created during the history of the human species is not all we need to explain symbolic consciousness. We need also some account of the processes that are responsible for the installation of the data structures or conceptual schemes that consciousness employs for representing the world. Evolutionary processes that occur during the life of the organism may be required to craft the brain’s representational repertoire to meet the challenge of negotiating with the world (Edelman 1987). Here evolution operates over populations of

representations, rather than over populations of organisms. This process is a brand of learning, for it employs information obtained from an interaction with the world to improve the organisms ability to negotiate within it. Evidence that the symbolic processing account is incompatible with evolutionary mechanisms that under gird conceptual learning would be a serious objection.

## 6. Two revised arguments

The Naive Argument fails because it equivocates on the word “software.” To make premise (1) plausible (complex software can’t evolve), “software” means programs. But to make premise (2) plausible (classicists require that the brain contain complex software) “software” means data in a language of thought, at least on the liberal construal. However, two new considerations concerning evolution have been raised that may challenge the symbolic processing account of consciousness in new ways. Classicism requires that the brain acquire representations of the world in a language of thought, and it requires an interpreter that puts the data so generated into action. So the Naive Argument may be revised in two different versions. The first concerns the evolvability of symbolic interpreters in the genetic history of the species.

The Interpreter Argument:

- I1. It would be difficult for an interpreter for a language of thought to evolve.
  - I2. The classical theory of consciousness requires that the brain contain an interpreter for a language of thought.
  - I3. The language of thought interpreter must be an evolutionary product.
- So the classical theory is not likely to be true.

The second argument concerns the acquisition of symbolic representations during the life of the organism.

The Representation Argument:

- R1. It would be difficult for symbolic representations to evolve.
  - R2. The classical theory of consciousness requires that the brain acquire complex symbolic representations of the world.
  - R3. Acquisition of such representations depends on evolutionary processes.
- So the classical theory is not likely to be true.

Neither of these two arguments is convincing as it stands, because several of

their premises lack empirical support. But a discussion of reasons in their favor will demonstrate why classicists should consider the problem of evolution a potential threat.

## 7. The interpreter argument

The first premise of the Interpreter Argument (the claim that interpreters have difficulty evolving) is the most controversial and for that reason the most interesting. But before we launch a discussion of this question, we should consider the other two premises. The reasons for accepting premise (I2) (that the classical view entails that the brain contain a language of thought interpreter) have been thoroughly discussed in Section 2. Premise (I3), however, needs a few more words.

### 7.1. Why interpreters are an evolutionary product

Premise (I3) claims that the brain's interpreter must be an evolutionary product. One objection to (I3) is that the interpreter might be (at least partly) learned. This reply, however, is not compatible with a classical article of faith. According to it, the data-procedure divide is fixed by the difference between the information that can be learned (data) and the procedures written into the brain's hardware. On this view the procedural part (and hence the interpreter) just *is* what evolution provides. So for the majority of card-carrying classicists, (I3) is correct.

### 7.2. What a language of thought interpreter is like

Let us now turn to the first premise (I1) of the Interpreter Argument, the claim that interpreters for a language of thought are not liable to evolve. The claim is made more plausible once we consider the job such interpreters must perform. A classical theory of consciousness is committed to symbolic structures that correspond to the propositional contents of our beliefs, desires, thoughts and plans. This means that the interpreter for the language of thought must process a syntax that is fairly complex. Not only must the syntax provide some device for labeled bracketing to indicate agent, patient, time, place, manner, etc., it must also include sentential connectives (like 'not' and 'or') and quantifiers. Even more importantly, language of thought syntax must allow for iterable

intensional operators. I can be conscious, for example, that you *believe* that I am *planning* to *trick* you into *believing* that I like you. To represent this thought symbolically, some provision must be made to capture the way in which the intentional verbs 'believe', 'plan', and 'trick' are nested. Since there is no principled way to put a limit on the depth of this sort of nesting, a classical interpreter should handle a syntax with recursive rules that would allow the repeated embedding of complex thoughts.

### 7.3. Can language of thought interpreters evolve?

The syntax for the language of thought has to be complex enough to record higher-order intentional states. To appreciate the problems faced in the evolution of interpreters for such a language let us try to imagine how an interpreter that can parse symbol strings for propositional attitudes such as: 'Bill believes that John loves Mary', might have arisen from an interpreter that only parses base level expressions such as 'John loves Mary'. If there is to be selection pressure for the acquisition of an ability to parse symbolic structures of the novel form: Noun-IntentionalVerb-Sentence, then the ability to frame this new kind of thought must contribute in some way to the cognitive fitness of the organism. But if the installation of the new syntactic innovation has a use, there must be a set of intentional verbs ready that represent intentional states. To be of use in turn, these verbs must already have a role in the processing that binds symbols to the tasks which they are to perform. But how can these verbs be so bound unless they are interpreted, and if they are interpreted correctly, then the parser must already accept the rule that allows them to take sentential complements.

I am not overly concerned about the apparent circularity here. One can imagine a scenario, for example, where at first, the interpreter handles verbs for the attitudes that take no complements: 'Bill fears' (for 'Bill is afraid'). So the linkages that preexisted in a language without verbs that take sentential complements might serve as a foundation for the invention of full blown propositional attitude verbs as in 'Bill fears John loves Mary'. However, there are still reasons for concern. On the ordinary picture of how parsers work, an input sentence is either accepted and placed in a causal role, or rejected thus depriving it of causal powers. This means that there is no way for evolution to "tune up" the processing of 'fears' so that it is "ready" to take sentential complements in organisms where the interpreter rejects the form: Noun-IntentionalVerb-Sentence. So the first lucky organism for whom this novel form is accepted must also be lucky enough

to contain novel processing links that give “fears” the right causal role in the presence of a sentential complement. My point is that the installation of syntactic novelties that make an evolutionary difference would seem to require extensive and coordinated changes in the way the interpreter functions.<sup>2</sup> It is hard to see how such an installation could proceed by gradual evolutionary stages. A full package of adjustments must be in place all at once or the innovation will not increase fitness.

There is a way out of the dilemma. It is to allow the possibility that symbolic structures that violate a given syntax still have causal roles to play in the life of the organism. After all, users of English chronically violate the syntax of their language. But that does not block their ability to communicate. Similarly, the language of thought might allow slang and poor grammar. As Andy Clark (1990: 76) points out, we can expect evolution to deviate from what we would take to be a clean logical design for solving problems. On this hypothesis, violation of syntactic structure is (at least partly) ignored by the rest of the system. Viewed as a whole, there are no hard and fast processing rules that link symbolic forms to action (Horgan and Tienson 1989). The interpretive process would have the feature that the degree to which a symbolic form obeys a given syntax is the degree to which it is able to play a causal role in the system. On this picture, there is no fact of the matter as to exactly which syntactic or semantical rules are in play in the brain. Syntactic and semantical rules are fuzzy in the sense that they are present to a certain degree, and the degree to which they are present is a feature that many gradually evolve from one generation to the next.

An interpreter operating under fuzzy rules would allow evolution to work by gradual stages, eventually bringing the full resources of the language of thought into focus. This idea seems to rescue the classical paradigm from worries about evolvability. But it does not, for the degree of fuzziness measures the degree to which the language of thought hypothesis must be abandoned. The reason is that the degree to which a syntactic item counts as (say) a propositional attitude verb is the degree to which it is interpreted as a propositional attitude verb (Garson 1994: 26ff.). So an interpreter running fuzzy rules fuzzifies the syntactic categories of the language being interpreted. But I take it that it is crucial to classical realism that there is a fact of the matter as to what the syntactic structure of a give symbol string is supposed to be. When the behavior of the interpreter is not characterized by one strict set of rules, the individuation of representations into their linguistic roles and parts is threatened in concert. It appears then that the requirements for gradual progress in evolution would conflict with the classicist’s thesis that there are identifiable

features of the brain with linguistic properties such as fixed syntactic categories and identifiable constituent structure.

One should object that the argument presented so far leans heavily on an adaptationist assumption, namely that the interpreter cannot evolve unless it displays a selective advantage. But we know of traits that began their evolutionary histories as spandrels, i.e. as features that are accidentally linked to other traits which were the true targets of selection. It is also possible for traits to evolve due to accidents in evolutionary history without their conferring fitness at all. Interpreters might have evolved in this way. Still, the classicist who lodges this objection owes us at least the beginnings of a story about how such a complex and structured “organ” as an interpreter could have been a historical accident, or genetically linked to some other structure with adaptive advantage. I cannot imagine how such a story would go. Perhaps I am too pessimistic. Our lack of imagination has always been a serious obstacle to providing evolutionary explanations of complex structures (such as the eye). The actual course of evolution of complexity may be equally complex and devious. So I have not provided a convincing case that interpreters for a fully formed language of thought cannot evolve. However, classicists should be concerned about the difficulties faced in understanding how this could happen. Speculation and research on this issue should be central to a defense of the language of thought paradigm.

Classicists may object that we already know that interpreters can evolve. For evolution has already provided us with two examples: ribosomes that interpret the genetic code, and our brain’s interpreter of spoken language. If an interpreter of spoken language can evolve, then why not an interpreter for the language of thought? One might even suggest that evolution had to coopt pre-existing machinery for the language of thought in order to produce natural language.

But neither case is convincing. If anything, the nature of ribosomal interpretation actually provides evidence *against* the idea that interpreters evolve easily. If the evolution of interpreters were easy, then we should have expected the parallel evolution of many species of genetic interpreters, and the rapid creation of innovations in their function. But the ribosomal machinery has been one of the most strongly conserved of all biological systems both through time and across organisms. As to the evolution of spoken language, its use as evidence for the evolution of interpreters merely begs the question. What is at issue is whether cognitive phenomena such as thought and language are founded on the presence of a classical interpretive mechanism. If this is in doubt, so is the thesis that

natural language processing must depend upon internal symbolic representations which are read by an interpreted mechanism.

## 8. The representation argument

It is time to evaluate the Representation Argument.

- R1. It would be difficult for symbolic representations to evolve.
  - R2. The classical theory of consciousness requires that the brain acquire complex symbolic representations of the world.
  - R3. Acquisition of such representations depends on evolutionary processes.
- So the classical theory is not likely to be true.

This argument is not as convincing as the Interpreter Argument. Although there should be little to complain of concerning the second premise (R2), the other two premises are speculative, especially the last one (R3). Nevertheless, exploring the argument will help raise further issues for classicists.

### 8.1. Is the acquisition of representations an evolutionary process?

The last premise (R3) asserts that evolutionary processes are involved in producing conscious representations of the world. Although our knowledge about the matter is limited, speculation about the nature and function of consciousness does provides some support for the idea. For example, it has been suggested (Dennett 1996: 88) that consciousness representation has arisen because of the advantages to be gained from planning future actions by trying them out “off line”, evaluating their likely outcomes, and choosing for action the plan with the best outcome. A randomizing process would generate a set of initial candidates. Those would be evaluated and the weaker ones culled from the population leaving the best plans for a round of “reproduction”. The process would continue until some candidate met the standards for success, at which point it could be implemented, or even stored for later use in the appropriate circumstances. This is a speculative but interesting hypothesis concerning how a repertoire of representations might develop that could be used to better adapt to the world.<sup>3</sup>

Edelman (1987) suggests that evolutionary mechanisms working over populations of synapses or neural clusters are important in the creation of conceptual structures that are essential to understanding the environment.

Connectionists have also emphasized the importance of mechanisms that can account for concept development through weight (synapse) adjustment. Of course the brand of representational or synaptic evolution which is postulated to account for the creation of plans and concepts differs in many respects from evolution over populations of organisms. In particular, the reproductive processes at issue may bear little resemblance to mutation or crossover. Nevertheless, the evolution of representations by a process that molds a population based on its success or failure in interacting with the environment provides an attractive story about the nature of cognitive development. If symbolic representation can be shown to conflict with mechanisms of this kind, then this should weigh against the classical hypothesis.

## 8.2. Can symbolic representations evolve?

But is there any reason to think that symbolic representation and evolution do not mix? Is there any reason to support premise (R1)? Let us look again at Koza's work. It may seem that Koza's experiments are not directly related to the question before us, because he was worried about the evolvability of symbolic programs while we are wondering about evolvability of representational structures. But from a slightly more general point of view, our question and Koza's are similar. We are both wondering about the evolvability of symbolic structures designed to solve problems for which a notion of fitness can be provided. In showing that symbolic structures can evolve, his work tends to refute (R1).

In Section 3, we discussed five objections to the faithfulness of Koza's models in meeting the demands of genetic evolution. Some of those worries do not apply to the case of the evolution of representations within the organism. For example, consider the first complaint, namely that he selected the best member of his populations as his measure of success, rather than average fitness, which was not very high. This problem does not arise in the present context. We presume, after all, that the brain is able to measure fitness of candidate plans. Once a good plan is found, it will not matter if the population from which it is drawn has a low average fitness.

It would appear then that Koza's results are better crafted for defeating the Representation Argument. Nevertheless, some of the objections to Koza's work are still of special concern. One was that primitives for a given task were selected according to the intuitions of the experimenter, not by a process of evolution. A related worry was whether the techniques that lead to Koza's

success (a reproductive process that always ensured well formed programs) might make it impossible for interpreters to evolve. In the context of symbolic representation, these concerns are related to the problem of explaining how concepts could have developed. The primitives of the language of thought, are, after all, the concepts from which all other symbolic representations are constructed. Coming upon the right set of primitive concepts amounts solving the fundamental epistemological problem of how to represent the world correctly. This is no small feat; one would expect it to require a long and arduous evolutionary negotiation with the world. But if success in mating representations to the solution of a given task depends on rigidity in the set of primitive concepts available, then there is strong evolutionary pressure against the creation of new concepts, and it is hard to see how a sophisticated conceptual scheme could have developed.

If it is presumed that the interpreter is part of our genetic endowment, the mystery of concept development becomes particularly severe. The “hypotheses” in a candidate set available for solution to a problem are all written in the fixed vocabulary for the language of thought determined by the brain’s interpreter. It follows that novel concepts (concepts not expressible in that vocabulary) simply cannot be learned (Clark 1993: 33ff.). The problem is not resolved by assuming that the brain is equipped with a massive primitive vocabulary ready for “triggering” by interaction with the world (Fodor 1981: 279–292). Now the burden of explaining the development of our concepts is simply shifted from learning to the process of genetic evolution where solution is all the more difficult.

Connectionists have provided new insights into mechanisms that the brain might coopt for processing an evolving language without becoming trapped in a fixed primitive vocabulary. Connectionist models allow representation of information in the pattern of weights or connection strengths at the synapses between neurons. These weights may be gradually adjusted during a process of training so that a novel class of representational patterns is constructed for a given task. Since there is no *a priori* commitment to a fixed vocabulary, the system is free to adjust the nature of its representation in interacting with its environment. The genetic contribution to the brain’s structure is modeled in the initial settings of the weights. Those settings provide a raw conceptual structure which can still be fine tuned or even reformed by experience (Clark 1993: 33ff.) This connectionist vision of the brain systematically rejects the classical division between symbols and their interpreter as unhelpful. The fact that non-symbolic forms of processing show promise for resolving questions

we are raising should underscore the importance to classicists of research on evolvability of symbolic structure.

## 9. Conclusion

I do not claim to have closed the question as to whether evolution is an obstacle for the classical paradigm. Neither the Interpreter Argument, nor the Representation Argument is adequately supported. Nevertheless, I believe this paper shows that the ball is now in the classical court. If genetic programming research is to support a symbolic theory of consciousness or cognition in general, experiments suggested in this paper must be designed and carried out. The hardest task will be to demonstrate the evolution of interpreters for a language as complex as the language of thought. Work on dynamical conceptions of the mind shows that symbolic processing models correspond to a small region in the space of possible systems that could model the brain's function (van Gelder, 1995: 365). We have some evidence now that in other regions of this space, fitness landscapes better suited to the development of consciousness may be found. It is part of the classicist's burden to show it isn't so.

## Notes

I am grateful to Greg Mulhauser for criticisms that greatly improved this paper.

1. Koza does explore automatic creation of *defined* functions (1993: 534), but the issue here is creation of the appropriate set of *primitive* functions.
2. We don't mean the argument here to depend on an ability to make a strong syntax-semantics distinction. Our point is simply that the creation of novelty in interpretation requires that a lot be put in place all at once.
3. Not all of the candidates that played a role in the process would need to be conscious; perhaps only those that are the best or nearly best would qualify.

## References

- Clark, A. 1990. *Microcognition*. Cambridge: MIT Press.  
 Clark, A. 1993. *Associative Engines*. Cambridge: MIT Press.  
 Dennett, D. C. 1996. *Kinds of Minds*. New York: Basic Books.  
 Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books.

- Elman, J. L. 1991. "Distributed representations, simple recurrent networks, and grammatical structure". In D. Touretzky (ed.), *Connectionist Approaches to Language Learning*. Dordrecht: Kluwer, 91–122.
- Fodor, J. A. 1981. *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge: MIT Press.
- Fodor, J. A. & Pylyshyn Z. W. 1988 "Connectionism and cognitive architecture: A critical analysis". *Cognition* 28: 3–71.
- Garson, J. W. 1994. "No representations without rules: The prospects for a compromise between paradigms in cognitive science". *Mind & Language* 9(1): 25–37.
- Lloyd, D. E. 1989. *Simple Minds*. Cambridge: MIT Press.
- Horgan, T. & Tienson, J. 1989. "Representations without rules". *Philosophical Topics* 17: 147–174.
- Koza, J. R. 1993. *Genetic Programming*. Cambridge: MIT Press.
- Ramsey, W., Stich, S. and Garon, J. 1991. "Connectionism, eliminativism, and the future of folk psychology". In W. Ramsey, S. Stich & D. Rumelhart (eds), *Philosophy and Connectionist Theory*, Hillsdale, New Jersey: Erlbaum, 199–228.
- van Gelder, T. 1995. "What might cognition be if not computation?". *Journal of Philosophy* 91(7): 345–381.

# Why did evolution engineer consciousness?\*

Selmer Bringsjord, Ron Noel and David Ferrucci

Rensselaer Polytechnic Institute / T. J. Watson Research Center

## 1. The question

You, the two of us, the editor of this volume, Plato, Darwin, our neighbors — we have all not only been conscious, but we have also at some point decided to go on living in large part in order to *continue* to be conscious (of that rich chocolate ice cream, a lover’s tender touch, the glorious feeling of “Eureka!” when that theorem is finally cracked!). For us, consciousness is, to put it barbarically, a big deal. Is it for evolution? Apparently; after all, we evolved. But

Q1 Why did evolution bother to give us consciousness?

In this chapter we refine this question, and then proceed to give what we see as the only sort of satisfactory answer: one which appeals to the intimate connection between consciousness and creativity. Because we three confess to a deep-seated inclination to abide by the dictum “If you can’t build it, you don’t understand it,” we end by relating our answer to Q1 to two *radically* different attempts at engineering artificial creativity in our laboratory, the second of which — arguably the more promising one — is evolutionary in nature.

## 2. Taming the mongrel

Ned Block (1995) has recently pointed out that the concept of consciousness is a “mongrel” one: the term ‘consciousness’ connotes different things to different people — sometimes *radically* different things. Accordingly, Block distinguishes between

- phenomenal consciousness (P-consciousness)
- access consciousness (A-consciousness)
- self-consciousness (S-consciousness)
- monitoring consciousness (M-consciousness)

This isn't the place to carefully disentangle these four breeds. It will suffice for our purposes if we manage to get a rough-and-ready characterization of Block's quartet on the table, with help from Block himself, and some others.

Block describes the first of these phenomena in Nagelian fashion as follows:

So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are "what it is like" to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste and have pains. P-conscious properties include the experiential properties of sensations, feelings and perceptions, but I would also include thoughts, wants and emotions. (Block 1995, p. 230)

According to this explanation, the list with which we began the paper corresponds to a list of P-conscious states, viz.,

- savoring the taste of rich chocolate
- taking pleasure in a lover's caress
- experiencing the joy of cracking a proof

A-consciousness admits of more precise treatment; Block writes:

A state is access-conscious (A-conscious) if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, i.e., poised to be used as a premise in reasoning, and (2) poised for [rational] control of action and (3) poised for rational control of speech. (Block 1995, p. 231)

A-consciousness seems to be a property bound up with information-processing. Indeed, as one of us has explained elsewhere (Bringsjord 1997), it's plausible to regard certain extant, mundane computational artifacts to be bearers of A-consciousness. For example, theorem provers with natural language generation capability would seem to qualify with flying colors. In recent conversation, Block has gladly confessed that computational systems, by his lights, are A-conscious.<sup>2</sup>

S-consciousness is said by Block to mean "the possession of the concept of the self and the ability to use this concept in thinking about oneself" (Block 1995, p. 235). There is a famous family of cases (see e.g. Perry 1979) which seem to capture S-consciousness in gem-like fashion: Suppose that you are sitting in the Collar City Diner looking out the window at the passing traffic, when you notice the reflection of a man sitting alone in the diner — a man in a

rumpled tweed jacket who is looking out the window with a blank, doleful expression. On the basis of what you see, you affirm, to put it a bit stiffly, this proposition: “The man with the tweed blazer is looking blankly out a window of the Collar City Diner.” But suppose you then suddenly realize that the man in question is *you*. At this point you affirm a *different* proposition, viz., “I am looking blankly out a window of the Collar City Diner.” In this case we say that the indexical is *essential*; and, following Pollock, we say that beliefs that the second sort of proposition hold are *de se* beliefs. We can then say that an agent having *de se* beliefs, as well as the capacity to reason over them (after your epiphany in the diner you may conclude that you need to stop philosophizing and go home and sleep), enjoys S-consciousness.<sup>3</sup>

Block tells us that M-consciousness corresponds to at least three notions in the literature: inner perception, internal scanning, and so-called “higher order” thought. The third of these has been explicated and defended through the years by David Rosenthal (forthcoming, 1986, 1989, 1990*b*, 1990*a*). According to Rosenthal, a state is conscious (in some for-now generic sense of ‘conscious’) just in case it is the target of a higher-order thought. Courtesy of (Rosenthal forthcoming), the view can be put in declarative form:

Def 1  $s$  is a conscious mental state at time  $t$  for agent  $a =_{df}$   $s$  is accompanied at  $t$  by a higher-order, noninferential, occurrent, assertoric thought  $s'$  for  $a$  that  $a$  is in  $s$ , where  $s'$  is conscious or unconscious.<sup>4</sup>

Def 1, as the **higher-order theory** of consciousness, is often abbreviated as simply ‘HOT.’ What sorts of examples conform to HOT? Consider the state *wanting to be fed*. On Rosenthal’s view, this state *is* a conscious state — and the reason it is is that it’s the target of a higher-order thought, viz., the thought that I want to be fed. Rosenthal’s Def 1, of course, leaves open the possibility that the higher-order thought can be itself unconscious.

With Block’s quartet characterized, it’s time to return to Q1, the question with which we began.

### 3. The tough question

Notice first that we now have a specification of Q1 for each member of Block’s quartet. For example, we have

Q1<sub>s</sub> Why did evolution bother to give us S-consciousness?

This would work similarly for the other breeds of consciousness, giving us  $Q1_M$ ,  $Q1_A$ , and  $Q1_P$ . Next, we direct your attention to a variant of  $Q1$ , viz.,

Q2 Why might an AI engineer try to give her artifact consciousness?

as well as the corresponding  $Q2_X$ , with the subscript set to a member of Block's quartet, and 'consciousness' therein changed to 'X-consciousness.' Now, it seems to us that the following principle is true.

P1 If  $Q2_X$  is easily answerable, then so is  $Q1_X$ .

The rationale behind P1 is straightforward: If the AI engineer has a good reason for giving (or seeking to give) her robot consciousness (a reason, we assume, that relates to the practical matter of being a productive robot: engineers who for emotional reasons want to give robots consciousness are of no interest to us<sup>5</sup>), then there is no reason why evolution couldn't have given *us* consciousness for pretty much the same reason.

The interesting thing is that  $Q2_S$ ,  $Q2_M$ , and  $Q2_A$  *do* appear to be easily answerable. Here are encapsulations of the sorts of answers we have in mind.

For  $Q2_M$ : It should be wholly uncontroversial that robots could be well-served by a capacity to have higher-order thoughts about the thoughts of other robots and humans. For example, a robot working in a factory could exploit beliefs about the beliefs of the humans it's working with. Perhaps the robot needs to move its effectors on the basis of what it believes the human believes she sees in front of her at the moment. For that matter, it's not hard to see that it could be advantageous for a robot to have beliefs about what humans believe about what the robot believes — and so on. Furthermore, it's easy to see that robots could benefit from (correct) beliefs about their own inner states. (Pollock (1995) provides a wonderful discussion of the utility of such robotic beliefs.) And for certain applications they could capitalize on a capacity for beliefs about their beliefs about their inner states. (To pick an arbitrary case, on the way to open a combination-locked door a robot may need to believe that it knows that a memory of the combination is stored in memory.<sup>6</sup>)

For  $Q2_S$ : Even a casual look at the sub-field of planning within AI reveals that a sophisticated robot will need to have a concept of itself, and a way of referring to itself. In other words, a clever robot must have the ability to formulate and reason over *de se* beliefs. In fact, it's easy enough to adapt our diner case from above to make the present point: Suppose that a robot is charged with the task of making sure that patrons leave the Collar City

Diner in time for the establishment to close down properly for the night. If, when scanning the diner with its cameras, the robot spots a robot that appears to be simply loitering near closing time (in a future in which the presence of robots is commonplace), it will certainly be helpful if the employed robot can come to realize that the loitering robot is just itself seen in a mirror.

For Q2<sub>A</sub>: This question can be answered effortlessly. After all, A-consciousness can pretty much be *identified* with information processing. Any reason a roboticist might have for building a robot capable of reasoning and communicating will be a reason for building a robot with A-consciousness. For this reason, page after page of standard textbooks contain answers to Q2<sub>A</sub> (see e.g. Russell & Norvig 1994).

Question Q2<sub>p</sub>, however, is another story. There are at least two ways to see that P-consciousness is quite another animal. The first is to evaluate attempts to reduce P-consciousness to one or more of the other three breeds. One such attempt is Rosenthal's HOT. This theory, as incarnated above in Def 1, didn't take a stand on what breed of consciousness is referred to in the definiendum. Rosenthal, when asked about this, bravely modifies Def 1 to yield this definition:

Def 1<sub>p</sub> *s* is a P-conscious mental state at time *t* for agent *a* =<sub>df</sub> *s* is accompanied at *t* by a higher-order, noninferential, occurrent, assertoric thought *s'* for *a* that *a* is in *s*, where *s'* may or may not be P-conscious.

Unfortunately, Def 1<sub>p</sub> is very implausible. In order to begin to see this, let *s''* be one of our paradigmatic P-conscious states from above, say *savoring a spoonful of deep, rich chocolate ice cream*. Since *s''* is a P-conscious state, "there is something it's like" to be in it. As Rosenthal admits about states like this one:

When [such a state as *s''*] is conscious, there is something it's like for us to be in that state. When it's not conscious, we do not consciously experience any of its qualitative properties; so then there is nothing it's like for us to be in that state. How can we explain this difference? ... How can being in an intentional state, of whatever sort, result in there being something it's like for one to be in a conscious sensory state? (Rosenthal forthcoming, pp. 24–25)

Our question exactly. And Rosenthal's answer? He tells us that there are "factors that help establish the correlation between having HOTs and there being something it's like for one to be in conscious sensory states" (Rosenthal forthcoming, p. 26). These factors, Rosenthal tells us, can be seen in the case of wine tasting:

Learning new concepts for our experiences of the gustatory and olfactory properties of wines typically leads to our being conscious of more fine-grained differences among the qualities of our sensory states ... Somehow, the new concepts appear to generate new conscious sensory qualities. (Rosenthal forthcoming, p. 27)

But Rosenthal’s choice of wine tasting tendentious. In wine tasting there is indeed a connection between HOTs and P-conscious states (the nature of which we don’t pretend to grasp). But wine-tasting, as a source of P-consciousness, is unusually “intellectual,” and Def 1<sub>p</sub> must cover all cases — including ones based on less cerebral activities. For example, consider fast downhill skiing. Someone who makes a rapid, “on-the-edge” run from peak to base will have enjoyed an explosion of P-consciousness; such an explosion, after all, will probably be the main reason such an athlete buys expensive equipment and expensive tickets, and braves the cold. But expert downhill skiers, while hurtling down the mountain, surely don’t analyze the ins and outs of pole plants on hardpack versus packed powder surfaces, and the fine distinctions between carving a turn at 20 mph versus 27 mph. Fast skiers ski; they plunge down, turn, jump, soar, all at incredible speeds. Now is it really the case, as Def 1<sub>p</sub> implies, that the myriad P-conscious states  $s_1, \dots, s_n$  generated in a screaming top-to-bottom run are the result of higher-level, noninferential, assertoric, *occurrent* beliefs on the part of a skier  $k$  that  $k$  is in  $s_1$ , that  $k$  is in  $s_2$ ,  $k$  is in  $s_4$ , ...,  $k$  is in  $s_n$ ? Wine tasters do indeed sit around and say such things as that, “Hmm, I believe this Chardonnay has a bit of a grassy taste, no?” But what racer, streaking over near-ice at 50 mph, ponders thus: “Hmm, with these new parabolic skis, 3 millimeters thinner at the waist, the sensation of this turn is like turning a corner in a fine vintage Porsche”. And who would claim that such thinking *results in* that which it’s like to plummet downhill?

|   |   |   |   |
|---|---|---|---|
| C | F | P | Y |
| J | M | B | X |
| S | G | R | L |

Figure 1. Sample 3 × 4 array for backward masking

Def 1<sub>p</sub> is threatened by phenomena generated not only at ski areas, but in the laboratory as well. We have in mind an argument arising from the phenomenon known as **backward masking**. Using a tachistoscope, psychologists are able to present subjects with a visual stimulus for periods of time on the order of milliseconds (one millisecond is 1/1000th of a second). If a subject is shown a 3 × 4 array of random letters (see Figure 1) for, say, 50 milliseconds (msecs),

and is then asked to report the letters seen, accuracy of about 37% is the norm. In a set of very famous experiments conducted by Sperling (1960), it was discovered that recall could be dramatically increased if a tone sounded after the visual stimulus. Subjects were told that a high tone indicated they should report the top row, a middle tone the middle row, and a low tone the bottom row. After the table above was shown for 50 msec, to be followed by the high tone, recall was 76% for the top row; the same result was obtained for the other two rows. It follows that a remarkable full 76% of the array is available to subjects after it appears. However, if the original visual stimulus is followed immediately thereafter by another *visual* stimulus in the same location (e.g., circles where the letters in the array appeared; see Figure 2), recall is abysmal; the second visual stimulus is said to backward mask the first (the seminal study is provided in Averbach & Coriell 1961). Suppose, then, that a subject is flashed a series of visual patterns  $p_p$ , each of which appears for only 5 msec. In such a case, while there is something it is like for the subject to see  $p_p$  it is very doubtful that this is because the subject thinks that she is in  $p_i$ . In fact, most models of human cognition on the table today hold that information about  $p_i$  never travels “far enough” to become even a potential object of any assertoric thought (Ashcraft 1994).

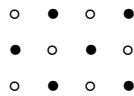


Figure 2. Sample visual  $3 \times 4$  array used as stimulus in backward masking experiments

So, for these reasons, Def  $1_p$  looks to us to be massively implausible. More generally, the point is that it's extraordinarily difficult to reduce P-consciousness to other forms of consciousness.

The *second* way to see that P-consciousness is much more recalcitrant than the other three breeds we have singled out is to slip again into the shoes of the AI engineer. Why would a roboticist strive to give her creation the capacity to experience that which it's like to, say, eat an ice cream cone? It would seem that any reason the robot might have for consuming chocolate fudge swirl in a waffle cone could be a reason devoid of any appeal to P-consciousness. (Perhaps the robot needs cocoa for fuel (other types of energy sources turned out to be a good deal more expensive, assume); but if so, it can be built to seek cocoa out when it observes that its power supply is low — end of story, and no need to

appeal to anything as mysterious as subjective awareness.) Evolution qua engineer should similarly find P-consciousness to be entirely superfluous.<sup>7</sup> Which is to say that we have moved from Q1 to what we call the “tough” question:

Q1<sub>P</sub> Why did evolution bother to give us P-consciousness?

This question can in turn be further refined through “zombification.”

#### 4. Zombifying the question

In order to zombify the tough question we need to restructure it so that it makes reference to zombies. The zombies we have in mind are *philosophers’* zombies, not those creatures who shuffle about half-dead in the movies.<sup>8</sup> Philosophers’ zombies, to use Stevan Harnad’s (1995) felicitous phrase, are bodies with “nobody home” inside. Such zombies are characters in a variation arising from a gedanken-experiment lifted directly out of the toolbox most philosophers of mind, today, carry with them on the job: Your brain starts to deteriorate and the doctors replace it, piecemeal, with silicon chip workalikes which flawlessly preserve the “information flow” within it, until there is only silicon inside your refurbished cranium.<sup>9</sup> John Searle (1992) claims that at least three distinct variations arise from this thought-experiment:

- V1 The Smooth-as-Silk Variation: The complete silicon replacement of your flesh-and-blood brain works like a charm: same mental life, same sensorimotor capacities, etc.
- V2 The Zombie Variation: “As the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior ... [You have become blind, but] you hear your voice saying in a way that is completely out of your control, ‘I see a red object in front of me.’ ... We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same” (Searle 1992, pp. 66–7).
- V3 The Curare Variation: Your body becomes paralyzed and the doctors, to your horror, give you up for dead.<sup>10</sup>

Scenario V2 seems to us to be clearly *logically possible* (a proposition written, using the possibility operator from modal logic, as  $\Diamond V2$ ); that is, V2 seems to us to be a scenario free from contradiction, perfectly coherent and conceivable. After all, Searle could, at the drop of a hat, provide a luxurious novel-length account of the scenario in question (or he could hire someone with the talents of a Kafka to do the job for him).<sup>11</sup>

Not everyone sees things the way we do. Daniel Dennett has registered perhaps the loudest and most articulate dissent. In fact, Dennett has produced an argument (based, by the way, on the Rosenthalian definition of M-consciousness discussed above) for  $\neg\Diamond V2$  in his recent *Consciousness Explained* (Dennett 1991, pp. 304–313).<sup>12</sup> One of us (Bringsjord) has formalized Dennett's argument (Bringsjord 1999), and found it wanting, but there isn't space here to recapitulate the argument. We don't ask that you regard this attempted refutation to be sound, sight unseen. We *do* ask that, for the sake of argument, you join the many prominent thinkers who affirm the likes of  $\Diamond V2$  (e.g., Dretske 1996, Block 1995, Chalmers 1996, Flanagan & Polger 1995, Harnad 1995). Moreover, we ask that you grant that V2 is *physically possible*, that is, that V2, though no doubt monstrously improbable, could come to pass without violating any laws of nature in our world. This seems to us to be a reasonable request to make of you. After all, why couldn't a neuroscience-schooled Kafka write us a detailed, compelling account of V2, replete with wonderfully fine-grained revelations about brain surgery and "neurochips"? Then we have only to change the modal operator to its physics correlate —  $\Diamond$  to  $\Diamond_p$ .<sup>13</sup> Each and every inch of the thought-experiment in question is to be devised to preserve consistency with neuroscience and neurosurgery specifically, and biology and physics generally. Our approach here is no different than the approach taken to establish that more mundane states of affairs are physically possible. For example, consider a story designed to establish that brain transplantation is physically possible (and not merely that it's logically possible that it's physically possible). Such a story might fix a protagonist whose spinal cord is deteriorating, and would proceed to include a step-by-step description of the surgery involved, each step described to avoid any inconsistency with neuroscience, neurosurgery, etc. It should be easy enough to convince someone, via such a story, that brain transplantation is physically possible.<sup>14</sup>

This last assertion will no doubt be challenged; we hear some readers saying: "Surely the two of you must be joking. To concede that such neural implantation is physically possible is (debatable) one thing, but to concede that

and that the result would be a V2-style zombie is absurd. In any case, if it is ‘perfectly reasonable’ to allow V2 as a physical possibility, then anything extra about logical possibility is superfluous, since the former entails the latter.”

The part of this objection which consists in observing that

$$\Diamond_p \phi \rightarrow \Diamond \phi$$

is certainly correct; this conditional is obvious and well-known. But why is it *absurd* that  $\Diamond_p V2$ ? Isn’t the progression to  $\Diamond_p V2$  quite sensible? We start with the story (from, e.g., Searle) designed to establish  $\Diamond V2$ ; and then when we look at this story we ask the question: What laws of nature are broken in it? Again, why can’t Kafka give us a novel *showing*  $\Diamond_p V2$ ?

Let us make it clear that we can easily do more than express our confidence in Kafka: We can provide an *argument* for  $\Diamond V2_1^{TM}$  given that Kafka is suitably armed. There are two main components to this argument. The first is a slight modification of a point made recently by Chalmers (1996), namely, when some state of affairs  $\psi$  seems, by all accounts, to be perfectly coherent, the burden of proof is on those who would resist the claim that  $\psi$  is logically possible.<sup>15</sup> Specifically, those who would resist need to expose some contradiction or incoherence in  $\psi$ . We think most philosophers are inclined to agree with Chalmers here. But then the same principle would presumably hold with respect to *physical* possibility: that is, if by all accounts  $\psi$  seems physically possible, then the burden of proof is on those who would resist affirming  $\Diamond_p \psi$  to indicate where physical laws are contravened.

The second component in our argument comes courtesy of the fact that V2 can be modified to yield  $V2^{NN}$ , where the superscript ‘NN’ indicates that the new situation appeals to artificial neural networks, which are said to correspond to actual flesh-and-blood brains.<sup>16</sup> So what we have in mind for  $V2^{NN}$  is this: Kafka really knows his stuff: he knows not only about natural neural nets, but also about artificial ones, and he tells us the sad story of Smith — who has his neurons and dendrites gradually replaced with artificial correlates in flawless, painstaking fashion, so that information flow in the biological substrate is perfectly preserved in the artificial substrate ... and yet, as in V2, Smith’s P-consciousness withers away to zero while observable behavior runs smoothly on. Now it certainly seems that  $\Diamond_p V2^{NN}$ ; and hence by the principle we isolated above with Chalmers’ help, the onus is on those who would resist  $\Diamond_p V2^{NN}$ . This would seem to be a *very* heavy burden. What physical laws are violated in the new story of Smith?

We are now in position to “zombify”  $Q1_p$ :

Q1<sub>P</sub><sup>Z</sup> Why did evolution bother to fashion us, bearers of P-consciousness, rather than zombies, creatures — courtesy of the right sort of information processing working in unison with sensors and effectors — having our behavioral repertoire, but lacking our inner lives?

## 5. Creativity as an Answer

There are at least three general ways to answer Q1<sub>P</sub><sup>Z</sup>:

- A1 “Look, evolution does allow for outright accidents, so maybe P-consciousness is just an adventitious ‘add on’ having no survival value ‘payoff’ — in which case there would be no particular reason why evolution fashioned us rather than zombies.”<sup>17</sup>
- A2 “The answer is that P-consciousness has some definite function, and though this function can be carried out by mere information processing (suitably symbiotic with the outside environment on the strength of sensors and effectors), evolution took the more interesting route.”
- A3 “P-consciousness has a definite function, yes, but one arguably not replicable in any system based on the standard information processing built into the gedanken-experiments taken to substantiate  $\Diamond V2$ .”

A1 is really not an answer to Q1<sub>P</sub><sup>Z</sup>; and as such it’s profoundly unsatisfying (if the informal poll we’ve taken is any indication). It even seems downright bizarre to hold that the phenomenon that makes life worth living (Wouldn’t you be depressed upon hearing that starting five minutes from now you would have the inner life of a slab of granite?) is a fluke. The second answer, A2, is given by a respondent who hasn’t grasped the problem: After all, if the function of P-consciousness can be carried out by computation, then why didn’t evolution take the programming route? This question is just Q1<sub>P</sub><sup>Z</sup> all over again, so A2 gets us nowhere.<sup>18</sup>

A3 is the answer we favor. This means we have to be prepared to step up to the challenge and show that certain behaviors do correspond to P-consciousness, and that obtaining these behaviors from ordinary computation isn’t possible. What behaviors might qualify? In a word: creativity. We conclude this section by providing reason to believe that P-consciousness’ role in us is to enable creativity. In the following section, when we discuss our engineering work, we return to the view that creativity requires more than standard information processing.

One of us (Bringsjord 1997) has recently tried to reconstruct a Searlean argument for the view that a — perhaps *the* — function of P-consciousness is to enable creativity. Searle's argument is enthymematic; its key hidden premise is a principle which unpacks the common-sense idea that if the advent of a psychological deficiency coincides with a noteworthy diminution of a person's faculty, then it's a good bet that the diminution is causally linked with the deficiency. With (a slightly more sophisticated version of) this principle (P2), we can produce a Searlean argument that is formally valid in first-order logic. The argument runs as follows.<sup>19</sup>

#### Argument 1

- P2 If S loses  $x$  over an interval of time during which S loses the ability to  $\phi$ , and there are substantive reasons to think  $x$  is centrally employed when people  $\phi$  (in part because (i) attempts to replicate  $\phi$ -ing in systems lacking  $x$  have failed, and show no appreciable promise of succeeding in the future; and (ii) subjects report that they need  $x$  in order to  $\phi$ ), then a function of  $x$  is to at least facilitate  $\phi$ -ing.
- (1) S loses  $x$  over an interval of time during which S loses the ability to  $\phi$ .
  - (2) There are substantive reasons to think  $x$  is centrally employed when people  $\phi$  (in part because attempts to replicate  $\phi$ -ing in systems lacking  $x$  have failed, and show no appreciable promise of succeeding in the future ... that they need  $x$  in order to  $\phi$ ).
- ∴ (3) A function of  $x$  is to facilitate  $\phi$ -ing. P2, 1, 2

Victorious instantiations of this schema seem to us to be at hand. (If  $x$  = 'P-consciousness,' and  $\phi$  = 'write belletristic fiction,' then it turns out that one of us has elsewhere (Bringsjord & Ferrucci 2000) explicitly defended the relevant instantiation. The basic idea underlying the instantiation is that creativity, for example the creativity shown by a great dramatist, requires P-consciousness.<sup>20</sup> The defense capitalizes on P2's parenthetical by including an observation that AI has so far failed to produce creative computer systems.

You may ask, "Yes, but what evidence have you for P2?" We haven't space to include here all of the evidence for this principle. Some of it is empirical (e.g., (Cooper & Shepard 1973)); some of it is "commonsensical." Evidence of the latter sort is obtained by remembering that all of us have experienced unplanned intervals of "automatism." To repeat the familiar example, you're driving late at night on the interstate; you're 27 miles from your exit ... and the next thing you know, after reverie about a research problem snaps to an end, you are but *seventeen* miles from your turnoff. Now, was there anything it was

like to drive those ten mysterious miles? If you're like us, the answer is a rather firm "No" (and we daresay the real-life cases are plentiful, and not always automotive). Now, why is it that such episodes invariably happen when the ongoing overt behavior is highly routinized? Have you ever had such an episode while your overt behavior involved, say, the writing of sentences for a short story, or the writing of inferences toward the proving of a theorem? These are rhetorical questions only, of course. But surely it's safe to say that P2 is no pushover, and that Argument 1 constitutes a respectable case for the view that a function of P-consciousness is to enable creative cognition.

## 6. Engineering creativity

The foregoing discussion, quite theoretical in nature, is related to two attempts on our part to "engineer creativity." The first attempt is, as we say, a **parameterized** one based in formal logic: a system designed to generate interesting short-short<sup>21</sup> stories with help from pre-set formalizations designed to capture the *essence* of such stories. As we explain, this attempt (at least as it currently stands) seems to be threatened by the apparent fact that the space of all interesting short-short stories cannot be captured in some pre-set formalism — the space may be, in the technical sense of the term, **productive**. (For now — we furnish the technical account below —, understand a productive set to be one whose membership conditions can't be formalized in computational terms.) Our second engineering attempt is evolutionary in nature, and steers clear of any notion that creativity consists in "filling in" the parameters in some pre-defined account of the desired output.

### 6.1. BRUTUS: A parameterized logicist approach

We are members of the Creative Agents project at Rensselaer,<sup>22</sup> which extends and enhances the Autopoeisis Project. Autopoeisis, launched in 1991 with grants from the Luce Foundation, with subsequent support from Apple Computer and IBM, is devoted to building an artificial storyteller capable of generating "sophisticated fiction." (A snapshot of the project's first stage was provided in (Bringsjord 1992).) Though we confess that no such AI is currently on the horizon (anywhere), the BRUTUS system suggests that the dream driving Autopoeisis may one day arrive. (BRUTUS is the overall system architecture. The first incarnation of that architecture is BRUTUS<sub>1</sub>. Details may be found in

(Bringsjord & Ferrucci 2000).) Though they aren't exactly Shakespearean, BRUTUS is able to produce stories such as the following one.

#### BETRAYAL.1

Dave Striver loved the university. He loved its ivy-covered clocktowers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth. He also loved the fact that the university is free of the stark unforgiving trials of the business world — only this *isn't* a fact: academia has its own tests, and some are as merciless as any in the marketplace. A prime example is the dissertation defense: to earn the PhD, to become a doctor, one must pass an oral examination on one's dissertation. This was a test Professor Edward Hart enjoyed giving.

Dave wanted desperately to be a doctor. But he needed the signatures of three people on the first page of his dissertation, the priceless inscriptions which, together, would certify that he had passed his defense. One of the signatures had to come from Professor Hart, and Hart had often said — to others and to himself — that he was honored to help Dave secure his well-earned dream.

Well before the defense, Dave gave Hart a penultimate copy of his thesis. The professor read it and told Dave that it was absolutely first-rate, and that he would gladly sign it at the defense. They even shook hands in Hart's book-lined office. Dave noticed that Hart's eyes were bright and trustful, and his bearing paternal.

At the defense, Dave thought that he eloquently summarized Chapter 3 of his dissertation. There were two questions, one from Professor Rogers and one from Dr. Teer; Dave answered both, apparently to everyone's satisfaction. There were no further objections.

Professor Rogers signed. He slid the tome to Teer; she too signed, and then slid it in front of Hart. Hart didn't move.

"Ed?" Rogers said.

Hart still sat motionless. Dave felt slightly dizzy.

"Ed, are you going to sign?"

Later, Hart sat alone in his office, in his big leather chair, saddened by Dave's failure. He tried to think of ways he could help Dave achieve his dream.

BRUTUS can generate stories like this one because, among other reasons, it "understands" the literary concepts of self-deception and betrayal via formal definitions of these concepts. (The definitions, in their full formal glory, and the rest of BRUTUS' anatomy, are described in (Bringsjord & Ferrucci 2000).) To assimilate these definitions, note that betrayal is at bottom a relation holding between a "betrayor" ( $s_r$  in the definition) and a "betrayed" ( $s_d$  in the definition). Then here is a (defective) definition that gives a sense of BRUTUS' "knowledge."

Agent  $s_r$  betrays agent  $s_d$  iff there exists some state of affairs  $p$  such that

- 1  $s_d$  wants  $p$  to occur;
- 2  $s_r$  believes that  $s_d$  wants  $p$  to occur;
- 3  $s_r$  agrees with  $s_d$  that  $p$  ought to occur;
- 4' there is some action  $a$  which  $s_r$  performs in the belief that thereby  $p$  will *not* occur;
- 5'  $s_r$  believes that  $s_d$  believes that there is some action  $a$  which  $s_r$  performs in the belief that thereby  $p$  will occur;
- 6'  $s_d$  wants that there is some action  $a$  which  $s_r$  performs in the belief that thereby  $p$  will occur.

BRUTUS also has knowledge of story structures in the form of story grammars. For example, BETRAYAL.1 conforms to the following story grammar (Table 1), taken from Thorndyke (1977). That which flanks '+' comes sequentially; the asterisk indicates indefinite repetition; parentheses enclose that which is optional; brackets attach to mutually exclusive elements.

Table 1. Story grammar of BETRAYAL.1 (Thorndyke 1977)

| Rule No. | Rule                                                                                                                        |
|----------|-----------------------------------------------------------------------------------------------------------------------------|
| (1)      | Story $\rightarrow$ Setting + Theme + Plot + Resolution                                                                     |
| (2)      | Setting $\rightarrow$ Characters + Location + Time                                                                          |
| (3)      | Theme $\rightarrow$ (Event)* + Goal                                                                                         |
| (4)      | Plot $\rightarrow$ Episode*                                                                                                 |
| (5)      | Episode $\rightarrow$ Subgoal(+ Attempt* + Outcome                                                                          |
| (6)      | Attempt $\rightarrow$ $\left\{ \begin{array}{l} \text{Event*} \\ \text{Episode} \end{array} \right.$                        |
| (7)      | Outcome $\rightarrow$ $\left\{ \begin{array}{l} \text{Event*} \\ \text{State} \end{array} \right.$                          |
| (8)      | Resolution $\rightarrow$ $\left\{ \begin{array}{l} \text{Event} \\ \text{State} \end{array} \right.$                        |
| (9)      | $\left. \begin{array}{l} \text{Subgoal} \\ \text{Goal} \end{array} \right\} \rightarrow \text{Desired State}$               |
| (10)     | $\left. \begin{array}{l} \text{Characters} \\ \text{Location} \\ \text{Time} \end{array} \right\} \rightarrow \text{State}$ |

Is Brutus creative? Perhaps not.<sup>23</sup> After all, Brutus is capable of generating only a small portion of the space  $I$  of all interesting short-short stories: the formalisms that make up BRUTUS' soul seem to leave out much of this space.<sup>24</sup> Even a future incarnation of BRUTUS, BRUTUS<sub>*n*</sub>, that includes knowledge of *all* presently deployed literary concepts (unrequited love, friendship, revenge, etc.), all known story grammars, and so on — even such a system, we suspect, would

inevitably fail to capture the essence of *I*. Our suspicion is based on the intuition that *I* is productive, in the technical sense: A set  $\Phi$  is productive if and only if (i)  $\Phi$  is classically undecidable (= no program can decide  $\Phi$ ), and (ii) there is a computable function  $f$  from the set of all programs to  $\Phi$  which, when given a candidate program  $P$ , yields an element of  $\Phi$  for which  $P$  will fail. Put more formally (following (Dekker 1955), (Kugel 1986), (Post 1944)):

- $\Phi$  is productive if and only if  $\exists f [f: \text{TM} \rightarrow \Phi \wedge \forall m \in \text{TM} \exists \phi \in \Phi (f(m) = \phi \wedge m \text{ cannot decide } \phi)]$ .

Put informally, a set  $\Phi$  is productive if and only if it's not only classically undecidable, but also if any program proposed to decide  $\Phi$  can be counter-exemplified with some element of  $\Phi$ . Evidence for the view that *I* is productive comes not only from the fact that even a descendant of BRUTUS<sub>1</sub> would seem to leave out some of *I*, but from what the Autopoeisis team has experienced when attempting an outright definition of the locution 's is an interesting short-short story.' Every time a definition of this locution is ventured, someone comes up with a counter-example. (If it's proposed that all members of *I* must involve characters in some form of conflict, someone describes a monodrama wherein the protagonist is at utter, tedious peace. If it's proposed that all members of *I* must involve one or more key literary concepts (from a superset of those already mentioned: betrayal, unrequited love, etc.), someone describes an interesting story about characters who stand in a novel relationship. And so on.) So a key question arises: What about attempts to engineer creativity *without* trying to pre-represent the space of the artifacts desired?

## 6.2. A non-parameterized evolutionary approach

We all know that processing and representation are intimately linked.<sup>25</sup> So, given this general fact, how does one get the representation correct for creativity? If the representations used by BRUTUS are inadequate, what might work? What about creative processes in evolutionary computation? And what about marrying a new mode of representation to processing that is evolutionary in character, rather than (as in the case of BRUTUS) processing that is essentially theorem proving?

The first response to such questions is to observe that present systems of evolutionary computation would seem to squelch their capacity for creative processes because of their own impoverished representation schemes. Consider, for example, one of the standard processes of an evolutionary system: epigenesis — the process in an evolutionary system that translates a given

genotype representation into the phenotype representation.<sup>26</sup> The present systems of evolutionary computation use parameterized computational structures to accomplish epigenesis: such systems use the genotype representation to encode levels of the parameters, and the phenotype representation becomes the direct output. The use of parameterized computational structures to formalize that which is within the reach of an evolutionary system works just as poorly as the methods behind BRUTUS. Pre-set parameterized computational structures are limited in their ability to map the levels of a finite number of parameters (even if the value of the parameters are infinite) onto a complete set of designs. Just as BRUTUS, for reasons discussed above, is tackling a space that probably can't be "mastered" via parameterized computational structures, so too present evolutionary systems face the same uphill battle. If we consider not *I*, the space of interesting short-short stories, but the space of interesting *paintings*, then Hofstadter has probably given us the key paper: he has argued in (Hofstadter 1982) that when Knuth (1982) applied such a system to the design of fonts, he was bound to fail: the system could only cover a small portion of the set of all 'As,' for example. (The wildly imaginative As drawn in Hofstadter's paper are reason enough to read it; see Figure 3.) The use of parameterized computational structures, if you will, requires describing the quintessence of the hoped-for design before the (evolutionary) system can seek a solution.



Figure 3. Various letter As

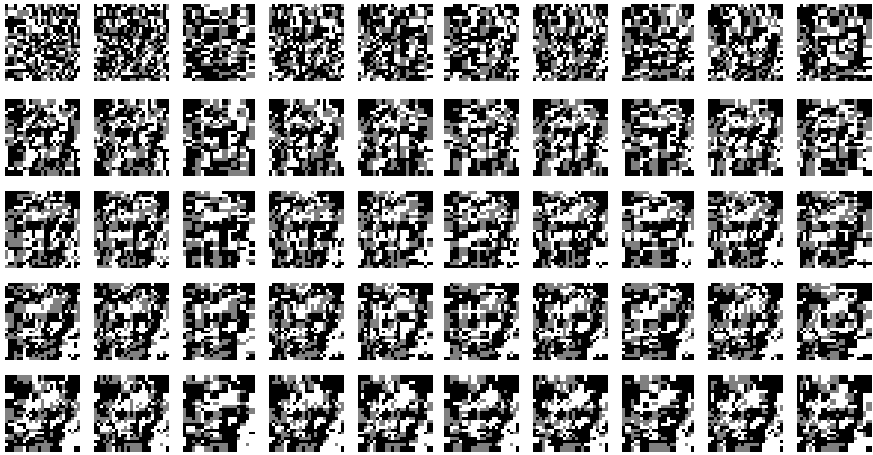


Figure 4. Evolution of a portrait of Lincoln



Figure 5. Evolved portrait of Lincoln

In our opinion, a system is creative only if it can somehow capture the quintessence of the space from which a particular design is to come. On this view, creativity occurs outside of (at least most) current evolutionary systems. This is so because such systems, like the logic-based BRUTUS, are based on a pre-selected and bounded design space.

One of us (Noel), working with Sylvia Acchione-Noel, has created a system — MONA-LISA — that changes things: MONA-LISA uses an information-based representation that affects the resolution of the system (the ability to describe a wave form) but forces no feature-level dimension on the system. The representation is based on atomic or molecular representation, similar to the notions of atomic or molecular decomposition by Fourier analysis or wavelets (Meyer

1993). This new evolutionary system evolves patterns of pixels into any desired image. The use of a sub-feature representation requires the units to evolve simultaneously *en masse* to generate both the features and the configuration of an image. The evolved image is not constrained at the feature level and can encode a dog, a tree, a car, or, just as easily, a face. (If there is anything to the view that the set of all As makes a productive set, such a view in connection with the set of all faces is hardly implausible. And of course we believe that this view about As is quite plausible. We direct readers to (Hofstadter 1982) for the data and evidence. This paper includes an interesting collection of faces. However, the resolution of the images is affected by the number of pixels or the atomic representation used in the system. For instance, the “portrait” of Abraham Lincoln shown in Figure 5 was evolved in a 25-by-25 pixel space which can only represent images with 12.5 lines of resolution or less.

Traditional evolutionary programs use parameterized modeling to map between the genotype and the phenotype. The use of such models results in a direct mapping between the set of genes that comprise a parameter, and the level of the feature that the parameter models in the phenotype. There is no noise, no pleiotropy, no polygeny in the mapping. The system can only create objects within the parameterized space, and all objects evolved are members of that space. For instance, one might model a face by making a model in which some genes selected types of eyes, nose, lips, ears, face shape, hair, etc., and other genes arrange the features within configurations of a normal face. If one were using such a system, the population of the first generation would all look like faces. One would select the faces that are most like the intended face and use them for breeding the next generation. The impact of this is that one must constantly compare faces with the intended face to make decisions. One face might have the eye shape and size right, while another might have the distance between the eyes correct.

In our technique, the desired or intended image is considered the signal, and all other images are considered noise. The elicitation of the image is done by a biased selection of the objects that generate the greatest recognition, or signal to noise ratio, in each generation. In keeping with our example, consider evolving a face image. The first generation is pure noise, or in other words, all possible images are equally likely. The task of the evolver is to select the images for breeding that have the greatest signal (most like the face to be evolved) and therefore the least noise. At first, the probability of any image looking like a face, any face, is extremely unlikely. Most images look like the snow on a TV tuned to a channel without a station. However, one can select images whose

pixels might give the impression of something rounded in the middle, dark in the area of hair, or light where cheeks may be. Since the images selected to parent the next generation have more of the signal and consequently less noise, it will give rise to a population of images whose mean signal to noise ratio will be greater than the previous generation. Eventually the signal to noise ratio is strong enough to elicit the intended image.

In our evolutionary system one only sees the face that one seeks. The face is seen in different amounts of noise, from high to low. In high noise conditions, only the lowest spatial frequency in low contrast can be imaged. As the image evolves, the level and contrast of detail increases. The end image is much like seeing the intended face in a cloud in that there are no distinct features, but a holistic percept. While the quality of the image is at present limited in resolution (about 15 to 20 lines of resolution), the reader should be reminded that the system can evolve any image that the pixels can represent. One could just as easily select to evolve a horse, a ball, a tree, etc. In the traditional approach one is limited to a domain and would require a new model for each new type of object to be evolved.

MONA-LISA is at present a two-dimensional system used by humans to create images, but it can be generalized to other domains. Humans were chosen to perform the judging and selecting.<sup>27</sup> Evolutionary systems are not necessarily, by themselves, creative. Creativity in evolution presumably occurs through the interaction of the objects and their environment under the forces of natural selection. However, as we have said, evolutionary systems *can* preclude creativity — by limiting and bounding the phenotype.<sup>28</sup> MONA-LISA gives complete representational power to the user. It starts from scratch or from randomness, and it is truly general since the image's features are not pre-determined.<sup>29</sup>

### 6.3. What next?

Where do we intend to go from here? Both BRUTUS and MONA-LISA will live on, both to be gradually improved. In the case of BRUTUS, we are only at BRUTUS<sub>1</sub>, and we will work sedulously as good engineers, clinging to a workday faith that some more robust system *can* capture all of *I*. (When the workday ends, we will permit our suspicion that *I* is productive to have free rein.) In connection with MONA-LISA, we plan to

1. Attempt to substitute artificial agents (treated as “percept to action” functions (Russell & Norvig 1994)) for the role of the human in the image elicitation process.
2. Transfer the image elicitation approach to the domain of stories (so BRUTUS can get some help!).
3. Explore more carefully the complexity implications of our image elicitation work.<sup>30</sup>

## 7. Conclusion

Is Q1<sub>P</sub><sup>Z</sup> answered? Does our A3, bolstered by our engineering, really satisfy? Probably not — because not all will be convinced that creativity calls for “super”-information processing beyond what a zombie is allowed in scenarios like V2. However, tireless toil on BRUTUS and MONA-LISA may eventually reveal, for everyone to clearly see, that these projects are, at root, impossible, because they would have computation do what only “super”-computation (= hypercomputation) can do. This would turn our engineering failure into philosophical gold.<sup>31</sup>

## Notes

\* We are indebted to Stevan Harnad for helpful electro-conversations, and Ned Block for *corporeal* conversations, on some of the issues discussed herein.

1. Things not necessarily to be ranked in the order listed here.
2. The problem is that probably *any* computational artifact will qualify as A-conscious. We think that does considerable violence to our pre-analytic concept of consciousness, mongrel or not. One of us (Bringsjord) has suggested, accordingly, that all talk of A-consciousness be supplanted with suitably configured constituents from Block’s definiens. All of these issues — treated in (Bringsjord 1997) — can be left aside without harming the present enquiry.
3. As Block points out, there are certain behaviors which seem to suggest that chimps enjoy S-consciousness: When colored spots are painted on the foreheads of anesthetized chimps, and the creatures wake and look in a mirror, they try to wipe the spots off (Povinelli 1997). Whether or not the animals really are self-conscious is beside the point, at least for our purposes. But that certain overt behavior is sometimes taken to be indicative of S-consciousness is relevant to what we are about herein (for reasons to be momentarily seen).
4. Def 1’s time index (which ought, by the way, to be a *double* time index — but that’s something that needn’t detain us here) is necessary; this is so in light of thought-experi-

ments like the following. Suppose that while reading Tolstoy's *Anna Karenina* you experience the state *feeling for Levin's ambivalence toward Kitty*. Denote this state by  $s^*$ ; and suppose that I (Bringsjord) have  $s^*$  at 3:05 pm sharp; and suppose also that I continue reading without interruption until 3:30 pm, at which time I put down the novel; and assume, further, that from 3:05:01 — the moment at which Levin and Kitty temporarily recede from the narrative — to 3:30 I'm completely absorbed in the tragic romance between Anna and Count Vronsky. Now, if I report at 3:30:35 to a friend, as I sigh and think back now for the first time over the literary terrain I have passed, that I feel for Levin, are we to then say that at 3:30:35  $s^*$ , by virtue of this report and the associated higher-order state targeting  $s^*$ , becomes a conscious state? If so, then we give me the power to change the past, something I cannot be given.

5. This is a bit loose; after all, the engineer could want to make a conscious robot specifically for the purposes of studying consciousness. But we could tighten Q2 to something like

Q2' Why, specifically, might an AI engineer try to give her artifact consciousness in order to make it more productive?

6. You may be thinking: Why should a robot need to believe such a thing? Why not just be able to *do* it? After all, a simple photoactive robot need not believe that it knows where the light is. Well, actually, the case we mention here is a classic one in AI. The trick is that unless the robot believes it has the combination to the lock in memory, it is irrational to for it to fire off an elaborate plan to get to the locked door. If the robot doesn't know the combination, getting to the door will have been a waste of time.

7. John Pollock, whose engineering efforts, he avows, are dedicated to the attempt to literally build an artificial person, holds that emotions are at bottom just timesavers, and that with enough raw computing power, the advantages they confer for us can be given to an "emotionless" AI — as long as the right algorithms are in place. See his discussion of emotions and what he calls **Q&I modules** in (Pollock 1995).

8. The zombies of cinematic fame apparently do have real-life correlates created with a mixture of drugs and pre-death burial: see (Davis 1985, 1988).

9. For example, the toolbox is opened and the silicon supplantation elegantly pulled out in (Cole & Foelber 1984).

10. This scenario would seem to resemble a real-life phenomenon: the so-called "Locked-In" Syndrome. See (Plum & Posner 1972) (esp. the fascinating description on pages 24–5) for the medical details.

11. Despite having no such talents, one of us (Bringsjord) usually spends twenty minutes or so telling a relevant short story to students when he presents zombies via V2. In this story, the doomed patient in V2 — Robert — first experiences an unintended movement of his hand, which is interpreted by an onlooker as perfectly natural. After more bodily movements of this sort, an unwanted sentence, to Robert's mounting horror, comes involuntarily up from his voicebox — and is interpreted by an interlocutor as communication from Robert. The story describes how this weird phenomenon intensifies ... and finally approaches Searle's "late stage" description in V2 above. Now someone might say: "Now wait a minute. Internal amazement at dwindling consciousness requires differing *cognition*,

a requirement which is altogether incompatible with the preservation (*ex hypothesi*) of identical 'information flow'. That is, in the absence of an argument to the effect that ordinary cognition (never mind consciousness) fails to supervene on 'information flow', V2 is incoherent". The first problem with this objection is that it ignores the ordering of events in the story. Robert, *earlier*, has had his brain supplanted with silicon workalikes — in such a way that all the same algorithms and neural nets are in place, but they are just instantiated in different physical stuff. Then, a bit later, *while these algorithms and nets stay firmly and smoothly in place*, Robert fades away. The second problem with the present objection is that it's a clear *petitio*, for the objection is that absent an argument that consciousness is conceptually distinct from information flow, the thought-experiment fails (is incoherent). But the thought-experiment is designed for the specific purpose of showing that information flow is conceptually distinct from consciousness! If X maintains that, necessarily, if  $p$  then  $q$ , and Y, in attempt to overthrow X's modal conditional, describes a scenario in which, evidently,  $p$  but  $\neg q$ , it does no good for X to say: "Yeah, but you need to show that  $p$  can be present without  $q$ ". In general, X's only chance is to grapple with the issue in earnest: to show that the thought-experiment is somehow defective, despite appearances to the contrary.

12. This is an argument on which Dennett has recently placed his chips: In his recent "The Unimagined Preposterousness of Zombies" (1995) Dennett says that the argument in question shows that zombies are not really conceivable.

13. For cognoscenti: we could then invoke some very plausible semantic account of this formalism suitably parasitic on the standard semantic account of  $\Diamond$ . For a number of such accounts, see (Earman 1986).

14. It is of course much easier to convince someone that it's logically possible that it's physically possible that Jones' brain is transplanted: one could start by imagining (say) a world whose physical laws allow for body parts to be removed, isolated, and then made contiguous, whereupon the healing and reconstitution happens automatically, in a matter of minutes.

15. Chalmers gives the case of a mile-high unicycle, which certainly seems logically possible. The burden of proof would surely fall on the person claiming that such a thing is logically impossible. This may be the place to note that Chalmers considers it *obvious* that zombies are both logically and physically possible — though he doesn't think zombies are *naturally* possible. Though we disagree with this position, it would take us too far afield to consider our objections. By the way, Chalmers (1996, pp. 193–200) refutes the only serious argument for the logical impossibility of zombies not mentioned in this paper, one due to Shoemaker (1975).

16. A quick encapsulation: Artificial neural nets (or as they are often simply called, 'neural nets') are composed of **units** or **nodes** designed to represent neurons, which are connected by **links** designed to represent dendrites, each of which has a numeric **weight**. It is usually assumed that some of the units work in symbiosis with the external environment; these units form the sets of **input** and **output** units. Each unit has a current **activation level**, which is its output, and can compute, based on its inputs and weights on those inputs, its activation level at the next moment in time. This computation is entirely local: a unit takes

account of but its neighbors in the net. This local computation is calculated in two stages. First, the **input function**,  $in_i$ , gives the weighted sum of the unit's input values, that is, the sum of the input activations multiplied by their weights:

$$in_i = \sum_j W_{ji} a_j.$$

In the second stage, the **activation function**,  $g$ , takes the input from the first stage as argument and generates the output, or activation level,  $a_i$ :

$$a_i = g(in_i) = g\left(\sum_j W_{ji} a_j\right)$$

One common (and confessedly elementary) choice for the activation function (which usually governs all units in a given net) is the step function, which usually has a threshold  $t$  that sees to it that a 1 is output when the input is greater than  $t$ , and that 0 is output otherwise. This is supposed to be “brain-like” to some degree, given that 1 represents the firing of a pulse from a neuron through an axon, and 0 represents no firing. As you might imagine, there are many different kinds of neural nets. The main distinction is between **feed-forward** and **recurrent** nets. In feed-forward nets, as their name suggests, links move information in one direction, and there are no cycles; recurrent nets allow for cycling back, and can become rather complicated. Recurrent nets underlie the MONA-LISA system we describe below.

17. As Ned Block has recently pointed out to one of us (Bringsjord), since at least all mammals are probably P-conscious, the accident would had to have happened quite a while ago.

18. This is eloquently explained by Flanagan & Polger (1995), who explain that some of the functions attributed to P-consciousness can be rendered in information-processing terms.

19. Some may object to P2 in this way: “*Prima facie*, this is dreadfully implausible, since each ( $x$  and  $\phi$ ) may be an effect of a cause prior to both. This has a form very similar to: provided there is constant conjunction between  $x$  and  $\phi$ , and someone somewhere thinks  $x$  is centrally implicated in  $\phi$ -ing,  $x$  actually does facilitate  $\phi$ -ing.”

Unfortunately, this counter-argument is very weak. The objection is an argument from analogy — one that supposedly holds between defective inferences to causation from mere constant conjunction to the inference P2 promotes. The problem is that the analogy breaks down: In the case of P2, there is more, *much* more, than constant conjunction (or its analogue) to recommend the inference — as is explicitly reflected in P2's antecedent: it makes reference to evidence from reports and from the failure of certain engineering attempts. (Some of the relevant reports are seen in the case of Ibsen. One such report is presented below.)

20. Henrik Ibsen wrote:

I have to have the character in mind through and through, I must penetrate into the last wrinkle of his soul. I always proceed from the individual; the stage setting, the dramatic ensemble, all that comes naturally and does not cause me any worry, as soon as I am certain of the individual in every aspect of his humanity. (reported in (Fjelde 1965), p. xiv)

Ibsen's *modus operandi* is impossible for an agent incapable of P-consciousness. And without something like this *modus operandi* how is one to produce creative literature?

At this point we imagine someone objecting as follows. "The position expressed so far in this paper is at odds with the implied answer to the rhetorical question, Why can't impenetrable zombies write creative literature? Why can't an impenetrable zombie *report* about his *modus operandi* exactly as Ibsen did, and then proceed to write some really great stories? If a V2 zombie is not only logically, but even physically possible, then it is physically possible that Ibsen actually had the neural implant procedure performed on him as a teenager, and no one ever noticed (and, of course no one *could* notice)."

The reply to this objection is simple: Sure, there is a physically possible world *w* in which Ibsen's output is there but P-consciousness isn't. But the claim we're making, and the one we need, is that internal behavior of the sort Ibsen *actually* engaged in ("looking out through the eyes of his characters") requires P-consciousness.

21. Our term for stories about the length of Betrayal.1. Stories of this type are discussed in (Bringsjord & Ferrucci 2000).

22. Information can be found at <http://www.rpi.edu/dept/ppcs/MM/c-agents.html>.

23. For reasons explained in (Bringsjord & Ferrucci 2000), BRUTUS *does* seem to satisfy the most sophisticated definition of creativity in the literature, one given by Boden (1995).

24. It may be thought that brute force can obviously enumerate a superset of *I*, on the strength of reasoning like this:

Stories are just strings over some finite alphabet. Given the stories put on display on behalf of BRUTUS.1, the alphabet in question would seem to be {Aa, Bb, Cc, ..., ;, !, ;, ...}; that is, basically the characters on a computer keyboard. Let's denote this alphabet by '*E*.' Elementary string theory tells us that though *E*\*, the set of all strings that can be built from *E*, is infinite, it's *countably* infinite, and that therefore there is a program *P* which enumerates *E*\*. (*P*, for example, can resort to lexicographic ordering.) From this it follows that the set of all stories is itself countably infinite.

However, though we concede that there is good reason to think that the set of all stories is in some sense typographic, it needn't be countably infinite. Is the set *A* of all letter As, countable? (Hofstadter (Hofstadter 1982) says "No.") If not, then simply imagine a story associated with every element within *A*. For a parallel route to the same result, think of a story about  $\pi$ , a story about  $\sqrt{2}$ , indeed a story for every real number!

On the other hand, stories, in the real world, are often neither strings nor, more generally, typographic. After all, authors often think about, expand, refine, ... stories *without* considering anything typographic whatsoever. They may "watch" stories play out before their mind's eye, for example. In fact, it seems plausible to say that strings (and the like) can be used to *represent* stories, as opposed to saying that the relevant strings, strictly speaking, *are* stories.

25. For instance, although both decimal and roman numeral notations can represent numbers, the process of multiplying roman numerals is much more difficult than the process for multiplying decimals. Of the two notations, decimal notation is in general the better representation to enable mathematical processes (Marr 1982).

26. We imagine some readers asking: “Mightn’t ‘morphogenesis’ fit better?” We use ‘epigenesis’ in order to refer to the complete process of mapping from the genotype to the phenotype. However, morphogenesis does capture the essence of the process that is used in our (Noel’s) work; frankly, we are smitten with the analogy. Choosing the atom features (say pixel for images, Hardy waves for sound) is similar to starting with a specialized cell, then forming the cell into some organization — morphogenesis.

27. Our intuition, for what it’s worth, is that humans here provide a holistic evaluation function that mimics the forces of nature.

28. Of course, even inspiration, insight, and phenomenal consciousness can preclude creativity, if one warps the example enough. But our point is really a practical warning: limiting and bounding the phenotype, *ceteris paribus*, can preclude creativity, so computational engineers beware!

29. Here, for cognoscenti, are some more details on MONA-LISA: The DNA code is a vectorized string in which each gene represents one of the pixels in an associated image (usually a 25 x 25 pixel array). The level of a gene encodes the color of the pixel (usually 4 grays, or 8 colors). Fifty images are in each generation, of which the evolver selects ten to be the parents for the next generation. Reproduction is accomplished by selecting two parents at random and generating the offspring’s DNA by randomly and uniformly selecting between the genes of the two parents at each allele site. Each population after the initial generation consists of the ten parents and forty offsprings allowing incest. MONA-LISA is a Boltzmann machine; as such its activation function is stochastic. Motivated readers may find it profitable to refer back to the brief account of neural nets given above, wherein the role of an activation function is discussed.

30. We conclude with some brief remarks on point 3: As one might expect, an increase in the creative capacity of a system causes an increase in the system’s complexity. Our evolutionary system creates representations with a new level of complexity over previous work in evolutionary computation. The increase in complexity is due to an increase in the cardinality of the relationships, increases in the level of emergent properties, and an increase in what Löfgren calls **interpretation and descriptive processes** (Löfgren 1974). The potential for complexity in a representation is determined by the relationships among the features. In the image elicitation system, the image is described at the atomic level. The low level description allows for an extremely large number of relationships as compared to systems that use a higher, feature-level representation. Image elicitation requires that features emerge along with the configuration, rather than serving to define the features with only the configuration evolving. As stated, our system promotes both polygenic and pleiotropic relationships between the genotype and the phenotype. Because of these relationships, the complexity of interpretation and description increases.

31. For a complete treatment of super-computation and related matters, including literary creativity, see (Bringsjord & Zenzen 2001) and (Bringsjord 1998).

## References

- Ashcraft, M. 1994. *Human Memory and Cognition*. HarperCollins, New York, NY.
- Averbach, E. & Coriell, A. S. 1961. 'Short-term memory in vision'. *Bell System Technical Journal* 40, 309–328.
- Block, N. 1995. 'On a confusion about a function of consciousness'. *Behavioral and Brain Sciences* 18, 227–247.
- Boden, M. 1995. Could a robot be creative? — and would we know? In K. Ford, C. Glymour & P. Hayes (eds), 'Android Epistemology'. MIT Press, Cambridge, MA, pp. 51–72.
- Bringsjord, S. 1992. *What Robots Can and Can't Be*. Kluwer, Dordrecht, The Netherlands.
- Bringsjord, S. 1997. 'Consciousness by the lights of logic and common sense'. *Behavioral and Brain Sciences* 20.1, 227–247.
- Bringsjord, S. 1998. Philosophy and 'super' computation. In J. Moor & T. Bynam (eds), 'The Digital Phoenix: How Computers are Changing Philosophy'. Blackwell, Oxford, UK, pp. 231–252.
- Bringsjord, S. 1999. 'The zombie attack on the computational conception of mind'. *Philosophy and Phenomenological Research* 59.1, 41–69.
- Bringsjord, S. & Ferrucci, D. 2000. *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*. Lawrence Erlbaum, Mahwah, NJ.
- Bringsjord, S. & Zenzen, M. 2002. *SuperMinds: People Harness Hypercomputation, and More*. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford, Oxford, UK.
- Cole, D. & Foelber, R. 1984. 'Contingent materialism'. *Pacific Philosophical Quarterly* 65(1), 74–85.
- Cooper, L. & Shepard, R. 1973. Chronometric studies of the rotation of mental images. In W. Chase (ed.), 'Visual Information Processing'. Academic Press, New York, NY, pp. 135–142.
- Davis, W. 1985. *The Serpent and the Rainbow*. Simon & Shuster, New York, NY.
- Davis, W. 1988. *Passage of Darkness: The Ethnobiology of the Haitian Zombie*. University of North Carolina Press, Chapel Hill, NC.
- Dekker, J. C. E. 1955. 'Productive sets'. *Transactions of the American Mathematical Society* 22, 137–198.
- Dennett, D. 1991. *Consciousness Explained*. Little, Brown, Boston, MA.
- Dennett, D. 1995. 'The unimagined preposterousness of zombies'. *Journal of Consciousness Studies* 2(4), 322–326.
- Dretske, F. 1996. 'Absent qualia'. *Mind & Language* 11(1), 78–85.
- Earman, J. 1986. *A Primer on Determinism*. D. Reidel, Dordrecht, The Netherlands.
- Fjelde, R. 1965. Foreward. In 'Four Major Plays — Ibsen'. New American Library, New York, NY, pp. ix–xxxv.
- Flanagan, O. & Polger, T. 1995. 'Zombies and the function of consciousness'. *Journal of Consciousness Studies* 2(4), 313–321.
- Harnad, S. 1995. 'Why and how we are not zombies'. *Journal of Consciousness Studies* 1, 164–167.

- Hofstadter, D. 1982. 'Metafont, metamathematics, and metaphysics'. *Visible Language* 14(4), 309–338.
- Knuth, D. 1982. 'The concept of a meta-font'. *Visible Language* 14(4), 3–27.
- Kugel, P. 1986. 'Thinking may be more than computing'. *Cognition* 18, 128–149.
- Löfgren, I. 1974. 'Complexity of descriptions of systems: A foundational study'. *International Journal of General Systems* 3, 197–214.
- Marr, J. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco, CA.
- Meyer, Y. 1993. *Wavelets: Algorithms and applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Perry, J. 1979. 'The problem of the essential indexical'. *Nous* 13, 3–22.
- Plum, F. & Posner, J. B. 1972. *The Diagnosis of Stupor and Coma*. F. A. Davis, Philadelphia, PA.
- Pollock, J. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA.
- Post, E. 1944. 'Recursively enumerable sets of positive integers and their decision problems'. *Bulletin of the American Mathematical Society* 50, 284–316.
- Povinelli, D. 1997. What chimpanzees know about the mind. In 'Behavioral Diversity in Chimpanzees'. Harvard University Press, Cambridge, MA, pp. 73–97.
- Rosenthal, D. M. 1986. 'Two concepts of consciousness'. *Philosophical Studies* 49, 329–359.
- Rosenthal, D. M. 1989. Thinking that one thinks, Technical Report 11, ZIF Report Zentrum f'ur Interdisziplinäre Forschung, Bielefeld, Germany.
- Rosenthal, D. M. 1990a. A theory of consciousness?, Technical Report 40, ZIF Report Zentrum f'ur Interdisziplinäre Forschung, Bielefeld, Germany.
- Rosenthal, D. M. 1990b. Why are verbally expressed thoughts conscious? Technical Report 32, ZIF Report Zentrum f'ur Interdisziplinäre Forschung, Bielefeld, Germany.
- Rosenthal, D. M. forthcoming. State consciousness and what it's like. In 'Title TBA', Clarendon Press, Oxford, UK.
- Russell, S. & Norvig, P. 1994. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Saddle River, NJ.
- Searle, J. 1992. *The Rediscovery of the Mind*. MIT Press, Cambridge, MA.
- Shoemaker, S. 1975. 'Functionalism and qualia'. *Philosophical Studies* 27, 291–315.
- Sperling, G. 1960. 'The information available in brief visual presentations'. *Psychological Monographs* 74, 48.
- Thorndyke, P. W. 1977. Cognitive structures in comprehension and memory of narrative discourse. In 'Cognitive Psychology'. Academic Press, New York, NY, pp. 121–152.

# Nothing without mind

Stephen R.L. Clark  
University of Liverpool

## 1. Why dualism seems to be a really bad idea

Philosophers exist to try out odd ideas. In what follows I shall offer a neglected form of panpsychism as a reputable alternative to a materialist metaphysics. In particular I shall suggest that panpsychists should look more sceptically at popular accounts of our evolutionary past, and that several currently fashionable notions lend support to panpsychism. The ‘odd idea’, in brief, is not so odd. Recent workers in the philosophy, and physiology, of mind sometimes suggest that, somewhere in the interval between Aristotle and Descartes, a grave mistake was made. Aristotle, we are assured, believed that people were living creatures with distinct capacities grounded in their physical makeup. Descartes erred by seeking to divide what he himself well knew was indivisible,<sup>1</sup> imagining that ‘minds’ were radically other things than ‘bodies’, and thereby making it impossible to say how one could act upon the other.

The problem is admirably expressed by Joseph Glanvill, a fervent Cartesian, in his *Vanity of Dogmatizing*, in the course of an enumeration of the limitations of human knowledge. “How the purer Spirit”, he writes, “is united to this Clod, is a knot too hard for fallen Humanity to untie. How should a thought be united to a marble statue, or a sun-beam to a lump of clay! The freezing of the words in the air in northern climes is as conceivable, as this strange union. ... And to hang weights on the wings of the winde seems far more intelligible” (Wiley 1934: 84, quoting Glanvill 1661: 20).

That it is indeed impossible for two such radically different things to act upon each other may be disputed. Thinking that cause and effect must be ‘alike’ may be the relic of a pre-Humean conception of causality. Being the cause of an effect is merely (or so Humeans suggest) to be a regularly associated precursor, such that some law-like generalization can be made about ‘cause’ and ‘effect’. Causes, on that account, are only occasions, and not explanations.

The man of science says “Cut the stalk, and the apple will fall”, but he says it calmly as if the one idea really led up to the other. The witch in the fairy tale says, “Blow the horn and the ogre’s castle will fall”; but she does not say it as if it were something in which the effect obviously arose out of the cause. ... The scientific men... feel that because one incomprehensible thing constantly follows another incomprehensible thing the two together somehow make up a comprehensible thing. ... A tree grows fruit because it is a *magic* tree. Water runs downhill because it is bewitched. The sun shines because it is bewitched (Chesterton 1961: 50f).

There is certainly a truth in that, and therefore in the thought that God could make whatever ‘mental’ property He pleased conditional on whatever ‘material’ one. But the scientific impulse is to rationalise such ‘mere conjunctions’. Leibniz was right to object to the notion that God might give matter a power of thought unconnected with its nature: this is an appeal, he said, to over-occult properties and inexplicable faculties (“helpful goblins which come forward like gods on the stage, or like the fairies in *Amadis*, to do on demand anything that a philosopher wants of them”: Leibniz 1981: 382 (4.3.7)). ‘Law-like generalization’ only attains a decently scientific status when it can be translated into mathematical form. For genuine causation (rather than magic or mere happenstance) there must be a mathematical equation between something lost and gained (motion, energy, weight or what you will). Descartes himself believed that God had placed a definite and unchangeable amount of ‘motion’ in the world at its beginning, but that ‘direction’ need not be similarly conserved.<sup>2</sup> Mind had its effect by altering the direction of movement, in effect by magic. If the projectile were deflected by another material object the motion of the two material objects would be differently distributed between them, or else be absorbed at a molecular level. If the relation between mind and matter is understood in the same way, then (*pace* Descartes), if the projectile is deflected by ‘an act of will’, a merely mental event, then the mind or will must also absorb or contribute energy. In all such changes the numbers must balance: acts of mind and matter must, somehow, be describable by the same measures.

And that, of course, is the problem. What could it even mean to imagine such an equation? Insofar as mind is essentially other than extended substance (which is equivalently space and matter) then there can be no common measure. Its causal activity must be merely magical, and hence (by modern standards) not real. It is an axiom of post-Galilean science that nothing happens differently simply because we use different words to describe it: two one-pound weights cannot, in reason, fall more slowly than one two-pound weight, since the difference is only verbal (Galileo 1952: 63). A universe in which it did

would be one where magic worked. That is why so many post-Cartesians adopted one version or another of occasionalism, and why so many others, dedicated to the mathematizing of reality, abandoned any idea that Mind could influence the world they sought to understand. Since that is, on the face of it, a paradigmatically self-refuting thesis (implying as it does that nothing that they said was caused by any thought), it must seem necessary to find another way of understanding Mind. Maybe, after all, 'minds' are no more than aspects of the real thinking bodies: to think is to engage in certain mathematizable activities, and is the result (strictly so called)<sup>3</sup> of other mathematizable movements. Minds, so some philosophers expect us to believe, are only very complex forms of motion, or of moving things. Either there are no 'minds' or 'mental acts' at all (we only *think* or *feel* there are), or else they are 'really' just the programs which run the biological computers we call animals (and we only think or feel that they are something more).

## 2. Why epiphenomenalism and reductivism are no better

At this point the most extreme materialist begins to sound exactly like a New Age Dreamer: minds are of a piece with matter, and what human minds do openly the whole of the material world does secretly. We may be unable to calculate the complex equation between obviously material (electro-magnetic, biochemical) events and apparently mental (intentional, qualitative, sentimental) ones, but that is a familiar problem. We cannot, after all, actually calculate what even obviously material things (the chemical reactions within the living cell, for example) will do, even if they do all and only what is chemically required: the result, for us, will always be one that we must find out by experience, not prior calculation (whereas the motion of most of the planets might be worked out in advance of astronomical observation),<sup>4</sup> because too many things are happening, too quickly and for too many reasons. The wetness of hydrogen oxide (that is, water), it is often said, is an 'emergent' property because it 'cannot be predicted from the properties of hydrogen and oxygen apart' (or rather, cannot be deduced from any properties that we could discover without combining hydrogen and oxygen). But 'wetness' is ambiguous: if what is meant is the feel of wetness, this is no more than an example of the problem posed by the relationship between mathematized and qualitative properties; if the mere behaviour of hydrogen oxide in a mindless, unexperienced world is what is meant, our failure

to see why the combined atoms behave like that need be no more than current ignorance. The atoms, we must presume, behave exactly as such entities are bound to, and a full explanation would consist of remarks about electron shells, quantum perturbations and the like. Of course, they do different things in company from what they do in solitude, but that is no more surprising than that a billiard ball has a different trajectory if it strikes other balls. ‘Emergence’, the appearance in a complex whole of properties the parts did not display, is always, for a strict materialist, an illusion: the parts were doing what the whole now does, but unobservably or else in ways too complex to count up. This is, of course, by its very nature, an article of faith, not demonstrated truth, but it is one that seems difficult to abandon without abandoning the scientific enterprise as it is presently understood.<sup>5</sup>

The reductive fantasy is no more coherent than the epiphenomenalist fantasy: epiphenomenalists imply that their words are not produced at the behest of conscious thoughts;<sup>6</sup> reductivists imply that the ‘thought’ that causes them is really nothing more than electro-chemical activity. Things happen differently when we seem to ‘think’ and ‘speak’, but not because of any meaning in what we ‘think’ or ‘say’ (for that would be the magic which post-Galilean science has abjured). And why should such motions result in anything that tracks the truth? “Why should anything go right; even observation and deduction? Why should not good logic be as misleading as bad logic? They are both movements in the brain of a bewildered ape.” (Chesterton 1961: 33). Natural selection of those movements that, *in evolutionary terms*, pay off, is no real answer. Few biologists, after all, suppose that ants need to be ‘thinking things’: it seems enough to suppose that they embody patterns of response that have some net advantage over other patterns. If that sort of adaptive movement is all that ‘thinking’ or ‘calculating’ is supposed to be, then ants, bees, earthworms and amoebae ‘think’: it doesn’t follow that they think correctly, or accurately, or phenomenally, or with an eye to truth. There is a sense in which even earthworms ‘map’ their surroundings (at any rate, they recreate their tunnel networks from ‘memory’), but how could they ever evolve to construct an accurate map of (say) the London Underground — a map whose construction could not possibly enhance their total genetic fitness? Why does science work? It may be no surprise that creatures like us can find our way across a dangerous landscape, but why should we be able to extend our reasoning (which somehow depends upon the hardwiring of our brains) to areas beyond any ancestral experience? Birds “are far more adept at exploiting the laws of mechanics than humans”, but have no more need of mathematical ability than most of us

(Davies 1992: 156). Yet we can do maths, and find it “unreasonably effective”, as Wigner put it (Davies 1992: 140; Penrose 1995: 415). Shouldn’t we be surprised? If we believe that we are ‘thinking things’ (even in a carefully reductive sense) and that we have some chance of mapping (say) the sidereal universe, it is difficult to see how we could also believe that the motions which produce that map are only those that could enhance genetic fitness.<sup>7</sup> “There can have been no selective advantage to our remote ancestors for an ability to reason about very abstractly defined infinite sets, and infinite sets of infinite sets, etc” (Penrose 1995: 148). The problem is exacerbated if, as some palaeo-anthropologists propose, the environment by which our ancestors were shaped was primarily the social one: we have the attributes that helped them to manipulate their neighbours, not the universe (see Dunbar 1996).

This is not a form of argument unique to anti-materialists. J.L.Mackie, for example, argues in his *Ethics* against the real existence of ‘objective moral values’ on the ground that such values would play no part in causing us to believe in them. We would still believe, he suggests, that infanticide was wrong even if there was no ‘moral fact of the matter’: accordingly, we have (he says) no reason to believe that, as a matter of fact, it is. Conversely, if we cannot abandon the claim that infanticide is wrong — or that bad arguments should be rebutted — we had better conclude that what caused us to believe these claims has something to do with their truth. Abandoning moral realism is a lot more difficult than some have thought (including Mackie): if there are no real obligations there is no obligation to accept the logical implications even of true premises, nor any obligation to prefer true premises to false. If we hold it necessary to believe that rationality is *right*, we must believe that there are real duties, and that we are the sort of creatures that can acknowledge and fulfil them.

‘Thinking’, considered as an activity constrained by laws of logic and epistemological duty, cannot sensibly be identified with the kind of complex motions that organisms with a central nervous system would have found advantageous. ‘Feeling’, considered as the experience of qualia, cannot be identified with the biochemical and mechanical effects produced in organisms by chemical or electromagnetic changes. Consistent materialists try to deny that anything does either think or feel (but never explain what it is that they are doing in denying it). Philosophers disinclined to deny the obvious (that we certainly do think and feel), but also inclined to accept materialist assumptions about true causality, adopt one version or another of epiphenomenalism. Thoughts and feelings have no effect, *qua* thoughts and feelings, on the real world: the effects that they may *seem* to have are really transformations of

material conditions which also happen to produce (for no reason anyone has yet identified, or could) both thoughts and feelings. How such thoughts and feelings could ever have been selected remains unclear. Hardly anyone manages to act on these conclusions.

From which we should conclude (if anything can be ‘concluded’) that we must, in reason, conceive minds and mental acts as something other than material. And the question of the relationship between such minds and matter, which cannot be simply mathematical, remains.

### 3. Why the material world won’t work

The thesis of many modern philosophers is that Descartes was wrong to divide, even in thought, what could not actually be divided. Our experience, as he acknowledged, is of living (and occasionally thinking) bodies: “if an angel were in a human body he would not have sensations as we do, but would simply perceive the motions which are caused by external objects and in this way would differ from a real man” (Descartes 1970: 128, January 1642). We are not angels in machines, but human beings. We could not even (it is now a commonplace to add) find out how to speak of our own thoughts and feelings if we could not (non-verbally and persuasively) recognize the feelings and the thoughts of others.<sup>8</sup> Why posit Minds (as thinking things) distinct from Bodies (as extended things), and thereby make their very thoughts, as well as their effect on Bodies, inconceivable or magical? Reductivists have sought to show that ‘minds’, after all, can be conceived as sorts of extended thing, even if we cannot actually calculate what human minds will do simply from our knowledge of what the bits that compose them might do. But if ‘minds’ are illusions, so also is matter, since ‘mind’ and ‘matter’ are each conceived as what is left of ordinary living when the other is removed.

The world we experience, what Jacob von Uexkuell called an *Umwelt* (von Uexkuell 1926; see also von Uexkuell 1957), and others have called a *life-world* (Schutz 1971: 205-59), is full of colours, scents, feelings, nuances, phantoms, memories, values, meanings, of a personal or species-specific or animal sort. The qualities we notice, that are real to us, are picked out by our evolutionary or personal history from an unobserved expanse of presently (maybe permanently) hidden qualities. Our very identification of things as the *same* things as we have perceived before is structured by our biology, our culture and our personal life. Why should our world be ‘truer’ than the lived worlds of the sheep

tick, bee or armadillo? Why should we think ‘the real world’ is composed of just the sort of things that hominids with our particular history find significant?

We may, if we like, by our reasonings, unwind things back to that black and jointless continuity of space and moving clouds of swarming atoms which science calls the only real world. But all the while the world we feel and live in will be that which our ancestors and we, by slowly cumulating strokes of choice, have extricated out of this, like sculptors, by simply rejecting certain portions of the given stuff. Other sculptors, other statues from the same stone! Other minds, other worlds from the same monotonous and inexpressive chaos! My world is but one in a million alike embedded, alike real to those who may abstract them. How different must be the worlds in the consciousness of ant, cuttlefish or crab! (James 1890: vol.1, 288f).

Did James really mean that we extracted our lived world from a “black and jointless continuity”, as if we or our remotest ancestors had once been immediately aware of that? Perhaps he did (on which see more below). Maybe he only meant that they emerged from a notional array of possible marks and tracks and sensa (and that the world defined by science as ‘real’ could constitute an explanation, somehow, of that process).

The triumph of mathematics, of course, has been to focus on such samenesses, such markers, as can be given numerical values, to uncover underlying samenesses that allow us to extrapolate to the presently unseen. Merely conventional meanings (as that purple is a royal colour) turn out not to dictate what may be true for other life-worlds. Even colours themselves (and other such so-called ‘secondary qualities’) turn out to make less difference than we thought, and to be the special creations, the special selections, of particular species. Other qualities (the ones that Aristotle thought could be perceived through several senses: such as size or shape or speed) are offered as the ones that would be required for any sensible account of what goes on. Even these primary qualities, notoriously, turn out not to be exactly what we first imagined. There is no privileged *scale*: and hence no absolute description of size, shape or speed. What turns out to remain *the same* throughout all life-worlds, and at every scale, is something only describable through mathematics, abstracted from all ordinary experience, a ratio. Purple is not, objectively, a royal colour; nor yet, objectively, a sensed colour redolent of blue and red, with echoes both of bruising and ripe plums. All that accompanies all ‘purple’, as that is scientifically, persuasively defined, is some calculable amount of transferable energies, something that can make a ‘real’, ‘objective’ difference. That idea of ‘the real world’, as something to be grasped through abstract, mathematical reasoning, takes its beginning from the marriage of Platonism and

Hebraic monotheism: whether or not that marriage is a good one, we may need to reconsider its effect.

Cartesian matter, or else the Lockean 'real world', on the far side of experienced being, is what can be described, and only described, in abstract numbers, irrespective of the lived experience from which, eventually, we have derived it. "The Wall is not White; the Fire is not Hot and so on", as Berkeley said (adding sardonically that "we Irishmen cannot attain to these truths") (Berkeley 1948: 1.47 (B392)). We carefully describe it in terms that do not carry any reference to merely subjective lives, even if those terms, historically, have themselves been derived from just such lived experience. The problem is that in choosing to describe it only in such terms we make it impossible to infer from that description anything at all of our own lived experience. A world that must, demonstratively, be experienced in one way rather than another (or in any way at all) would not be the Lockean Real World: that was the point of the Lockean Real World, and of Cartesian Matter,<sup>9</sup> that it is not well described in terms that carry such an implication. Values and secondary qualities (to select the two main sorts of qualia that materialists prefer to relegate)<sup>10</sup> are only accidentally associated with the primary qualities, the real properties of things-in-motion. Nothing in those latter properties implies the former. So it is hardly surprising that a Real World, so described, fails to explain our lived experience.

The final catch is that such accidental associations, ones that could not be deduced from any understanding of their parts, can only be identified through experience. Bachelors are certainly unmarried, and anyone who understands what bachelors are, what 'bachelor' means, will know as much. That bachelors are untidy (say), can only be discovered by experience: by being able to identify bachelors, and to see that they (or some significant proportion of them) are indeed untidy. Correspondingly, the association of Lockean Real World or Cartesian Matter and the world we live, can only be an empirical discovery: a brute fact. The world described in terms that expressly exclude all reference to subjective qualia, to mind, cannot explain the mind, because 'the mind' has been excluded from it. But neither have we any grounds to believe in such a world by mere experience: nothing we *experience* is that real world, and we cannot therefore notice that 'the real world' is frequently or universally conjoined with 'the experienced world'. We can certainly notice that some aspects of our experienced world are conjoined: damage to Wernicke's area in the human brain, for example, is associated with a particular variety of aphasia. "Patients utter fluent streams of more-or-less grammatical phrases, but their speech makes no sense and is filled with neologisms and word substitutions"

(Pinker 1994: 310). Observations of this sort are what usually convince people that there is a real connection between 'material' and 'mental' events. But they are, precisely, *observations*: all that we can actually observe are changes in our experienced world. This only provides reason to believe in an association between 'material' and 'mental' episodes if we already believe in the association, if we believe (that is) that our observation of the damage 'reflects' or 'represents' real alterations in a merely material object. Speech (and maybe thought) requires an intact brain, just as sight requires intact eyes: it doesn't follow that either speech or sight are nothing but the operation of brain and eyes. All that we can justly conclude (and even this by faith) is that all life-worlds turn out to have a certain mathematical sub-structure. We cannot conclude that such a structure would enable us to say what it is actually like to be an ant, a cuttlefish or crab, or even a victim of Wernicke's aphasia, nor that there somehow 'is' a world of purely mathematized reality, without any qualia of the excluded kind at all.

If the mathematized world, 'the world without qualities', is neither something that we can experience separately and note its constant conjunction with our lived reality, nor yet a world from which we could deduce the existence or the nature of that lived reality, why bother to hypostatise that 'real world' at all? The actual, historical answer is that, for some important purposes (say, firing projectiles) we can choose to concentrate on certain calculable properties of our lived world: what colour the projectile is, or what the gunners' pet-name for it is, will make no difference to where it falls. Again, we are so easily misled by our projections that we sometimes need to concentrate on simple properties of the people, or the animals, we seek to understand (and use). But none of this, except for political purposes, suggests that a projectile's speed or weight is all that really matters, nor that a human or non-human creature has really no existence outside the formulaic description of its weight or speed. The actual reality from which we all begin our calculations and extrapolations is the shared, lived world: to deny that world in the name of an abstract fantasy that does not even explain the world we live is only politics. Empiricists, which is what most scientifically minded persons claim to be, must in the end conclude that the world of (open) experience is the real world, and the imagined world of 'mere matter in motion' is a useful, or occasionally dangerous, phantom: "the false appearance which appears to the reasoner as of a Globe rolling thro' Emptiness" (Blake 1966: 516). There are great advantages to mathematizing our experience, but also great advantages in not doing so at the expense of other significant explanations or descriptions of what is going on. Intentional

explanation is often just as good, as Aristotle also held. The ‘Aristotelian’ approach to these issues, which philosophers who think that Descartes was mistaken now prefer, rests on the conviction that the apparent world, the world that sane people live, is itself the real world, and that mathematics is not, after all, the only way to truth.<sup>11</sup>

It may now be possible to see that the difficulty I raised for dualism (that there was no easy, mathematizable equation between acts of will and bodily changes) begged the question. The reason it seemed impossible to include mind and matter within a single causal system was that the system in question was only appropriate to an imagined matter. If we could instead suppose that there were decent explanations of another sort than the mathematical, we might reasonably suggest that mental acts sometimes explain bodily motions, even though it makes no sense to think that ‘mind’ could be measured by the same unit as ‘matter’. Descartes was wrong to suppose that there was a fixed quantum of ‘motion’, unaffected by the directions that mind imposed on matter — but not altogether wrong. A given state of matter (say, the present condition of the neural network of my brain) will always be a proper transformation of a previous state, without any need to suppose that energy has been added to, or subtracted from, the system by an immaterial other. At the same time, whatever thoughts require that state may be explained by my willingness to follow the argument where it leads (or not). The working of our brains, it seems, depends on quantum indeterminacies: that is, there may be many different outcomes of any particular neural state, each equally compatible with natural law. Which actually occurs need not be random (nor need we indulge the fantasy that all such possible outcomes happen, in alternate histories): what mathematically possible outcome happens is decided by our immaterial act of will (which is what Descartes suspected). Penrose (1995: 330), though himself opposed to the naively materialist and computational models of consciousness that are currently fashionable, rejects this thought, that “a conscious act of will could influence the result of quantum-mechanical experiment”, but his reasons seem to be, first, that such influence involves an ‘external mind-stuff’, and second, that we should then be able, logically, to influence far more than our own brains (1995: 350). Neither objection is compelling: the mind is not a stuff at all, and the reach of any individual mind is limited by the choices of innumerable others. If we cannot acknowledge that kind of explanation (that we sometimes say and do things simply because we have concluded that we should) then we might as well give up: what I had thought were thoughts were indeed only “movements in the brain of a bewildered ape” (Chesterton 1961: 33).

#### 4. How to imagine the past

What I have so far offered is a familiar, Kantian, story: we must conceive ourselves as material beings, but also as thinking things, capable of recognizing epistemological duties, and of having real experiences. Mind cannot be explained, or explained away, as a function only of material events, since such an explanation would require exactly the sort of mathematical equation that is appropriate only to matter. Can we turn the tables, and explain what happens to and in matter by reference to mind? On this account, any explanation of particular properties that exist here and now should aim to show how they have been chosen, by God or by his creatures. What reasons God and his creatures may have for their choices, however, remain thoroughly obscure: that is Descartes's reason, for example, for abandoning the quest for 'final causes'. "We shall entirely reject from our Philosophy the search for final causes; because we ought not to presume so much of ourselves as to think we are the confidants of His intentions".<sup>12</sup> It would not be absurd to rely on revelation in this matter, but I shall instead propound a 'bottom-up' account of how things got to be. Perhaps the material structures and relationships we see are the outward and visible sign of successive, mental acts and attitudes. That 'the world of physics' is one that 'we' have constructed, for particular purposes, and that our primary reality must always be that of human experience, is a thought quite commonly encountered: what is puzzling is that those who proclaim it usually have no qualms about endorsing naively realist accounts of the world as it was before us. If 'we' constructed it, who is this 'we'? And what were 'we' doing before this age set in?

Aristotle himself would probably have denied that there were earlier, different ages. The world of sane experience is, for him, the unchanging datum, even though there may be cycles of growth and decay within it. Later theorists, influenced by Christian insistence that the world did have a beginning, and a narrative history, must grapple with the thought that things have not always been as they now, to us, appear, and that how things look here and now have in some way developed out of 'simpler' beginnings. Whether the assumption that things were once 'simpler' can be given a consistent sense is itself a moot point: the usual attempt to postulate as few 'entities' as possible in one's explanation is regularly confounded by attributing yet more 'powers' to each of them. Explanations work by ruling out what otherwise might seem real possibilities — but the more such possibilities are excluded the more 'complex' is the explanans, in that it has more implications (see Clark 1991). Attempts to locate

the supposedly ‘simple’ elements even of material structures have been notably unsuccessful, in that each new range of ‘elementary particles’ turn out to be systematically related in ways that suggest that they are compounds after all. Each new range is also less intelligible to the eye of common sense. Nonetheless, I shall assume that the notion of a “black and jointless continuity” has certain strengths: distinctions have increased over time, and so has our capacity for grasping them.

Consider the following paraphrase (of the first chapter of the *Book of Genesis*). In the very beginning there were no distinctions: nothing was marked out as here or there, this or that, now or hereafter. The earth — whatever heaven was like — was “without form, and void”. Until ‘markers’ came to be, there were no identities through time at all, nor any boundary lines drawn at a single time. Nor was there any ‘direction’ of time. In that ‘*tohu-bohu*’ (the Hebrew term for chaos) there was suddenly a division between ‘now’ and ‘then’: the now was luminous, and the rest was dark. Further distinctions ramified: between this and that, up and down, and dry and wet. Those divisions were successively mirrored in many different sorts of creature, each with its own angle on the world, and each contributing another feature to the worlds of others. Kinds differentiated themselves, and lived on over generations. Seasons came, and motives; sameness in difference, and differences in the same. Each transformation created its own image, which lived on in a changing world, until at last the possibility of self-referential knowledge came about, and consciousness came to address itself as something other, and the same. This is as far, perhaps, as we have gone: that this world here is the very end of differentiation is an unwarranted assumption, but we can hardly guess what happens next (or it would have happened already). Nor should we suppose, of course, that ours is the only outcome of the process up to date: ants, cuttlefish and crabs may have their own worlds waiting. The story, in short, is a ‘subjective’ one: an account of what the *living* world was like, not an imagined metaphysical realm outside all experience.

‘Literalism’ is a theological as well as a scientific error, and *Genesis* has always been understood to mean more than that an unidentified godling ‘constructed’ things in the first six ‘days’ of time, a few millenia ago. Notoriously, there were no literal ‘mornings and evenings’ before the sun was made. But a similar error occurs in common evolutionary histories: their authors seem to take it for granted that we are entitled to suppose that ages ago there were just the sort of things that we now see — even though it is obvious that those things depend for their existence on our seeing them like that.

On the one hand it is said that the earth as we know it, including its “secondary” quality of solidity, is a construct of the human brain; and, thus, that the brain and solidity are correlatives; on the other hand it is assumed that the earth as we know it was there for billions of years before there was a human, or any other brain in existence (Barfield 1963: 157).

Scorpio is, in a sense, a real constellation, but no-one (not even astrologers) supposes that Scorpio could feature in a serious account of how things were before there were astrologers to ‘map’ the constellations. We now see evidence that once-upon-a-time there were large individual animals (say, dinosaurs): but what counts as individual existence in a world where, probably, there were no self-conscious individuals to identify themselves or others? “In reality, there are only atoms and the void: all else is convention”, said Democritus:<sup>13</sup> and a realistic explanation should not appeal to conventions.

It can do no harm to recall occasionally that the prehistoric evolution of the earth, as it is described for example in the early chapters of H.G.Wells’s *Outline of History*, was not merely never seen. It never occurred. Something no doubt occurred, and what is really being propounded by such popular writers ... is this. That at the time the unrepresented was behaving in such a way that if human beings with the collective representations characteristic of the last few centuries of western civilization had been there, the things described would also have been there. (Barfield 1965: 37)

But not only were there, as a matter of fact, no human beings of our sort present at the time, there couldn’t have been: the very conditions that, perhaps, obtained, precluded the presence of our sort of life. What then happened? Not only were there no secondary qualities before there were things to sense them, there were no identities at all (as Descartes realised: his ‘matter’ is extended stuff, equivalent to space, and there are no important boundaries in it). Individuality is an invention, even a social artefact: so why suppose that there were clearly ‘things’ *the same* from egg to brain-death? Multicellular organisms (and even eukaryotic cells) can be conceived as colonies: what, except the description, makes them one rather than many?<sup>14</sup> How was it from within? What experience was there ‘then’? And how long did it last? Popular accounts suggest that ‘billions of years’ dragged past before there was life, or multicellular life, or hominids: without a mind to bring successive moments together, there was no such real time. Until there are minds, there are neither periods nor distinguishable places, no sizes, no speeds, no shapes. Until there are minds like ours nothing of our experienced world exists at all. So what was ‘there’? I have, of course, no certain answer, except to suggest that we contain the stages of that

process, and can recall or think ourselves back into them. By doing so we may be able to perceive exactly how each stage of a developing awareness grows and maps itself into the world of sense. Rudolf Steiner, whose enterprise I have here summarised, was sufficiently successful as an educational reformer, architect, agriculturalist and psychologist to make it reasonable to take his theories seriously. That they have *not* been taken seriously (except by his disciples) is a rebuke to twentieth century thought.<sup>15</sup>

The stars themselves, according to Frank Ramsey (a more mainstream thinker), should be interpreted as marks upon mammalian retinas, and the history of their coming-to-be is a biological one (Ramsey 1931: 291). It is not a view for which I usually have much sympathy. That there are realities, like stars, that do not depend on our perception of them, and have the real properties they do for reasons that have nothing to do with our development, is as close to being an axiom as any. But if there were, as I believe, such stars — or things that have evolved into such stars — ‘before’ there were eyes, mammalian or otherwise, to see them, it must be by virtue of there being a pre-mammalian, maybe a pre-terrestrial awareness, since without such a real perspective there was no definite division between ‘star’ and spatial eddy, nor was there anything that world then was ‘like’. The history that has given us ‘stars’ is giving us signs of an unearthly life before us. In that we ourselves are star-stuff it would not be absurd to think that we could find within ourselves what it is like to be such stars: ‘we’, the very beings that have devised the worlds, have left ourselves with evidence of what we were.

Steiner suggested that there were worlds before us, and chose to give them the names of present-day planets: before ‘we’ lived on what is now the Earth we lived on ‘Saturn’, ‘Sun’ and ‘Moon’ (Steiner 1981: 154ff). After this Earth, we shall live on ‘Jupiter’, ‘Venus’ and ‘Vulcan’. It is a pity that he chose such names. Whatever ‘Saturn’ it was that ‘we’ inhabited it cannot be the planetary body that we now call by that name, as though our ancestors migrated thence. His point was rather that the forms of experience which were available to our predecessors (or our earlier selves) did not always include the possibility of discrete, enduring individuals. It is all too easy to imagine that ‘we’ now have the concepts which are wholly adequate to explain an experience which ‘they’ did not. But what grounds this conviction, that our forms of experience reach out to a ‘real world’ in ways that theirs did not? This would have, as Barfield has pointed out, the “too difficult corollary that, out of all the collective representations which are found even today over the face of the earth, and the still wider variety which history unrolls before us, God has chosen for His delight the

particular set shared by Western man in the last few centuries” (Barfield 1965: 38). That conviction is not supported even by the physical theories which we have developed out of our own forms of understanding. For physicists, our ordinary concepts of causation, identity, location and duration are no more than locally useful illusions: why then employ them in a strict account of what once was? The evolutionary picture sketched by Steiner has a different conclusion: it is the change in experience that has scattered the stars across the sky, and across the past as we now conceive it. There were no numbers, or no numbered things, until there were numbering intelligences. Such counting operations, accordingly, do not identify real causes which existed earlier. If there are no numbers until we can count, it is pointless to explain our own existence, and ability, by referring to such non-existent numbers. We should instead look back at how things were before we counted, individuated or remembered, and explain our present state (non-mathematically) by reference to what came before.

“Can we really believe”, Penrose (1995: 330) enquires, “that the weather patterns on some distant planet remain in complex-number superpositions of innumerable distinct possibilities — just some hazy mass quite distinct from an actual weather — until some conscious being becomes aware of it, at which point, and *only* at which point, the superposed weather becomes an actual weather?” The enquiry, directed at those speculative physicists who think that the quantum possibilities are only actualised by particular observations, is of a piece with the usual response to idealism: how can we believe that the truth about worlds long ago and far away depends on *us*? Berkeleians can answer that those distant facts are in God’s keeping (whether He has given them a determinate reality or not), and that our task is always to uncover the reality that God has already made. Steiner’s response, more sceptical of our ‘collective representations’, is that the distant weather is indeed ‘a hazy mass’ until some finite mind makes it distinct. We cannot look back at that haze (for our particular minds have access only to the presently differentiated world), but we can perhaps remember what it was.

Interior awareness, empathy, *participation mystique* all have their problems, and have been rather ill-regarded by our mainstream science. That has been, no doubt, a wise self-discipline.

The Western mind, since the time of Galileo, is not logically idiotic, but .. out of its unconscious drives, it has *chosen* to see the universe about it as mindless and to refuse, in the face of all the evidence, to see it in any other way. Isn’t it possible that we chose to do so, because it is to that very choice that we owe this valued

‘solitude’, on which our spiritual integrity and our freedom seem to depend? (Barfield 1963: 179)

By separating ourselves off from ‘nature’, and abandoning the search for final causes, we have made it possible to choose new ways. At the same time, we have made it *possible* to notice that other forms of life have other ends than ours (see Clark 1995). We are all easily deceived, especially by the thought that ‘we’ are a form of consciousness, of being, enormously superior to others, and somehow equipped to understand them all. But scientific objectivism itself (the attempt to reckon only with mathematized properties) shares those faults, in practice. Maybe, after all, it is time to reckon with the constructive power of empathy, of insight. Things as they first appear to us are infused with value. “On the earlier view, black bile doesn’t just cause melancholy; melancholy somehow resides in it. The substance embodies the significance” (Taylor 1989: 189; see also 379, 421ff and 478ff). We may need to move beyond the world we first encounter, but only in the sense that our vision becomes wider, deeper, more inclusive. We need to move from what is obvious to us, as Aristotle said, to what is “obvious by nature” (Aristotle *Physics* 1.184a17ff), and to the sane observer. Blake was only wrong to suggest that “the Globe rolling thro’ Emptiness” was a “false appearance” if he meant that there was no possible living world where we might see that globe. As it happens, the icon of the living earth, as it is seen from orbit, is indeed a lived reality, with its own moral and imaginative force.

## 5. Conclusion

Hypothesising a ‘real world’ apart from mind, and seeking to explain the mind’s existence and character by reference to that hypothesis, is wasted labour — nor is it obscurantist to say so, any more than it is obscurantist to resist demands for research grants to build perpetual motion machines, or square the circle. ‘Matter’ and ‘mind’ (identified by those with Cartesian sympathies) can have no common measure, and there can therefore be no way of equating either to the other. Either the two are only extrinsically connected (but we have then no reason to hypostatise the world of matter), or the division must be abandoned. That fashionable nostrum (normally invoked to denigrate the claims of mind) can be read otherwise: to denigrate the claims of matter. Either there is no ‘material world’ at all, or it is a simple aspect of *this* world, the world of changing experience. Cartesians have always suspected (as Platonists

before them) that ‘matter’ had no real, substantive, independent being: ‘mere matter’, ‘mere extension’, ‘mere countability’ is as close to being nothing at all as we can well imagine. The materiality of the things we experience, their availability for counting and the like, is a product of our dealings with each other and the world. ‘Matter’ so called is only that set of properties abstracted from the world of our experience for certain (largely political) purposes. It is convenient, for example, to insist that ‘animals’ are only ‘matter in motion’, since, as really perceived animals, they would offer far more obstacle to our casual, callous use of them. We are not thereby put in touch with the real world: our real, experienced world is merely impoverished.

Let every soul, then, first consider this, that it made all living things itself, breathing life into them. ... Before soul [the heaven itself] was a dead body, earth and water, or rather the darkness of matter and non-existence, and “what the gods hate”, as a poet says [that is, the Unseen] (Plotinus *Enneads* V.1.2, 1-2, 25-8).

Testing such hypotheses may also seem wasted labour: what could be specified as a success, or what as failure? My own suspicion is that the very success of ordinary science is partial confirmation of the theory. Materialists, of whatever school, have no real explanation of our power to find out the truth about the world at large, nor any reason to expect that we have it, nor any account of why we *ought* to seek it. The use of controlled dreaming (which is what Steiner relied upon) is also attested in accounts of ordinary scientific discovery: such dreams need to be checked before they are finally endorsed, but the fact that they work at all may still suggest that we do indeed ‘remember’ the foundations of the world, as Galileo learnt from Plato.<sup>16</sup>

Mathematics is a tool that works because it reflects realities, or cuts things “at the joints”. It is not the only tool. The ‘hard problem’ about consciousness — how such a thing could come about within an essentially unthinking world — is one that we have made up, but not by pretending to be conscious when we aren’t. Our error was to pretend that only the unconscious could be really real.

Platonists and their successors (like Penrose 1995: 412ff) think that there is a real world distinct from any world that we experience, and that we can discover at least the outlines of that real world by careful discipline, detaching ourselves from ordinary sensual hopes and fears. The discovery was possible, for Platonists, because reality and our minds alike were founded upon eternal Reason. I have argued, here and elsewhere, that modern materialists who believe that we can find out ‘Truth’ have abandoned Plato’s reason for believing this, without providing any other, and have thus created an insoluble

problem. How could an unthinking and unfeeling world generate such feeling and thinking things as we must think we are? Aristotelians, and their successors, have supposed instead that there was no ‘real world’ beyond what sane experience contains, and that there is therefore no problem how ‘unthinking matter’ can give rise to feeling and thinking things. On the contrary, ‘unthinking matter’ is no more than an abstraction from the lived realities we know. Some have concluded that the impossibility of deriving Real Consciousness from Matter demands that we abandon either science or common sense: common sense, if we deny the real existence of consciousness, and science, if we deny that consciousness can be accommodated within a materialist framework. Steiner’s insight was that it was possible to investigate the mental *scientifically*, by acknowledging its distinctive being. He concluded that the Aristotelian picture must be accommodated to the revelation that the world has changed: things have not always been as they now, truly, appear. We have lived in different worlds, created by our different attitudes. “Other sculptors, other statues from the same stone! Other minds, other worlds from the same monotonous and inexpressive chaos!” (James 1890: vol.1, 288f) The world where magic does *not* work is one that we have chosen to create, as a solid framework for our lives.

## Notes

1. Descartes did not suppose, as popular textbooks claim, that each individual human being was made of two separately identifiable things, a mind and a body: his problem was that what we conceive as a single living person must also be conceived as a composite of mind and matter, but we cannot hold these two conceptions together ‘because to do so it is necessary to conceive them as a single thing and at the same time to conceive them as two things, which is self-contradictory’ (28 June 1643, cited by Grene 1985: 19). As Plotinus argued long before: there is a single body only because there is a single, unextended soul to make it single.
2. The Cartesian principle of constancy of motion is derived (Descartes 1983: 2.37) from the notion that each and everything as far as it can always continues in the same state — and *being in motion* is (paradoxically, and *pace* Aristotle) a state: thus we have an ‘explanation’ of the projectile’s continued motion (see Donagan 1988: 66f). Descartes thence deduced seven rules of collision and said that their “demonstration was so certain that if experience would seem to prove the contrary we would be obliged to trust more in our reason than in our senses. Unfortunately, six out of the seven rules, as well as his version of the fundamental law, turned out to be false” (Hooykass 1972: 43).
3. The ‘result’ of two or more motions bears a mathematical relationship to its originating

motions: this is to be contrasted with an 'emergent' property of a whole, bearing no mathematical relationship to the properties of the parts.

4. It turns out that even planetary motions are not so easily calculated: Pluto's orbit cannot be predicted or retrodicted beyond twenty million years, since present conditions cannot be specified so exactly as to distinguish between wildly different consequences, even if nothing influences that orbit except the astronomical bodies that we know about. But our knowledge of what planetary orbits would be is still better than our knowledge of the living cell. The chemical changes of a single cell cannot be followed so exactly that we could know what they would be if nothing interfered with them: consequently, we do not know that nothing does.

5. Nagel 1979: 181ff has briefly sketched the case against 'emergence'. As he acknowledges, a merely Humean account of causality eliminates the problem: put A and B together and we just do get C, every time we do it. But decently scientific theories can't consist of those bare associations. See Clark 1984: ch.6.

6. Experiments conducted by Libet and others purport to show that decisions take effect, and observations made, *before* the agent is aware of the decision or perception. I share Penrose's doubt that conscious decisions or perceptions can, even in principle, be so securely timed (Penrose 1995: 387). I would add, in anticipation of later argument, that we can be far more confident of the efficacy of conscious decision than of the metaphysical claim that effects must always follow causes. The notion that such experimental observations could give us reason to think that we don't need to be conscious to have reasons is not a scientific, but a metaphysical claim.

7. I argued this point at greater length in Clark 1984. There are many other, but related arguments to a similar point, including C.S.Lewis's. Anyone interested in the history of that resilient mental microbe 'Anscombe's comprehensive refutation of Lewis's amateur philosophising' should read Anscombe's own account, in the second volume of her collected papers (1981: ix). The first version of Lewis's argument was published in the first edition of his *Miracles*. Anscombe quite properly criticised the formulation in her first published philosophical work (*Socratic Digest* 1947: reprinted in Anscombe 1981: 224-32). Lewis worked harder at the argument, and published a much improved version in the second edition of *Miracles*. Anscombe recorded, in 1981, that she thought her criticisms of the first version were just, as Lewis acknowledged, but that they lacked 'any recognition of the depth of the problem'. She also testified to Lewis's 'honesty and seriousness' in reconsidering the argument, and to 'the depth and difficulty' of the issue. She added that descriptions of the 1947 meeting as 'a horrible and shocking experience which upset [Lewis] very much' did not fit her own, or others', memory. I am told (by Anthony Kenny) that Lewis himself did something to propagate the story, but no such angst is visible in his rewritten version. Nor need it have been.

8. Which is why it is truly ridiculous to claim that non-humans have no feelings because they cannot tell us what they are: if we could not recognize such feelings without the aid of language we could never learn, or teach, a language.

9. Which are not quite the same: Lockean substances are usually understood to be something like Epicurean atoms in a void; Cartesian matter does not occupy, but is identical

with, space. The two sorts of non-mental entity are alike just in that mind is not a relevant explanation of anything they do.

10. Other writers may prefer to call 'qualia' only such phenomenal properties as redness, painfulness or loudness: even they may have evaluative elements. Experienced 'values' (such as beauty, danger, virtue) may themselves depend on qualia of that kind, or may constitute real qualities distinct from those identified as 'secondary'. I think that values are as much a feature of our experience as colours, and as closely related to our evolutionary history.

11. Plato agreed with Democritus, that "truth lies in the depths" (68 B117 DK), and with Heraclitus, that it "loves to hide" (22 B123 DK), though he resisted the Democritean suspicion that the real world consisted only in "atoms and the void" (68 B9 DK). Aristotle, acknowledging that we did not all see things straight, nonetheless supposed that truth was what the wise, or sane, person sees, and that there was no gap between truth and perception (see Clark 1975: 191ff; Clark 1991: 50ff).

12. Descartes 1983: 14 (I.28); and see 1970: 117: 'we cannot know God's purposes unless God reveals them'. Socrates seems, according to Plato's *Phaedo*, to have reached a similar conclusion.

13. According to Sextus Empiricus *Against the Mathematicians* 7.135.

14. A currently popular idea associates 'self-hood' with the immune system: organisms, it is said, defend themselves against 'the Other', and exist by virtue of their protective boundaries. I suspect that this will prove to have been a false trail, followed for ideological reasons: do pregnant mammals protect themselves against their young?

15. For an elegant introduction to Steiner's thought see Barfield 1963 and 1965. It must be admitted that most of Steiner's own writings are so laced with his idiosyncratic mythology that it is easy to forgive those who have given up the task of interpretation. As McDermott (Steiner 1984: 168) remarks, 'readers who are new to spiritual science and to esoteric thought may well find Steiner's writings on such topics either unfounded, or ludicrous, or both'. It is worth remarking that ideas which have seemed thus ludicrous in Steiner often receive a respectful hearing when couched in other terms: that people of an earlier age had a different form of consciousness, for example, or that self-conscious individuality is a social invention, or that children acquire adult habits of reasoning and self-awareness by predictable degrees, or that emotions are modes of perception.

16. "We find and discover [fundamental laws of motion] not in Nature, but in ourselves, in our mind, in our memory, as Plato long ago has taught us": cited by Koyré (1968: 13, 42). See further Clark 1990: chs. 2 and 5.

## References

- Anscombe, G.E.M. 1981. *Metaphysics and the Philosophy of Mind*. Oxford: Blackwell.  
 Barfield, O. 1963. *Worlds Apart*. London: Faber.  
 Barfield, O. 1965. *Saving the Appearances*. New York: Harcourt, Brace & World.

- Berkeley, G. 1948-56. *Collected Works*, A.A. Luce & T.E. Jessop (eds). Edinburgh: Thomas Nelson.
- Blake, W. 1966. *Complete Writings*, G. Keynes (ed.). London: Oxford University Press.
- Chesterton, G.K. 1961. *Orthodoxy*. London: Fontana.
- Clark, S.R.L. 1975. *Aristotle's Man: Speculations upon Aristotelian Anthropology*. Oxford: Clarendon Press.
- Clark, S.R.L. 1984. *From Athens to Jerusalem*. Oxford: Clarendon Press.
- Clark, S.R.L. 1990. *A Parliament of Souls: Limits and Renewals II*. Oxford: Clarendon Press.
- Clark, S.R.L. 1991. "Limited explanations". In D.Knowles (ed.), *Explanation and its Limits*. Cambridge: Cambridge University Press, 195-210.
- Clark, S.R.L. 1995. "Objective Values, Final Causes". *Electronic Journal of Analytical Philosophy* (<http://www.phil.indiana.edu/ejap/>) 3: 65-78.
- Clark, S.R.L. 1996. "Plotinus: Body and Mind". In Lloyd Gerson (ed.), *Cambridge Companion to Plotinus*. Cambridge: Cambridge University Press, pp.275-91.
- Davies, P. 1992. *The Mind of God*. London: Simon & Schuster.
- Descartes, R. 1983. *Principles of Philosophy*, R.P. & V.R. Miller (eds). Reidel: Dordrecht.
- Descartes, R. 1970. *Philosophical Letters*, A. Kenny (ed.). Oxford: Clarendon Press.
- Donagan, A. 1988. *Spinoza*. London: Harvester.
- Dunbar, R. 1996. *Grooming, Gossip and the Evolution of Language*. London: Faber.
- Galileo Galilei. 1952. *Dialogues concerning Two New Sciences*, H. Crewe & A. de Salvo (tr.). New York: Dover Publications: New York. Translated 1914; 1st published 1638.
- Grene, M. 1985. *Descartes*. London: Harvester.
- Hooykaas, R. 1972. *Religion and the Rise of Modern Science*. Edinburgh: Scottish Academic Press.
- James, W. 1890. *The Principles of Psychology*. London: Macmillan.
- Koyré, A. 1968. *Metaphysics and Measurement*. London: Chapman & Hall.
- Leibniz, G. 1981. *New Essays on the Human Understanding*, P. Remnant (tr.), J. Bennett (ed.). Cambridge: Cambridge University Press.
- Lewis, C.S. 1947. *Miracles*. London: Bles.
- Lewis, C.S. 1960. *Miracles*. London: Fontana.
- Mackie, J.L. 1977. *Ethics: Inventing the Difference between Right and Wrong*. Harmondsworth: Penguin.
- Nagel, T. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.
- Penrose, R. 1995. *Shadows of the Mind*. Vintage: London.
- Pinker, S. 1994. *The Language Instinct: the new science of language and the mind*. London: Allen Lane.
- Plotinus. 1984. *Enneads*. tr. A.H. Armstrong. (Loeb Classical Library). London: Heinemann.
- Ramsey, F.R. 1931. *Foundations of Mathematics*. London: Kegan Paul.
- Rand, B. 1914. *Berkeley and Percival*. Cambridge: Cambridge University Press.
- Schutz, A. 1971. *Collected Papers I: The Problem of Social Reality*, M. Natanson (ed.). The Hague: Nijhoff.
- Steiner, R. 1981. *Cosmic Memory*, Karl.E. Zimmer (tr.). San Francisco: Harper & Row; 1st published 1959.

- Steiner, R. 1964. *The Philosophy of Freedom*, M.Wilson (tr.). London: Rudolf Steiner Press: London (7th edition; 1st published as *Die Philosophie der Freiheit* in 1894).
- Steiner, R. 1984. *The Essential Steiner*, R.A. McDermott (ed.). San Francisco: Harper & Row.
- Taylor, C. 1989. *Sources of the Self*. Cambridge: Cambridge University Press.
- von Uexkuell, J. 1926. *Theoretical Biology*, D.L. Mackinnon (tr.). London: Kegan Paul.
- von Uexkuell, J. 1957. "A stroll through the worlds of animals and men". In C.H. Schiller (ed.), *Instinctive Behaviour*. New York: International University Press, 5-80.
- Wiley, B. 1934. *The Seventeenth Century Background*. London: Chatto and Windus.



PART III

**Artificial consciousness**



# Studying the emergence of grounded representations

## Exploring the power and the limits of sensory-motor coordination

Stefano Nolfi and Orazio Miglino

Institute of Cognitive Sciences and Technologies, CNR /  
University of Naples II

### 1. Introduction

The word *consciousness* is often used by referring to different meanings. Moreover different level of consciousness may be identified (see Chalmers 1995). In this chapter we will deal with the very first levels of consciousness from an evolutionary point of view. In particular we will investigate in which conditions we can expect the emergence of internal representations in populations of evolving agents that interact with an external environment.

Also the word *representation* can be used with different meaning. In this paper we will use the term *pure sensory-motor agents* to indicate agents that do not have any internal representation. An example of pure sensory-motor agent is a robot controlled by a neural controller in which sensory neurons are directly connected to motor neurons. The sensory neurons are activated by the external environment. The sensory patterns received from the environment and the connection weights determine the activation of the output neurons and as a consequence the way in which the agent acts in the environment. Pure sensory-motor agents may adapt to the environment by changing phylogenetically. However they cannot change their connection weights during lifetime. Therefore, they react always in the same way to the same sensory patterns during lifetime.

We will use *agents that have internal representations* to indicate agents in which sensory neurons are not directly connected to motor neurons. These agents, for example, might have internal neurons and transform sensory pat-

terns into internal patterns that are then transformed into motor patterns. An example of agent that has an internal representation is a mobile robot controlled by feed-forward neural network with a layer of sensory neurons, a layer of internal neurons, and a layer of motor neurons. Like pure sensory-motor agents, agent that have internal representations cannot change their connection weights during lifetime and as a consequence always react in the same way to the same sensory patterns during lifetime.

We will use *dynamical agents or agents with an internal dynamics* to indicate agents that have internal states that change during lifetime. These agents change both phylogenetically and ontogenetically. An example of dynamical agent is a robot controlled by a neural network with recurrent connections and/or by a neural network in which the connection weights change during lifetime. Dynamical agents may respond to the same sensory state in different ways during lifetime.

Different type of agents may exhibit different forms of behavior and different levels of consciousness. In particular, agents that have internal representations may exhibit more complex forms of behavior than pure sensory-motor agents. For instance they can produce sharp different motor reactions for quite similar sensory states. Similarly, dynamical agents may exhibit more complex forms of behavior than sensory-motor agents (with or without internal representations). For example they might display an ability to adapt to changes occurring during their lifetime.

In this paper we will investigate in which conditions we can expect the emergence of agents that exploit internal representations and/or internal dynamics in populations of evolving agents.

### 1.1. Embodied systems

An organism is a system that lives and acts in an environment. Unfortunately, this obvious statement is often overlooked in cognitive science research. Also biologically inspired approaches like connectionism treat cognition as a structure that develops in a formal and abstract environment. These approaches try to understand the computational style of the brain and implicitly assume that we will know the intrinsic nature of cognition (and consciousness) when the laws that govern our brain will be discovered. However an organism is not only a collection of neurons. Organisms “live” in an external environment and actively extracts information from the environment by coordinating the sensory and motor processes (cf. Piaget 1971).

Unlike researchers in other fields of A.I., researchers in mobile robotics have been forced to deal with systems that interact with an external environment. However, only limited results have been obtained in this area in the last years. Aside few notable exceptions (e.g. Nilsson 1984) “such robots seemed to have reached a ceiling in terms of what they can do” (Sharkey & Heemskerk 1997). This partial failure demonstrates that it is difficult to apply traditional cognitive models to mobile robots. As we can see in everyday life in fact, there are efficient computer programs that play chess or solve formal problems but there are no intelligent mobile robots in our houses or towns. In other words, this shows that it is difficult to link models of high-level aspects of intelligence with the dynamic and unpredictability of the real world, at least by treating mobile robots as computer with wheels.

This difficulty is one of the reasons why a new approach called *behavior based robotics* that is firmly grounded in the real world has been proposed and has gained much attention (Brooks 1991). This approach proved to be more successful than previous attempts. However, most of the applications developed within this approach involve systems that are reactive (i.e. system in which sensors and motors are almost directly linked) and in which internal representation and/or internal dynamics, if any, play a limited role.

Behavior based robotics is a new field of research and it is certainly too early to draw conclusions on how far this approach can lead us. However, we have the feeling that a ceiling effect is going to be observed here too. Systems displaying behavior of remarkable complexity have been obtained (see Pfeifer & Scheier 1999). However, most of the research works within this field describe different algorithms or architectures applied to the same type of relatively simple behaviors and only few really address the problem of scaling up to significantly more complex cases.

One can conclude that something important is missing in this approach and that it is this missing characteristic that prevents the possibility to scale up to significantly more complex forms of behavior. It can be claimed that while traditional A.I. was an erroneous attempt to capture complexity, behavior based robotics, by dealing with systems that are grounded and situated, is a first step in the right direction. However, successive steps must be taken to obtain more complex behaviors. Sharkey & Heemskerk (1997) for example, claimed that most of the research within behavior based robotics can be classified at the lower end of the four successive stages in the development of human cognition hypothesized by Piaget: sensory-motor, pre-operational, concrete operationism and formal operations (Piaget 1952). Following this line of thought one

can try to identify the most appropriate second step to be taken: dynamics agents (Beer 1995); agents that are able to build an internal representation by interacting with the external environment (Asada 1990); agents that learn to predict the consequence of their actions and can use this knowledge to choose between alternative actions (Tani 1996); agents able to learn by imitation (Bakker & Kuniyoshi 1996).

We believe that all these research lines are valuable and can produce important insights. However, we also believe that more fundamental questions must be answered. In particular, in what conditions are pure sensory-motor systems insufficient? What is the best experimental framework for studying more complex systems like those described above? These questions are particularly compelling for approaches like evolutionary robotics (Cliff et al. 1993; Nolfi et al. 1994) where one expects that progressively more complex behaviors can be obtained through a self-organization process in which agents have a high degree of autonomy.

In the following sections we will present the results of some experiments in which a mobile robot is trained to perform three different tasks that apparently require the use of internal representation and/or internal dynamics. To train the controller of the robot we will use an evolutionary approach. As we will see, this approach leaves evolving agents free to develop their own strategy to solve the task. Then we will analyze the solutions found by the evolutionary process that, contrary to what can be expected, rely on pure sensory-motor controllers. Finally, we will discuss the implication of the obtained results.

## **2. Exploring an arena surrounded by walls**

Lund & Hallam (1996) reported a simulation in which a Khepera robot was trained to perform an exploratory task (i.e. to visit as much of the environment as possible within a given time). We present the results of a replication of this experiment in almost identical conditions.

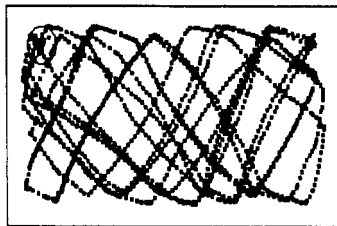
Khepera is a miniature mobile robot developed at E.P.F.L. in Lausanne, Switzerland (Mondada et al. 1993). It has a circular shape with a diameter of 55 mm, a height of 30 mm, and a weight of 70g. It is supported by two wheels that can move in both directions at 10 different speeds and has eight infrared proximity sensors (six sensors are positioned on the front and two on the back of the robot). The robot was placed in a rectangular arena of 60x30 cm idealized as divided into cells of 2x2 cm. The robot was trained in simulation

and then tested on the physical robot (see Miglino et al. 1995).

To control the robot we used neural networks with different architectures. All neural controllers had 8 sensory neurons connected to the 8 infrared sensors of the robot and 2 output neurons connected to the left and right motor of Khepera. The first architecture consisted of a pure sensory-motor neural network (a perceptron) with 8 sensory and 2 motor neurons. The second architecture consisted of a dynamical neural network with two additional output units whose activation was copied back into two additional input units.

The architecture of the controllers was kept fixed. The connection weights were encoded in a genotype string made up of binary digits (each weight was encoded with 8 bits and could assume 256 different values ranging from  $-10.0$  to  $+10.0$ ) and subjected to an evolutionary process. Each generation consisted of 100 individuals. The first generation was created by randomly generating the genotypes of 100 corresponding individuals. Each individual was then evaluated for its ability to control a robot placed in a randomly selected position of the environment for 3 epochs, each epoch consisting of 1500 actions of 100 ms each. The 20 best individuals of each generation were allowed to reproduce by generating 5 offspring each with their same genotype but with 10% of their bits randomly changed. The process was repeated for 100 generations. Individuals were scored by counting the number of cells visited for the first time during each epoch.

Surprisingly enough we found that pure sensory-motor robots proved capable of displaying quite efficient behavior. Figure 1 shows the typical behavior of one evolved individual. As can be seen, the robot is able to visit most of the environment even if it cannot explicitly know where it has been before. As in Lund and Hallam's experiments this is accomplished "by using a specific turning angle each time the robot approaches a wall, so that the perceptron

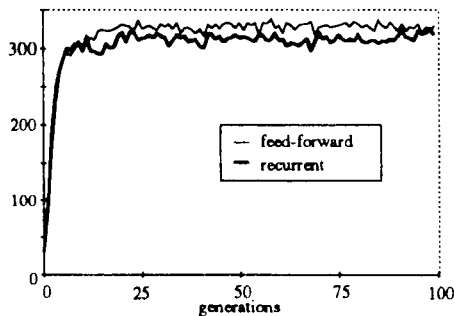


**Figure 1.** Trajectory of a typical evolved individual left free to move in the environment for 1500 cycles (150 seconds). Almost identical behavior can be observed by downloading the individual's control system into the real robot and testing it in the real environment (see Lund & Hallam 1996).

navigates the robot through the environment exactly so as to have the robot encountering previously non visited cells most of the times” (Lund & Hallam 1996: 11).

If we compare the performance of individuals with the two architectures we see how the addition of recurrent connections does not allow individuals to achieve better performance. On the contrary, slightly worse performance are observed (see Figure 2). This is due to the fact that pure sensory-motor controllers are enough to solve this task and, on the other hand, the space to be searched is larger in the case of the experiment with dynamical neural controllers. A further proof of the fact that the possibility to have an internal dynamic is not exploited by evolved individuals is that their connection weights are selected so to keep the activation value of the internal neurons always at the same level. In other words, these individuals do not exploit the possibility of having an internal dynamic.

Notice how in principle slightly better performance may be achieved by using dynamical controllers. By storing in internal states some information about the areas of the environment that have been already visited in fact, one can minimize the time spent to cross cells that have been already visited. From the evolutionary point of view however, given that near optimal performance is achieved with a pure sensory-motor solution and that the sensory-motor solution is discovered first, there is not enough pressure to select a more complex, albeit more efficient, solution.

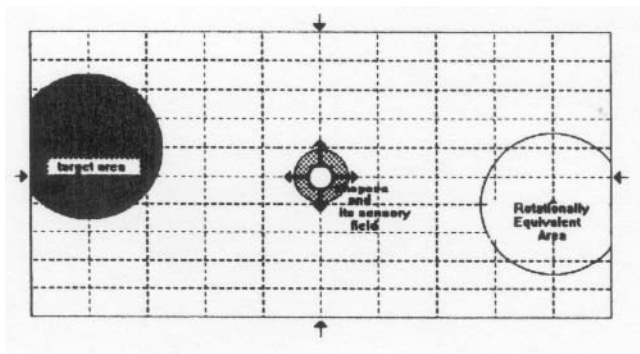


**Figure 2.** Number of cells visited for the first time in populations of individuals with and without recurrent connections throughout generations. Performance of the best individuals of each generation. Each curve represents the average result of 10 replications.

### 3. Navigating in an open field box

During the last decade Gallistel (Gallistel 1990; Margules & Gallistel 1988), Cheng (1986) and other researchers have used the experimental framework described in Figure 3 to show that rats build cognitive maps of the external environment and use them to navigate. In a first phase of their experiments, Gallistel and Cheng let rats learn how to locate a visible food patch in an open field box. Then they tested the rats for their ability to find a buried food patch located in the same position.

Ten rats (divided into two groups) were analyzed in three different experimental conditions: (a) in an box with and without angular cues (i.e. different visual landmarks located in the corner of the box); (b) in an box with and without angular cues but with the orientation of the box randomly changed; (c) in a box with angular cues, with the orientation of the box randomly changed, and with the top of the box covered to prevent the use of cues external to the box itself. The rats were able to find the buried food most of the time in condition (a) and (b) in which they could rely on cues external to the box but not in condition (c). Interestingly, in this condition they were able to locate the correct food location in about one third of the trials while about the same number of times they looked for food in a position that had the same geometric relationships of the correct position with respect to the rectangular shape of the box. The authors called these systematic errors rotational errors because the rats confused the correct food location with a location that would be correct if



**Figure 3.** The open field box experiment. One possible target position and the corresponding rotational error are shown. Arrows indicate the 8 different starting positions and orientations used during the training and testing phases.

the box was rotated 180 degrees. For example, if the food was located in the target position (see the gray area in Figure 3), rats looked half of the time in the right area and half of the times in the rotationally equivalent area (see the empty circle in Figure 3).

According to the authors the rotational errors observed in condition (c) show that rats base their heading only on the box shape. In addition, given that rotational errors can be explained only in Euclidean terms, the existence of these errors indicate that rats have a representation of the shape of the box based on Euclidean geometry. Finally, the fact that angular cues are not taken into account might imply the presence in rats of a neural module (c.f. Fodor 1983) of the shape of the environment.

Given the elegance of this experimental framework and the large amount of experimental data collected by Gallistel, Cheng and others we decided to try to replicate the experiment by using a mobile robot. We thought that if we could succeed in solving this problem we would obtain interesting insights on how an internal representation of the external environment could emerge and how environmental information are represented in biological organism and can be represented in artificial organisms.

We set up an experiment in which a Khepera robot was placed in a rectangular environment of 120x60 cm ideally divided into 100 cells of 12x6 cm. During both training and testing the robot was asked to reach the target area (food) from 8 different starting positions and orientations after 80 input/output cycles of 100 ms each. As in the case of the original experiments on rats, the initial positions were located in the middle of each of the four walls facing the center of the environment and in the center of the environment facing the four different walls (see Figure 3). Robots were trained by using artificial evolution and the experiment was repeated for each possible position of the target among the 80 intersections of the lines dividing the floor into 100 cells, with the exception of the central intersection.

The robot was controlled by a pure sensory-motor neural network with 8 sensory neurons connected to the 8 infrared sensors of the robot and 2 output neurons connected to the two motors that controlled the left and right wheels. To train the network, an evolutionary approach of the same type as described in the previous section was used. Individuals were scored for the number of times (out of 8) they were able to reach the target area (i.e. to reach a distance below a threshold of 7.5 cm from the intersection corresponding to the position of the food) after 80 input/output cycles. The initial generation was made

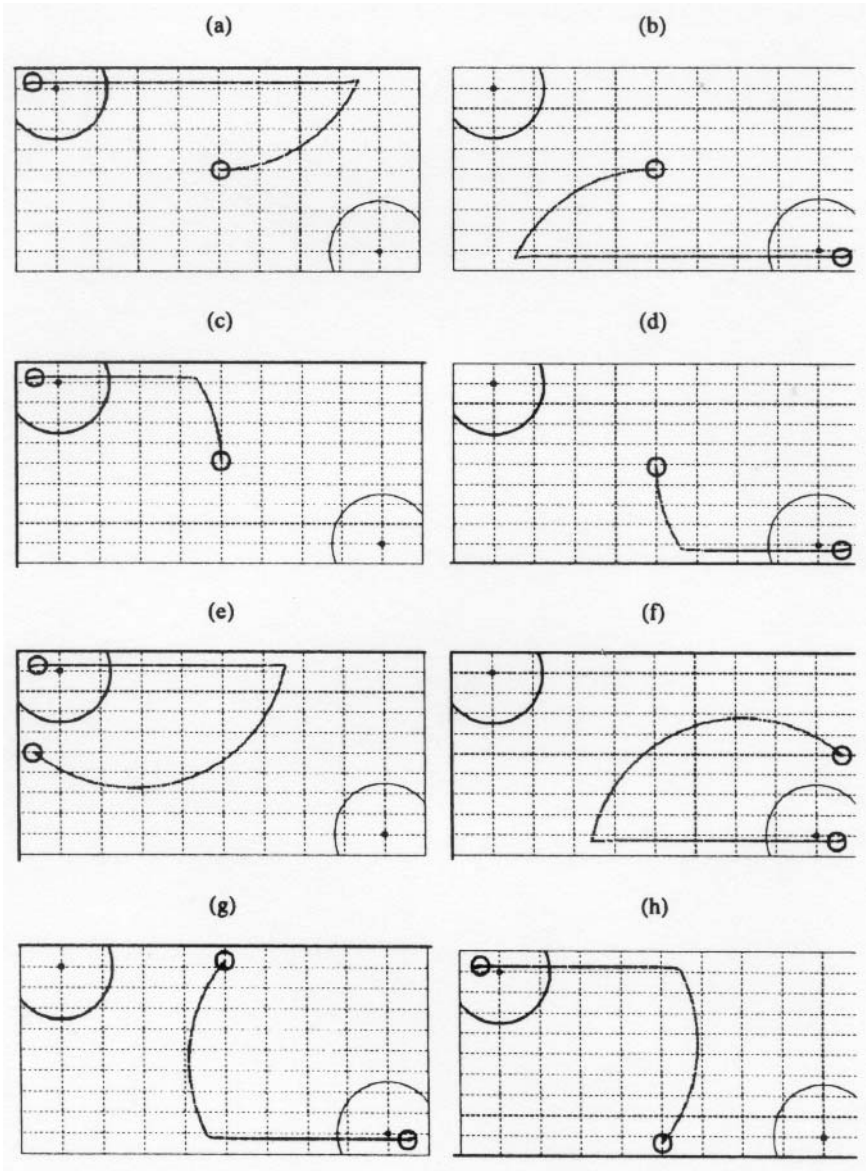
up of 100 individuals with randomly generated genotypes. The 20 best individuals of each generation were allowed to reproduce by generating 5 offspring with their same genotype but with 10% of their bits replaced by new randomly generated values. Each evolutionary process lasted 50 generations.

By looking at the result of the simulations we found that individuals with pure sensory-motor neural controllers were able to locate the food area or the rotational equivalent area in 82% of the cases (see Table 1). As in the case of rats, robots navigated toward the target area and the rotational equivalent area the same number of times. Overall performance were comparable to those displayed by rats even if robots scored slightly fewer misses.

It required some effort to understand how pure sensory-motor neural controllers with short distance sensors (the infrared sensors can detect walls up to a distance of about 4 cm) were able to achieve such a good performance. Figure 4 shows the behavior of a typical evolved individual trained in an environment in which the target area was located in a corner. At a first sight we might think that the robot discriminates between long and short walls and navigates toward the target area by keeping the long walls on its left. On the other hand, given that infrared sensors can only detect obstacles from a close distance the robot cannot discriminate between walls with different length. The actual strategy developed by the robot is much simpler. If we carefully analyze the behavior of this individual and we bear in mind the characteristics of its sensory systems we can describe its behavior in the following way: (a) when sensors are inactive the robot produces a curve trajectory by turning on its left side with a given angle; (b) when the robot perceives a wall with its left-side sensors it avoids the wall by turning left; (c) when the robot perceives a wall with its right-side sensors it performs a wall following; (d) when it perceives a wall both with its left-side and right-side sensors it stops. This strategy is able to produce 50% of correct navigation and 50% of rotational errors. Of course this strategy is valid only for this food location. Nevertheless for other food locations artificial evolution was able to select other strategies of the same

Table 1. Percent of navigation toward the correct area (correct), the rotationally equivalent area (rotational errors), and other wrong locations (misses) for robots and rats in experimental condition (c). Data on rats are taken from Margules & Gallistel (1988).

|         | Correct | Rotational Errors | Misses |
|---------|---------|-------------------|--------|
| Rats    | 35      | 31                | 33     |
| Khepera | 41      | 41                | 18     |



**Figure 4.** Behavior of a typical evolved robot. The target area is located in the upper left corner. In conditions (a), (b), (c), and (d) the robot is initially placed in the center of the environment with four different orientations. In conditions (e), (f), (g), and (h) the robot is initially placed along the four walls facing the center of the environment. The robot scored 4 correct navigation, and 4 rotational errors.

type (i.e. strategies that rely on sensory-motor coordination but do not require internal representations and/or internal dynamics).

It is important to clarify that these results do not imply that rats do not use internal representations and/or internal dynamics to solve this task. This problem can be solved in different ways and some of these ways may rely on internal representations and/or internal dynamics. On the other hand these results demonstrate that this task can be solved by pure sensory-motor systems.

#### 4. Recognizing objects with different shapes

In this section we describe another experiment in which a Khepera is placed in an arena of 60x35 cm surrounded by walls containing a cylindrical object with a diameter of about 2.3 cm located at a random position and is asked to discriminate between walls and cylindrical objects by finding and remaining close to the cylinder (Nolfi 1996).

To discriminate between walls and cylinders is difficult given that the sensory patterns corresponding to the two objects largely overlap in sensory space. This can be shown by training a neural network to classify the two objects at different orientations and distances. We used 3 different architectures: (a) a perceptron with an input layer with 6 neurons corresponding to the 6 frontal infrared sensors of the robot and one output neuron; (b) an architecture with an additional layer with four hidden units; (c) an architecture with an additional layer with 8 units. For each architecture we ran 10 training processes starting with different randomly assigned initial weights. Networks were trained by back-propagation for 5000 epochs. During each epoch, networks were exposed to the sensory patterns perceived by the robot at 20 different distances and at 180 different angles with respect to a wall and to a target, for a total of 7200 different patterns. The network was asked to produce an output of 0 or 1 for sensory patterns corresponding to walls or cylinders respectively. A learning rate of 0.02 and no momentum were used.

In the most successful replications, networks trained with back-propagation were able to classify correctly the two types of stimuli (i.e. to produce an activation value below 0.5 in the case of a wall and above 0.5 in the case of a target) only about 22% of the cases for networks without hidden units and only about 35% of the cases for networks with the additional layer of 4 or 8 hidden units.

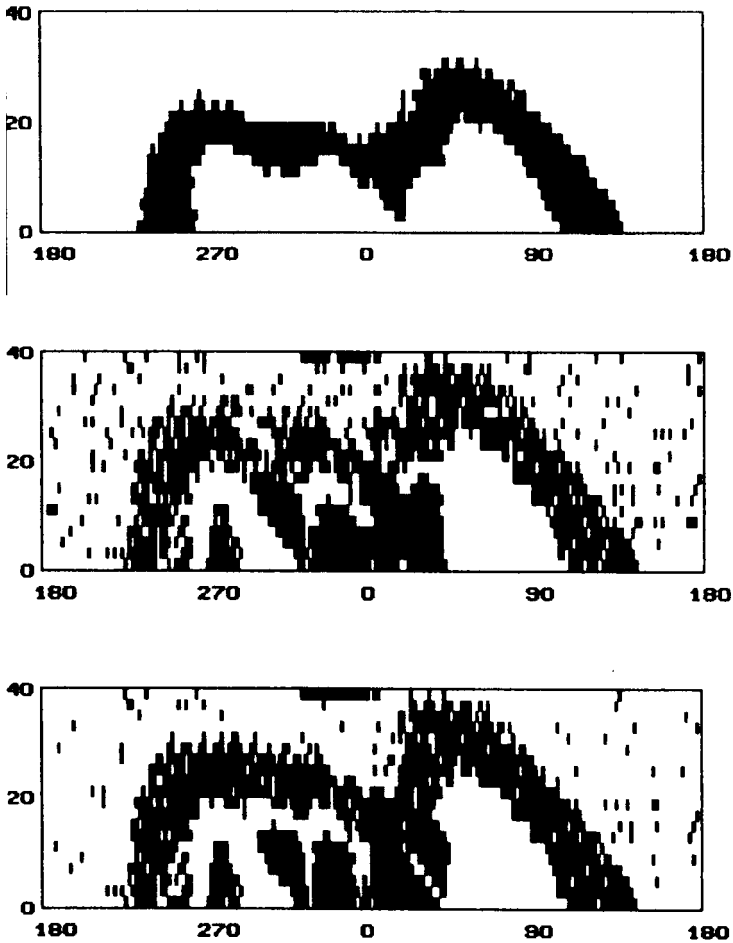


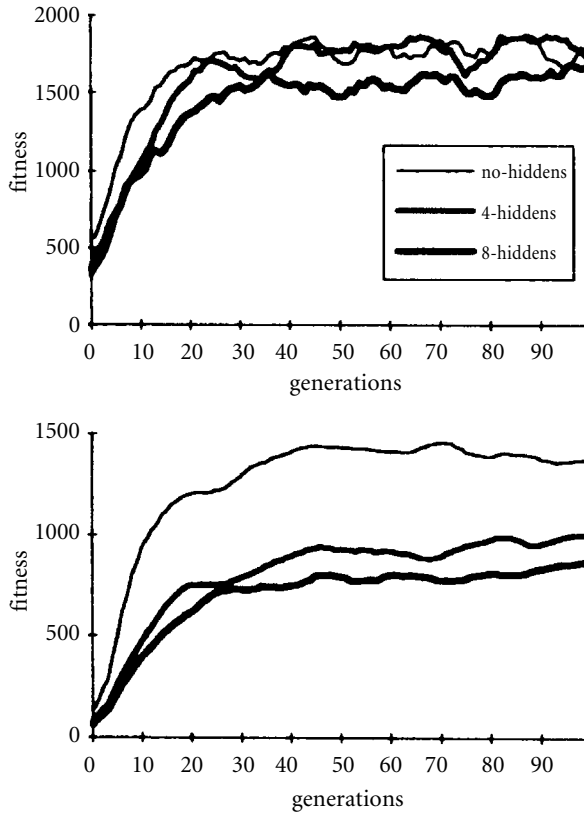
Figure 5. The areas in black represents the relative positions from which networks are able to correctly discriminate the sensory patterns belonging to walls from those belonging to small cylindrical objects. The three pictures (from top to bottom) represent the result for the best simulation with no hidden, 4 hidden, and 8 hidden units, respectively.

The fact that only a minority of the sensory patterns can be correctly classified is due to the fact that, given the sensory apparatus of the robot, the two classes of patterns largely overlap in the sensory space. If we look at Figure 5, that represents the positions (i.e. the combination of angles and distances) from which the networks are able to correctly discriminate between sensory patterns belonging to walls and cylinders, we see that the networks produce correct answers when the objects are not more than  $120^\circ$  to the left or the right of the robot face and no more than 32mm away. This result is not surprising if we consider that the network relies only on its 6 frontal infrared sensors, and that when the distance is high the infrared sensors are only slightly activated. However, there are two areas in which objects cannot be correctly discriminated even though they are “within sight”: in the figure these are the white areas enclosed by the black outer stripes.

Given the impossibility of discriminating objects in most of the robot/environment situations and given that the addition of a layer of internal units allowed these network to produce better performance in the experiment described above we expected that evolving robots would benefit from the possibility of exploiting internal representations in this case. We then ran three set of simulations in which robots were evolved for their ability to find and remain close to the cylindrical object and were provided with the three different architectures described above (a simple perceptron with 6 sensory neurons encoding the state of the frontal infrared sensors, a network with an additional layer with internal neurons, a network with an additional layer with 8 internal neurons. All networks had two motor neurons that were used to control the speed of the two corresponding wheels. Lifetime consisted of 5 epochs of 500 sensory/motor cycles for a total of 2500 motor cycles.

Also in this case however, we observed that individuals with pure sensory-motor neural controllers were able to solve the task. If we look at the fitness of individuals throughout generations in fact, we see how after few generations the best individuals spend most of their lifetime close to the target objects. These individuals in fact are able to avoid walls, explore the environment and keep close to the cylinder as soon as they find it (Figure 6). Notice also how individuals with additional internal units do not achieve better performance but on the contrary achieve significantly lower performance on the average.

Figure 7 shows the behavior of a typical evolved individual. The robot, after being placed in front of a wall on the right hand side of the environment, recognized and avoided the wall, recognized and avoided the new wall it encountered, and finally found and kept close to the target. All individuals, like

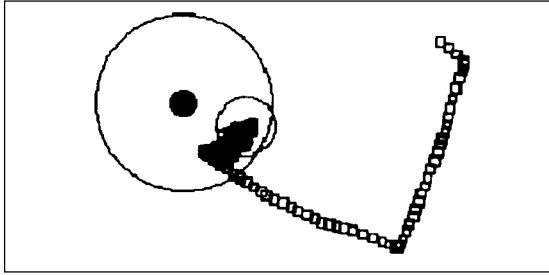


**Figure 6.** Number of cycles spent close to the target (out of 2500) for simulation without hidden units and with 4 and 8 internal units. The top graph shows the performance of the best individual of each generation. The bottom graph shows the average performance of each generation. Each curve represents the average result of 10 replications.

the one shown in the figure, never stopped in front of the target, but start to move back and forth as well as slightly left and right, while remaining at about the same angle and distance with respect to the target.

## 7. Discussion

We presented a set of experiments in which mobile robots were evolved for the ability to perform different tasks: exploring an arena, navigating toward a



**Figure 7.** Behavior of a typical evolved individual without internal units. The lines represent walls, the filled circle in the center of the arena represents the target object, the large empty circle around the target represents the area in which the robot is rewarded, the small empty circle represents the position of the robot after 500 cycles. Finally the outline path on the terrain represents the trajectory of the robot.

certain position, and discriminating between objects with different shapes. By comparing the results obtained with robot provided with pure sensory-motor neural controllers, with neural controllers with internal representations and/or internal dynamics we observed that in all cases pure sensory-motor controller were able to develop close to optimal performance. Moreover, the addition of internal layers or recurrent connections did not produce any advantage.

These results show the power of sensory-motor coordination. By coordinating the sensory and motor processes in fact, these robots were able to solve rather complex task that from the point of view of an external observer apparently require internal representation and/or internal dynamics. This does not imply that internal representations are useless or that any task can be solved with a pure sensory-motor controller. However, the obtained results suggest that care must be taken in claiming that a particular task necessarily requires solutions relying on internal representations or internal dynamics.

The real question therefore becomes: under what conditions one could expect internal representations and/or internal dynamics to emerge without preventing the use of sensory-motor coordination? There are at least three possible directions that can be investigated:

- a. Using experimental frameworks that have quantitatively less regularity and therefore require more general solutions. As we have seen, pure sensory-motor solutions are simple and effective and can solve rather complex problems. However, they tend to heavily rely on the regularities of the current robot/environment situation. It is possible that by reducing the number of

regularities the effectiveness of pure sensory-motor solutions decreases and as a consequence the pressure to select strategies relying on internal representations or internal dynamics increases. Consider for example the exploration task that has been described in Section 2. We saw that if evolving individuals are trained in an environment with a fixed shape and dimension evolution can easily find pure sensory-motor solutions. In an experimental framework in which individuals are exposed to environments with varying shapes and sizes, pure sensory-motor solutions may become significantly less effective than solutions relying on internal representations or internal dynamics. In such a scenario, it is more likely that the latter will be selected.

b. Progressively increasing the richness of the experimental framework or adapting individuals in ever changing environment. Another scenario in which one may expect that sensory-motor solutions become ineffective is the case of tasks that are quantitatively (although not qualitatively) more rich. Consider for example a discrimination task like that described in section 4 in which the number of objects to be discriminated is much larger or in which the shape of the object might slightly change throughout generations. Pure sensory-motor strategies might be less effective in this case and therefore the selection pressure to develop strategies that rely on internal representations and/or internal dynamics will increase.

c. Including additional processes such as lifetime learning that may force the system to develop internal representations or internal dynamics. To enhance their adaptation level, evolving individuals should develop useful internal representations and/or internal dynamics and, at the same time, develop an ability to translate this information into more effective actions. Having good internal representations but not being able to use them or viceversa having the ability to use internal representations but not having the right representations do not produce any adaptive advantage and unfortunately the probability that these two mechanisms can be developed at the same time through random mutations or recombination is rather low. Good internal representations and/or internal dynamics however can be developed through lifetime learning. If internal representations and/or internal dynamics are developed by an independent process, such as ontogenetic learning, evolutionary changes that increase the ability to exploit these internal structure will produce immediately an adaptive advantage and will be retained. For example, evolving individuals may be selected for the ability to perform a certain task and at the same time

can be asked to predict the consequences of their action through prediction learning (c.f. Nolfi et al. 1994). In order to predict these individuals have to develop internal representations and internal dynamics that are coupled with the environment. These internal structures can be exploited by evolution to increase the adaptation level of evolving individuals with respect to the evolutionary task.

## References

- Asada, M. 1990. Map building for a mobile robot from sensory data. *IEEE Transaction on Systems, Man, and Cybernetics*, 6, 1326–1336.
- Bakker, P., & Y. Kuniyoshi 1996. Robot see, robot do: an overview of robot imitation. In N. Sharkey (Ed.), *Proceedings of the AISB Workshop on Learning in Robots and Animals*. Brighton, UK: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
- Beer, R. 1995. A dynamical system perspective on agent-environment interaction, *Artificial Intelligence*, 1, 173–215.
- Brooks, R. A. 1991. New approaches to robotics. *Science*, 253, 1227–1232.
- Chalmers, D. 1995. Facing up the problem of consciousness. *Journal of Consciousness Studies*, 2, 2–19.
- Cheng, K. 1986. A purely geometric module in the rat's spatial representation. *Cognition*, 23, 149–178.
- Cliff, D., I. Harvey, & P. Husband. 1993. Exploration in evolutionary robotics. *Adaptive Behavior*, 2, 73–110.
- Fodor, J. A. 1983. *The Modularity of Mind*. Cambridge, MA: MIT press.
- Gallistel, C. R. 1990. *The Organization of Learning*. Cambridge, MA: MIT Press.
- Lund, H. H., & J. H. Hallam 1996. Sufficient neurocontrollers can be surprisingly simple. Technical Report. University of Edinburgh: Department of Artificial Intelligence.
- Margules, J., & C. R. Gallistel. 1988. Heading in the rat: determination by environmental shape. *Animal Learning and Behavior*, 16, 404–410.
- Miglino, O., K. Nafasi, & C.E. Taylor. 1994. Selection for wandering behavior in a small robot. *Artificial Life*, 2, 101–116.
- Miglino, O., H.H. Lund, & S. Nolfi. 1995. Evolving mobile robots in simulated and real environments. *Artificial Life*, 4, 417–434.
- Mondada, F., E. Franzi, & P. Ienne. 1993. Mobile robot miniaturization: A tool for investigation in control algorithms. In T.Y. Yoshikawa, & F. Miyazaki (Eds.), *Proceedings of the Third International Symposium on Experimental Robots*. Berlin: Springer-Verlag.
- Nilsson, N. 1984. *Shakey the robot*. Technical Note 323, SRI International, Menlo Park, California.
- Nolfi, S. 1996. Adaptation as a more powerful tool than decomposition and integration. In T. Fogarty & G. Venturini (Eds.), *Proceedings of the Workshop on Evolutionary*

- Computing and Machine Learning, 13th International Conference on Machine Learning, Bari, Italy: University of Bari.
- Nolfi, S., J. Elman, & D. Parisi. 1994. Learning and evolution in neural networks. *Adaptive Behavior*, 3, 5–28.
- Nolfi, S., D. Floreano, O. Miglino, & F. Mondada. 1994. How to evolve autonomous robots: Different approaches in evolutionary robotics. In R. A. Brooks, & P. Maes (Eds.), *Artificial Life IV. Proceedings of the Fourth International Workshop on The Synthesis and Simulation of Living Systems*. Cambridge, MA: MIT Press.
- Nolfi, S.. 1997. Using emergent modularity to develop control systems for mobile robots. *Adaptive Behavior*, 3–4, 343–364.
- Piaget, J. 1952. *The Origins of Intelligence in Children*. New York: International University Press.
- Piaget, J. 1971. *Biology and Knowledge: An Essay on the Relations Between Organic Regulations and Cognitive Processes*. Chicago: Chicago University Press.
- Pfeifer, R., & C. Scheier. 1999. *Understanding Intelligence*. Cambridge, MA: MIT Press.
- Sharkey N.E., & J.N. Heemskerk. 1997. The neural mind and the robot. In A.J. Browne (Ed.), *Neural Network Perspectives on Cognition and Adaptive Robotics*. Bristol U.K.: IOP Press.
- Tani, J. 1996. Model-based learning for mobile robot navigation from the dynamical system perspective. *IEEE Transactions on Systems, Man, and Cybernetics*, 3, 421–436.

# Ago Ergo Sum\*

Dario Floreano

Swiss Federal Institute of Technology at Lausanne (EPFL)

## 1. Paving the road to robot consciousness

Could the ordered list of operations which forms any computer algorithm constitute the basis of consciousness for a robot? “Of course not”, would be the sensible answer probably given by most readers with a scientific or engineering oriented mind. The popular justification for such a view is that computer programs are instructions used to control machines which must carry out a specific task, such as a lengthy computation or assembly of car components. A robot controlled by such an algorithm would be nothing more than an automaton, a mechanical device which blindly and repetitively executes a pre-specified set of instructions (see Penrose 1994, chapter 1] for elegant elaborations of this view). In this paper, I shall consider a different position and argue that it is not possible to exclude that some of today’s robots might indeed possess a form of *proto-consciousness* whose substrate is a mere algorithm, or something that can in principle be described in terms of algorithmic computation (e.g., an artificial neural network, a rule-based system, a classifier system, etc.). I will do so by showing why this might be possible at all and by describing the basic ingredients that, if accurately mixed together, might give rise to forms of conscious behaviour. I will support my line of reasoning with examples of behaviours recorded during experiments of artificial evolution performed on mobile robots at our laboratory.

All of what follows assumes that consciousness exists, if not as a reality, at least as a scientific hypothesis worth considering. Before proceeding any further, though, it is necessary to explain exactly what I mean by consciousness here in order to make my ideas testable and concretely debatable.

### 1.1 Consciousness as awareness

Consciousness is difficult to define in scientific terms because it has many facets and attempts to provide an exhaustive description often involve a good

deal of introspection. Whatever consciousness might be, it certainly is a product of natural evolution, the result of several generations of selective reproduction. Given the fierce competition for energy and resources that characterises all aspects of life, it is reasonable to assume that consciousness provided some survival advantage for organisms that began to exhibit it, even though its primordial manifestations might have been an interesting byproduct of evolution with no negative impact on inclusive fitness. If we consider consciousness as an ability for better coping with the difficulties faced by a living organism, then it becomes inevitable that its characterisation should bear some relationship to the basic challenges on which survival and success of reproduction is based. These include, in increasing order of complexity (from Clark 1989): sensori-motor coordination and proprioception, object recognition and recognition of spatio-temporally invariants, spatial navigation, multi-modal sensory fusion and logical processing, social behaviour. The common factor underlying all these abilities is the required capacity of an organism to position itself with respect to the external environment, that is to achieve increasingly complex levels of situatedness.

Therefore, among all the possible definitions, here consciousness is taken to be *awareness* of one's own internal state with respect to the environment. Being conscious is the process by which an intelligent control system performs spontaneous self-monitoring of internal states (which could take into account not only neurally generated activity, but also physiological states) by putting them in relation to the external environment. Thus, although consciousness is an internal reflexive activity, it is not purely subjective and disjointed from the external environment (which would rather correspond to hallucination), but is intimately linked to purposeful behaviour. Within this framework, consciousness becomes useful because it allows comparison and processing of information coming from several sensory modalities and it provides the system with the ability to make predictions in order to change its course of actions. By spontaneously monitoring, cross-correlating, and variously arranging several internal processing states in relation to what happens in the environment, one can anticipate different behavioural outcomes and act accordingly (see also Dennett 1991). This type of consciousness also requires active exploration. When the system realises that there is not sufficient information for performing an action, or feedback from the environment does not correspond to the internally anticipated situation, the organism will engage in exploratory behaviour in order to actively seek missing information or in order to change some of its own internal parameters. I think that this type of consciousness is one of the

characteristics that distinguishes purely reactive organisms from intelligent organisms. In simple words, consciousness is a way of continuously refreshing and updating one's position in the world, a way of making sense of what is happening by coordinating relevant information and of anticipating future situations for appropriate action.

The definition given above is quite reductive, but I think that it captures the essential elements of what consciousness is, the basic purposes of conscious activity in humans as well as possibly in other organisms. To this extent, my analysis of consciousness as awareness of one's own state refers to a type of "proto-consciousness", similar for several aspects to the definition of "primary consciousness" given by Edelman 1989, which does not exclude other phenomena, such as feelings, sensations, etc., as long as they participate to the goal of establishing and maintaining the position of the organism in its own environment. However, I deliberately do not consider phenomenal consciousness here, which is largely based on introspection and might be characteristic of human beings only or a pure artifact of our reasoning abilities.

In the remainder of this paper, I will try to show that a careful consideration of some evolutionary issues might indeed give us a methodology to create with our current technology conscious — that is, aware — artificially living organisms.

## 2. Prerequisites

One reason why attributing consciousness to an algorithm sounds weird is that we tend to see algorithms as abstract instruction sets designed for a specific computational purpose. I agree that such an algorithm cannot exhibit any form of awareness. The type of interactions between such an algorithm and the external world —if any— is precisely defined and regulated by the conditional statements (e.g., if-else) which form the algorithm itself and guide its operation toward the achievement of a pre-specified goal. There are no reasons whatsoever why such an algorithm should be aware of its position in the environment, change its course of action or its processing modality.

### 2.1. Embodiment

Things might change when the algorithm becomes *embodied*, that is when it is part of a body equipped with a sensori-motor system; now the algorithm is called

a “control system”. A body is a physical entity which interacts with a physical environment in forms which might drastically affect the processing modality of its own controller. For example, let us consider information processing in a certain type of algorithm, an artificial neural network which maps input vectors into appropriate output vectors. Here the sequence of input-output vectors (both during the learning phase and during normal functioning) is quite arbitrary, usually random or in accordance with a certain schedule which has been carefully chosen by the external user in order to maximise a well-defined objective function. Instead, when the neural network is embodied in a sensorimotor system (wherein it is often called a neurocontroller), the *time sequence* of the input vectors (sensory information) depends on the output (motor actions) of the network itself, which in turn depends on the input. In this case the flow of information between the network and the environment is coherent, meaningful, and self-contained (in the sense that it is partially or completely generated by the information-processing system itself) (Parisi et al. 1990). The interaction with the environment affects not only the time sequence, but also the *type* and *frequency* of information which is available to the network. That means that the type of experiences to which such a network is exposed greatly depends on the network itself. Such a neural network could actively avoid or seek certain situations as well as potentially affect its own learning modality. If the control system was not part of a physical body struggling for survival, all this potential freedom could degenerate in un-interesting behavioural loops where the system closes on itself and minimises interactions with the external world. However, bodies have physiological demands — such as keeping an adequate temperature level and maintaining a sufficient amount of energy — which continuously urge the organism to develop better and more efficient behavioural strategies in order to satisfy them.

If we sum up the consequences of embodiment mentioned above, it becomes evident that such a control system would gain considerable advantage from the type of conscious activity described in section 1.1, that is the ability to monitor and regulate internal and physiological states; to seek, integrate, and coordinate the multi-modal flow of information in order to establish appropriate correlations, highlight important bits, and suppress the rest; to distinguish between environmental changes that are due to its own actions and those that are due to external causes (in other words, to understand the sensory consequences of its own actions Parisi et al. 1990); to anticipate future situations in order to find a timely solution and avoid dangerous situations. These consider-

ations apply both to living organisms and to artificial organisms, such as robots where the physiological demands might correspond to the ability to maintain an appropriate reserve of electrical power or to minimise mechanical wear.

## 2.2. Autonomy

However, embodiment alone is not sufficient to make us believe that some type of computation, be it biological or artificial, might exhibit consciousness. After all, there are several examples of robots that are precisely pre-programmed to respond in precise ways to incoming information; e.g., the already mentioned robots working on industrial assembly lines. In that case, a human analyst designs the control system of the robot so that it performs exactly and precisely certain actions in response to a restricted class of environmental conditions in order to carry out a pre-defined task. Here, the robot does not have freedom of exploration, it cannot establish its own goals and sub-goals, or change the way in which it acts; in other words, such a robot would not display any awareness because the flow of information from sensors to actuators has been rigidly specified in detail by an external designer for the purpose of achieving an externally defined goal. In fact, another important pre-requisite of conscious activity — both in living and artificial organisms — is *autonomy*. Autonomy implies freedom from external control (McFarland 1992). An autonomous agent is a system that is capable of defining its own goals and regulating its own behaviour accordingly; in other words, it must be capable of self-control (Steels 1993). Autonomy is often associated with self-sufficiency, but the two notions are in fact quite different. For example, a bird raised in a cage is not self-sufficient, and if put in the wild it will probably die because it has not developed efficient foraging strategies; nevertheless, it is an autonomous agent. Similarly, a robot might be autonomous even though it depends on external provision of electrical power. The definition of autonomy is concerned with the locus of control, i.e. with who decides the goals of the agent, not with the capacity for self-sufficiency.

## 2.3. Adaptation

The notion of autonomy clashes with the traditional method of engineering behaviours for robots, which could be summarised in the following logical and temporal steps: (1) Create a formal model of the robot and of the environment;

(2) Analyse the task requirements and decompose it in appropriate subtasks; (3) Design an algorithm which steers the robot in accordance with the formal understanding gained in the previous steps. Such an agent could hardly be described as autonomous. As McFarland has put it (McFarland et al. 1993), creation of autonomous agents requires a shift in behaviour engineering from a goal-directed to a goal-seeking and self-steering control system (see also Pfeifer et al. 1992). An autonomous agent cannot be programmed in the traditional way described above. Rather, its control system should be equipped with *learning* mechanisms that give the agent the ability to adapt to the environment and to develop its own strategies in order to maintain itself in a viable state. A very popular choice consists in using neurocontrollers — artificial neural networks where the input units are clamped to the robot sensors and the output units control the effectors, such as wheels, grippers, etc. — which are well suited structures for learning mechanisms (for examples, see some recent special journal issues on robot learning Gaussier 1995; Dorigo 1996). Learning consists in a gradual modification of the internal parameters (synaptic connection strengths, neuron thresholds, and architecture) while the robot interacts with the environment. However, one might envision a variety of different processing metaphors (such as classifier systems) that are capable of autonomous self-configuration. What really matters in these adaptation techniques is to what extent learning proceeds according to an externally defined goal and schedule rather than resulting from a process of self-organisation. These two learning modalities are often referred to as supervised and unsupervised, respectively (e.g., see Hertz et al. 1991). Probably, it is impossible to draw a sharp line between them because in both cases adaptation takes place in order to optimise an objective function (e.g., a cost function or an energy function), which seems to hold for learning and behaviour regulation in living systems too (McFarland et al. 1993). The difference between learning automata and learning autonomous agents, then, lies in the type of information which is made available to the agent for learning and in the way it is used. For example, modifying the synaptic strength by backpropagation of error (Rumelhart et al. 1986) between the agent's actions and the actions defined by a human observer, is an efficient — when feasible — way of programming automata (e.g., see Pomerleau 1993); there, learning is aimed at achieving a specific goal by implementing a detailed behavioural schedule predefined by an external user. Instead, a more suitable technique for creating autonomous agents might be provided by reinforcement learning (for a clear review, see Barto et al. 1990), where learning takes place only when the agent receives a reward signal from the environment. Reinforcement

learning algorithms try to maximise the probability of repeating behaviours which lead to positive reinforcement signals. Despite being a promising approach, these algorithms require an extensive initial exploration of the possible behavioural outcomes in order to develop stable controllers, which does not seem always to be the case in living organisms.

In this paper, I will consider a different adaptation mechanism, *artificial evolution of neurocontrollers*, as a viable methodology to develop autonomous agents that could exhibit conscious abilities. Artificial evolution differs from other learning schemes because it works on a population of different individuals and is based on a selectionist, rather than goal-directed, approach (Steels 1993).

### 3. Artificial evolution of autonomous robots

Application of evolutionary techniques for automating the development of robotic control systems has been tried with success by several researchers (see Mataric et al. 1996 for an extensive review). Also referred to as “Evolutionary Robotics” (Cliff et al. 1993), this approach relies on at least two interesting issues: (a) the possibility of replicating and understanding basic principles of adaptation in artificial forms of life (Langton 1990), and (b) the generality and the power of evolutionary techniques to develop control systems exhibiting complex behaviours which would otherwise be difficult to design by hand. In several cases, the basic building blocks specified in the artificial genomes are networks of real-time recurrent neurons, which are parallel processing systems well-suited for operating in real-world conditions (Harvey et al. 1993) and for potentially exhibiting complex and minimally cognitive behaviours (Yamauchi et al. 1995; Beer 1996). Although the specific algorithmic instantiation of the control system is not crucial for creating autonomous agents, the definition of consciousness given above requires some type of feedback and lateral flow of information among the parallel processes that compose the control system. These information channels, which are easily implemented as recurrent connections in artificial neural networks, might serve the purpose (often rethorically attributed to an internal *homunculus*) of establishing appropriate multi-modal correlations, of coordinating internal states, of monitoring and changing the course of actions on the basis of previous experience, and of anticipating future internal and sensory states (see also Edelman 1992) for an extensive review of the relevance of “re-entrant connections” in living and artificial conscious systems).

Recently, the well-known issue of agent simulation versus physical implementation has been re-examined with respect to evolutionary robotics. All evolutionary techniques operate and rely on populations of different individuals, several of which display maladaptive behaviours, especially in the initial generations. This represents a heavy burden for physical implementations on real robots in terms of time and resources. Therefore, most of the experiments of evolutionary robotics are carried out in simulations, and only in few cases is the evolved controller transferred on the real robot (e.g., see Nolfi et al. 1994; Miglino et al. 1996). This procedure has been criticised because it might not scale well as soon as the robot geometry and dynamics become more complex (Mataric et al. 1996). Although for the purpose of the arguments offered in this paper it does not really matter whether the agent is simulated or physically implemented, all the experiments that I will report below have been entirely carried out on a real mobile robot.

### 3.1. Artificial evolution vs. evolutionary optimisation

Computationally speaking, evolutionary methods represent a very general technique because they can be applied to any aspect of the agent (synaptic connections and number of neurons (Yao 1993), growing factors (Cangelosi et al. 1994), body specifications (Lee et al. 1996)), etc. and to any type of algorithm instantiation (neural networks Rudnick 1990, classifier systems Dorigo et al. 1993, graph systems Kitano 1990, programming language functions Koza 1992, etc.) as long as they can be mapped into a genetic description. It has been empirically shown that genetic algorithms outperform other search methods when the space of the possible solutions is high-dimensional and non-differentiable (see Goldberg 1989 for a simple description of where and why evolutionary search outperforms other search methods). These two properties, generality and efficiency, are often exploited to optimise unknown parameters of a complex system so that it will exhibit a well-defined behaviour (e.g., see Baluja 1996). To the extent that behaviour generation is externally defined and driven to a specific goal, evolutionary optimisation is not very different from the supervised learning approach described in section 2.3. However, natural evolution does not have the notion of teleology which is so familiar to biologists (Atmar 1993) and engineers, but is rather an open-ended process (see also Harvey 1992) where the concept of optimum cannot be universally defined. It would be wrong to assume that a specific organ, such as the retina, has evolved in order to maximise a specific function, such as transmission of reflected light.

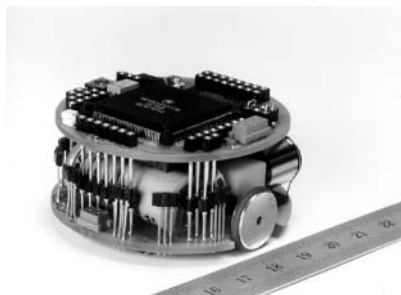
Similarly, there are no goals driving evolutionary development of cognitive systems; where goals exist, they are autonomously created by the agent itself as a more efficient way of organising its own survival strategies. If taken to its extreme consequences, this reasoning would imply that artificial evolution should not make use of any externally defined fitness function and perhaps it should substitute selective reproduction with elimination of individuals which cannot keep themselves viable. However, also in this case, it would be inevitable to introduce some external bias by deciding what viability is in the specific artificial implementation of the ecosystem. Indeed, it is very difficult to draw a line between evolutionary optimisation and artificial evolution. With respect to the issue of evolvability of conscious control systems, we can re-state these differences by saying that the more consistent is the shift from evolutionary optimisation to artificial evolution, the bigger is the probability that the evolved controllers exhibit consciousness.

In the next sections, I will try to support this statement by describing two experiments of evolutionary robotics where more complex neurocontrollers can be evolved by simply lifting some constraints on the fitness function and considering the robot as an organism with its own physiological requirements. Although still simple, these controllers exhibit several of the computational and behavioural characteristics listed above as indicators and prerequisites of conscious activity.

### 3.2. Some experiments in evolutionary robotics

The experiments described here have been carried out on a real mobile robot without human intervention. In all cases we employed a single physical robot which served as body for several populations of individuals which were tested one by one. Although this procedure is biologically implausible because in biological life control systems and bodies co-evolve, hardware evolution is still a challenging technical issue (see Lee et al. 1996 for some initial attempts in this direction). Each individual control system was a neural network where the input units were clamped to the robot sensors and the output unit activations were used to set the rotation speed of the wheels. Different neural architectures were used in the various experiments, but they all had sigmoid activation function and recurrent connections.

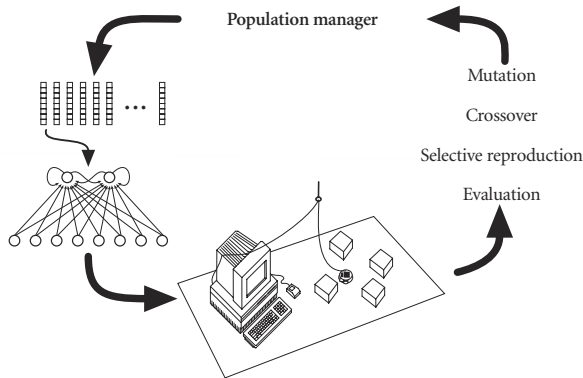
The robot employed in the experiments is Khepera, a miniature mobile robot (Mondada et al. 1993b). It has a circular shape (Figure 1), with a diameter of 55 mm, height of 30 mm, and weight of 70 g, and is supported by



**Figure 1.** Khepera, the miniature mobile robot used for the experiments of artificial evolution.

two wheels and two small Teflon balls. The wheels are controlled by two DC motors with incremental encoder (10 pulses per mm of advancement of the robot), and can rotate in both directions. In the basic configuration, the robot is provided with eight infrared sensors, six positioned on one side of the robot (front), the remaining two on the other side (back). Each sensor can function in two modalities: it can emit and measure reflected infrared light and it can measure the infrared light component of the ambient light. When functioning in the first modality, it is called a proximity sensor because it is more active when it is closer to an object; when functioning in the second modality, it is called a light intensity sensor because it gives a rough indication of the ambient light intensity at its location. A Motorola 68331 controller with 256 Kbytes of RAM and 512 Kbytes ROM manages all the input-output routines and can communicate via a serial port with a host computer.

Khepera was attached via a serial port to a Sun SparcStation 2 by means of a lightweight aerial cable and specially designed rotating contacts (see Mondada et al. 1995 for more detailed descriptions of technical issues related to the experiments described in this paper). In this way we could exploit the power and disk size available in a workstation by letting high-level control processes (genetic operators, neural network activation, variables recordings) run on the main station while low-level processes (sensor-reading, motor control, and other real time tasks) run on the on-board processor (Figure 2). Thus, while the robot was operating, we could keep track of all the populations of organisms that were born, tested, and passed to the genetic operators, together with their “personal life files”. At the same time, we could also take advantage of specific software designed for graphic visualization of trajectories and sensory-motor status while the robot was evolving (Cheneval 1995). As



**Figure 2.** Operating methodology

already stated in section 2.2. the fact that the robot depended on an external source of energy does not affect its autonomy. Also, for what concerns Khepera, the robot is not aware of where its own “brain” is located as long as it is connected to its own sensors and motors.<sup>1</sup>

The evolutionary procedure was a standard genetic algorithm as described by Goldberg (Goldberg 1989) with fitness scaling and roulette wheel selection, biased mutations (Montana et al. 1989), and one-point crossover. Each individual of a population was in turn decoded into the corresponding neural networks, the input nodes connected to the robot sensors, the output nodes to the motors, and the robot was left free to move for a given number of steps (motor actions) while its fitness  $\Phi$  was automatically recorded. Each motor action lasted 300 ms. Between one individual and the next, a pair of random velocities was applied to the wheels for 5 seconds.

### 3.3. Automatic evolution of behavior

The first experiment was chiefly aimed at testing the evolutionary approach on a real mobile robot and assessing advantages and difficulties (see Floreano et al. 1994 for more details). We decided to evolve a neural network with a single layer of weights to perform straight navigation and obstacle avoidance. The robot was put in an environment consisting of a sort of circular corridor whose external size was approx. 80x50 cm large (Figure 3). The walls were made of light-blue polystyrene and the floor was a thick gray paper. The robot could sense the walls with the IR proximity sensors. Since the corridors were rather narrow (8–12 cm), some sensors were slightly active most of the time. The



Figure 3. Environment of the experiment.

environment was within a portable box positioned in a room always illuminated from above by a 60-watt bulb light. A serial cable connected the robot to the workstation in our office a few rooms away.

The fitness function  $\Phi$ , was explicitly designed for straight navigation and obstacle avoidance

$$\begin{aligned}\Phi &= V(1 - \sqrt{\Delta v})(1 - i) \\ 0 &\leq V \leq 1 \\ 0 &\leq \Delta v \leq 1 \\ 0 &\leq i \leq 1\end{aligned}\tag{1}$$

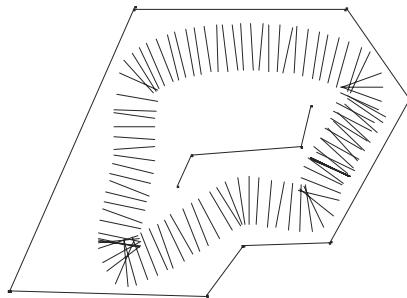
where  $V$  is a measure of the average rotation speed of the two wheels,  $\Delta v$  is the absolute value of the algebraic difference between the signed speed values of the wheels (positive is one direction, negative the other), and  $i$  is the activation value of the proximity sensor with the highest activity. The function  $\Phi$ , has three components: the first one is maximized by wheel speed (but not direction of wheel rotation), the second by straight direction, and the third by obstacle avoidance. The second component was very important. Without it, a robot could maximise its own fitness by simply spinning on itself at very high speed on a location far from obstacles, a trivial solution which can be found by simply setting a very high threshold value for one of the motor neurons and a strong synaptic value from any of the sensors to the other motor neuron. The square root was introduced after a series of trials and errors and was aimed at emphasising penalisation of small differences between the two wheel rotations for that specific training environment. Without it, the controller would set in sub-optimal solutions consisting in large circular trajectories.

In less than 50 generations, corresponding to approximately 24 hours of uninterrupted evolutionary adaptation, the best neurocontrollers of the population displayed smooth trajectories around the arena without hitting the walls (Figure 4). Thanks to the generalisation properties of artificial neural networks, these neurocontrollers could navigate in any type of environment (in different light conditions and with obstacles of different reflectance). These results were replicated several times by restarting the experiment from new randomly initialised populations. The resulting controllers cannot be properly called autonomous, but evolution did autonomously find a number of interesting solutions to the problem of navigation. For example, although the robot had a perfectly circular shape and the wheels could rotate in both directions, all the best neurocontrollers developed a frontal direction of motion corresponding to the side with more sensors, which gave a better resolution of the sensory information permitting better obstacle detection.

### 3.4. Evolution of an autonomous system

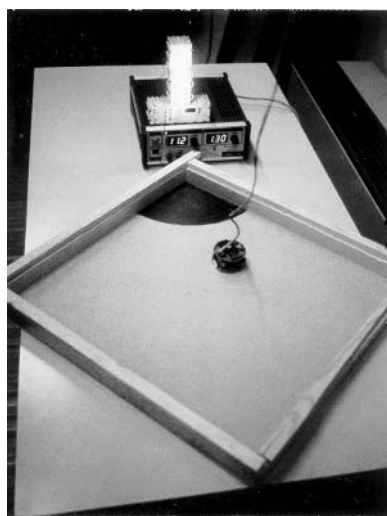
Having ascertained that evolutionary techniques can be successfully applied to a real mobile robot, we decided to give the system more autonomy of development by applying the following changes:

- simplify the fitness function by eliminating the middle component;
- provide the robot with its own physiology, that is a limited life duration controlled by a rechargeable battery;
- make the environment more “interesting” by introducing a battery charger and a landmark.



**Figure 4.** The trajectory performed by one of the evolved robots. Segments represent successive displacements of the axis connecting the two wheels. The direction of motion is anti-clockwise.

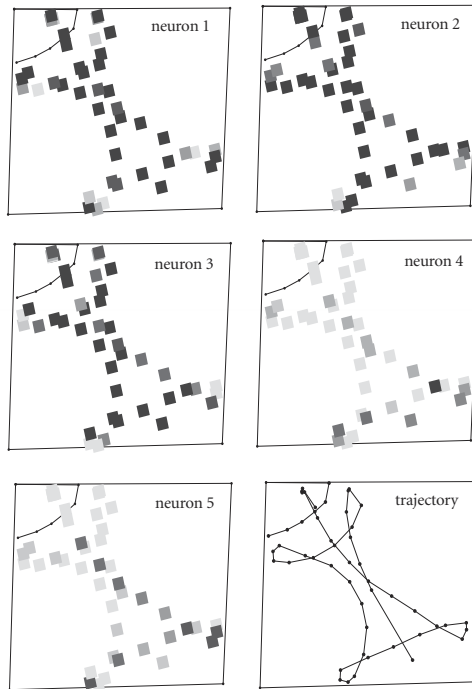
The environment employed for evolutionary training consisted of a 40x45 cm arena delimited by walls of light-blue polystyrene and the floor was made of thick gray paper (Figure 5) as in the previous experiment. A 25 cm high tower equipped with 15 small DC lamps oriented toward the arena was placed in one corner. The room did not have other light sources. Under the light tower, a circular portion of the floor at the corner was painted black. The painted sector, which represented the recharging area, had a radius of approximately 8 cm and was intended to simulate the platform of a prototype battery charger. When the robot happened to be over the black area, its simulated battery became instantaneously recharged. The simulated battery was characterised by a fast linear discharge rate (max duration: approx. 20 seconds). In order to truncate the lives of individuals who would just sit on the battery charger or manage to recharge themselves, we set an upper life limit of 60 seconds for each robot, after which the next individual in the population was tested. By simulating the battery instead of using the onboard available batteries (which lasted much longer and also required much longer recharging time), artificial evolution produced interesting results in approximately 240 hours instead of 6 years! The neurocontroller received input from the eight proximity sensors, from



**Figure 5.** The environment of the experiment on battery recharge. The light tower is positioned in the far corner over the recharging area which is painted black. There are no other light sources in the room.

two ambient light sensors (each positioned on one side of the robot), from a detector of floor brightness placed underneath the robot platform, and from a sensor of battery level; these inputs projected to five internal units interconnected by lateral and recurrent connections, which then projected to two motor neurons controlling the wheels. The simplicity of the fitness function employed in this experiment may be exploited by a robot quickly spinning on the same spot far from the walls. The fitness function returned almost null values when the robot was on the battery charger (because it was positioned on a corner near the walls).

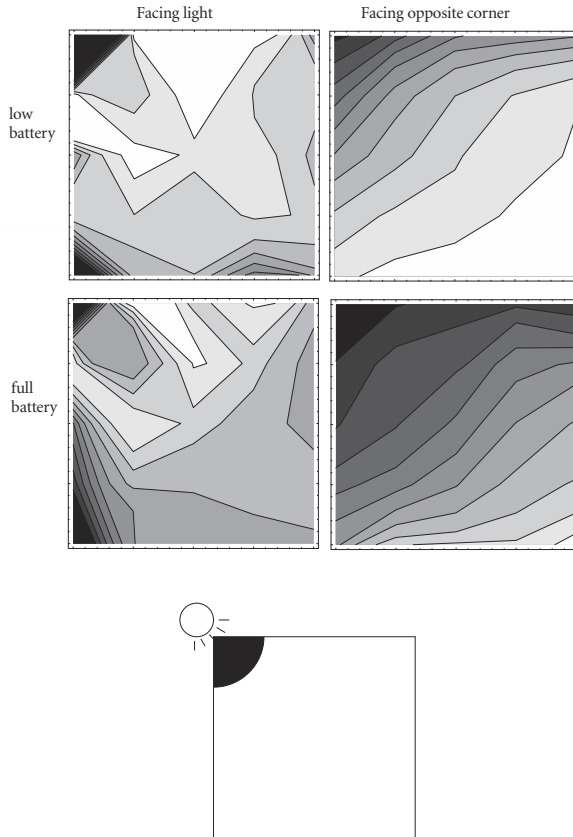
After ten days of continuous evolution (approx. 240 generations), the best neurocontroller of the population displayed interesting homing behaviour. By providing Khepera with a small “helmet” which gives us precise indication of where it is in the environment, we can correlate internal neural activity with its behaviour while it is autonomously moving in the environment. Figure 6 plots a typical trajectory of the robot (bottom-right) along with the activation of the five internal hidden units. Starting with a fully charged battery, the robot moves around the arena covering the whole area without hitting the walls and accurately avoiding the recharging zone. However, as soon as the battery level reaches a certain minimum value, the robot orients itself toward the recharging area and proceeds towards it on a straight line. Once recharged, it quickly exits the recharging area and resumes its large trajectories across the arena. Arrival to the recharging area is always extremely precise, approximately 1 second before total discharge. Since the trajectories around the environment are never the same, the robot autonomously decides what is the residual battery level when it must move toward the recharging area, depending on where the robot is in the environment. When it is far away from the recharging area, Khepera will begin to turn and move toward it at a much higher residual level than when it is closer to the area. From these and other tests reported in Floreano et al. 1996, it becomes clear that the robot accurately avoids the recharging area if the battery level has not yet reached a certain minimum level, but it goes toward it if that level has been surpassed. By looking at the internal dynamics of the neurocontroller while the robot freely moves in the environment (Figure 6), we notice that neuron 5 becomes active shortly before the robot starts orienting toward the recharging area. Other neurons display significant changes of activity levels only when the robot is close to the walls and are thus responsible for the quite automatic behaviour of obstacle avoidance. By positioning the robot at various locations in the environment and measuring the activity of neuron 5 for a single instant, we can see that it has developed



**Figure 6.** Visualisation of the activity of hidden units while the robot moves in the environment. Darker squares mean higher node activation. The robot starts in the lower portion of the arena. The bottom-right window shows the trajectory only. The recharging area is in the top left corner (Adapted from Floreano et al. 1996).

a map of the environment which varies depending on the orientation of the robot, but *not* on the battery level (Figure 7). Thus, neuron 5 creates an internal representation of where the robot is and correlates it with battery level information in order to switch the robot behaviour from simple navigation and obstacle avoidance into a precise and timely homing trajectory.

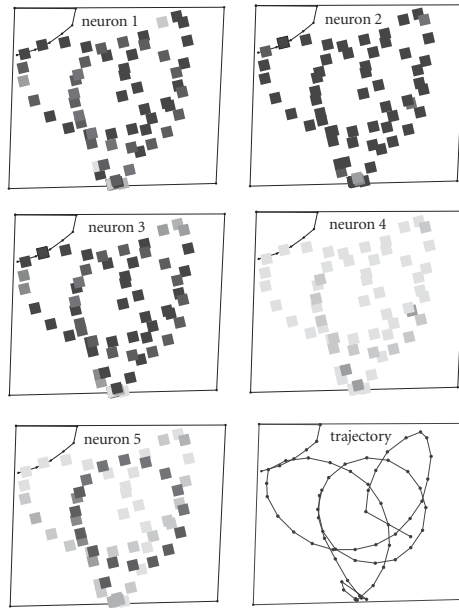
The switch from a rather automatic behaviour into an active “home-seeking” behaviour can be seen in a simple test where the robot is put in the environment with the light off. Now, its infrared proximity sensors can still detect the walls, but there is no more information for homing. Figure 8 shows the trajectory and neuron activations for this test. Initially, the robot is not disturbed by the dark environment and it performs the usual navigation and obstacle avoidance across the arena. As soon as the battery reaches a certain residual charge level (which now is much higher than in normal light condi-



**Figure 7.** Contour map of the activation of neuron 5 for 2 different battery levels. The measurements were taken by positioning the robot at several evenly-spaced locations in the environment (depicted at the bottom). When the robot is facing the recharging area, the activity map does reflect the paths taken by the robot (Adapted from Floreano et al. 1996).

tion), neuron 5 becomes active and the robot starts to perform a perfectly circular trajectory in the middle of the environment until it eventually stops operating. This circular path is a very smart exploratory strategy because, if there was some weak light source in the arena, the robot could detect it and follow its gradient.

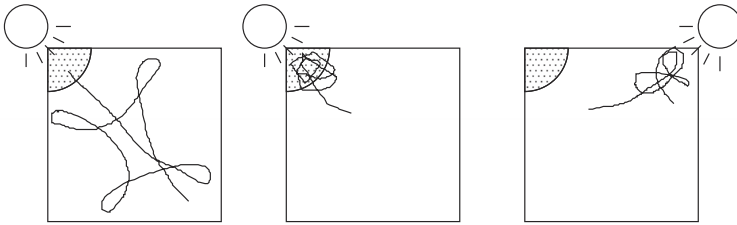
We can do some further tests to check the reactions of the robot in unexpected situations (Figure 9). If the battery is not automatically recharged once the robot has arrived to the black area, it will continue to stay on it waiting



**Figure 8.** Visualisation of the activity of hidden units while the robot moves in the environment. Darker squares mean higher neuron activation. The robot starts in the lower portion of the arena. The bottom-right window shows the trajectory only. The recharging area is in the top left corner (Adapted from Floreano et al. 1996).

for the recharge until it will eventually stop operating (Figure 9, centre). Similarly, if the light tower is moved to a different corner of the environment, but the recharging area remains in the same position, the robot will move toward the light; as soon as it does not detect the black painted surface, it will start searching in the surroundings until it will eventually stop operating (Figure 9, right).

Although re-adaptation to changes of landmark position can be easily achieved in a few generations of continued evolution (Floreano 1996), the system would greatly benefit from a combination of phylogenetic evolution and ontogenetic learning. The basic idea is that, instead of evolving neurocontrollers with fixed synaptic weights, it might be possible to evolve plastic networks that change their synaptic strengths according to evolved plasticity rules while the agent interacts with the environment. Preliminary successful results have already been achieved using the environment and fitness function described in section 3.3 (Floreano et al. 1996), but they are not reported here



**Figure 9.** Trajectories of best individual of generation 240 in three environmental conditions. Left: Test in training conditions, from Figure 6. The robot starts with a full battery in the bottom right corner (only the first 50 actions are displayed). Centre: The robot starts in the centre of the environment with an almost discharged battery; the battery is not automatically recharged when the robot arrives on the charging area. Right: The light source is positioned on the top right corner, but the charging area remains at the original location.

because they were mainly aimed at assessing technical feasibility rather than at developing a fully autonomous agent.

None of the abilities and behaviours described above were explicitly described in the fitness function. However, the fact that longer life duration could correspond to the accumulation of more fitness points, evolutionary pressure selected individuals that:

- autonomously discovered the presence of the recharging area;
- correlated internal physiological requirements with the robot location in the environment to decide when to switch behaviour so as to avoid battery discharge;
- created an internal representation of the environment to perform efficient homing navigation;
- could substantially change their behaviour depending on their physiological state (battery level);
- autonomously decomposed the global behaviour in subgoals, i.e. automatic navigation with obstacle avoidance and homing navigation, which were managed by different internal processes;
- could actively search for missing information;
- displayed meaningful behaviours when “environmental reward” (battery recharge) was not available, in one case waiting for the expected recharge, and in the other case searching for the sensory cues associated with recharge.

These abilities reflect the type of activity and behavioural outcomes that were listed as indicators of conscious awareness in section 1.1; also, the algorithm

that controls the robot was created in accordance with the prerequisites discussed in Section 2.

### **What is wrong with artificial consciousness?**

In this paper, I have emphasised a specific aspect of consciousness, i.e. awareness of one's own position with respect to the environment, using evolutionary and behavioural arguments; I have then argued that some of today's robots might indeed exhibit such conscious ability even when their internal activity can be described by an algorithm, provided that the prerequisites of embodiment, autonomy, and adaptation are met. Finally, I have given some examples of artificially evolved neurocontrollers whose internal dynamics and corresponding behaviours match the requirements for conscious awareness. In my arguments, purposeful interaction with an environment plays a fundamental role. Whereas Penrose, who thinks that consciousness cannot be understood and replicated by means of current algorithmic computation, considers the environment as a potential source (quickly dismissed) of non-computability for an algorithmic system which would then raise the possibility of conscious phenomena (Penrose 1994, pp. 152–154), the environment here is important because it gives meaning to an otherwise abstract processing system and it drastically affects its processing modalities. Behaviour, not thinking, is the basis of conscious activity.

My conclusions are intentionally provocative. The essence of my argument is that if there is something wrong with the notion of consciousness in an algorithm-driven machine, the problem is not to be found in the algorithm or in the machine, but in the definition of consciousness. In all the definitions of consciousness I come across, there is always some element of introspection and subjectivism, if not poetry, which makes difficult any scientific conclusion on this issue and complicates the debate. Although here I have tried to take a definition of consciousness (or rather “proto-consciousness”) which could form the basis for debate or at least scientifically-motivated criticism, it is not clear whether one should introduce a new label like “consciousness” instead of simply using well-understood concepts like attention, cross-modal fusion, coordination, planning, etc.

What I have attempted to show is that current scientific and technological methods are sufficient to recreate artificial forms of life that display characteristics of biological intelligence. Therefore, it cannot be ruled out that these

same artificial organisms could display forms of consciousness, whatever consciousness might be. After all, my conclusion is not a big conceptual advancement with respect to what Thomas Huxley said in 1874 during an invited address at the meeting of the British Society in Belfast under the title “On the hypothesis that animals are automata, and its history”

One does not battle with drummers; but I venture to offer a few remarks for the calm consideration of thoughtful persons. It is quite true that, to the best of my judgement, the argumentation which applies to brutes [animals] holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain substance. [...] We are conscious automata, endowed with free will in the only intelligible sense of that much abused term — inasmuch as in many respects we are able to do as we like — but nonetheless parts of the great series of causes and effects which, in unbroken continuity composes that which is, and has been, and shall be — the sum of existence [from Huxley 1874, partially reprinted in Boakes 1984, p. 20]

## Acknowledgments

I would like to thank Francesco Mondada, Edo Franzi and André Guignard for designing and building the Khepera robot; I am especially grateful to Francesco Mondada for technical assistance and fruitful discussions on biologically inspired methods for engineering artificial systems.

## Notes

\* Ago, first person singular of the latin verb of Greek root, meaning “to lead, to play and active role, to behave”.

1. The software implementing the genetic operators and the neurocontroller (Floreano 1993) could be easily slimmed down and downloaded into the robot processor.

## References

- Atmar, W. 1993. Notes on the Simulation of Evolution. *IEEE Transactions on Neural Networks*, 5: 130–148.
- Baluja, S. 1996. Evolution of an Artificial Neural Network Based Autonomous Land Vehicle Controller. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 26: 450–463.
- Barto, A. G., R. S. Sutton, and C. J. C. H. Watkins. 1990. Learning and sequential decision making. In M. Gabriel and J. W. Moore, editors, *Learning and Computational Neuroscience*, pages 539–602. MIT Press-Bradford Books, Cambridge, MA.

- Beer, R. D. 1996. Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *From Animals to Animats IV: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 421–429. MIT Press-Bradford Books, Cambridge, MA.
- Boakes, R. 1984. *From Darwin to behaviourism*. Cambridge University Press, Cambridge.
- Cangelosi, A., D. Parisi, and S. Nolfi, 1994. Cell division and migration in a genotype for neural networks. *Network*, 5: 497–515.
- Cheneval, Y. 1995. Packlib, an interactive environment to develop modular software for data processing. In J. Mira and F. Sandoval, editors, *From Natural to Artificial Neural Computation, IWANN-95*, pages 673–682, Malaga, Springer Verlag.
- Clark, A. 1989. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. MIT Press, Cambridge, MA.
- Cliff, D., I. Harvey, and P. Husbands. 1993. Explorations in evolutionary robotics. *Adaptive Behavior*, 2: 73–110.
- Dennett, D. C. 1991. *Consciousness Explained*. Little, Brown and Company, USA.
- Dorigo, M. 1993. Editorial introduction to the special issue on learning autonomous robots. *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 26: 361–364.
- Dorigo, M. and U. Schnepf. 1993. Genetic-based machine learning and behavior based robotics: a new synthesis. *IEEE Transactions on Systems, Man and Cybernetics*, 23: 141–154.
- Edelman, G. M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books, New York.
- Edelman, G. M. 1992. *Bright Air, Brilliant Fire. On the Matter of the Mind*. Basic Books, New York.
- Floreano, D. 1993. Robogen: A software package for evolutionary control systems. Release 1.1. Technical report LabTeCo No. 93-01, Cognitive Technology Laboratory, AREA Science Park, Trieste, Italy.
- Floreano, D. 1996. Evolutionary re-adaptation of neurocontrollers in changing environments. In P. Sincak, editor, *Proceedings of the Conference Intelligent Technologies*, volume II, pages 9–20. Efa Press, Kosice, Slovakia.
- Floreano, D. and F. Mondada. 1994. Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural-Network Driven Robot. In D. Cliff, P. Husbands, J. Meyer, and S. W. Wilson, editors, *From Animals to Animats III: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 402–410. MIT Press-Bradford Books, Cambridge, MA.
- Floreano, D. and F. Mondada. 1996. Evolution of homing navigation in a real mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 26: 396–407.
- Floreano, D. and F. Mondada. 1996. Evolution of plastic neurocontrollers for situated agents. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *From Animals to Animats IV: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 402–410. MIT Press-Bradford Books, Cambridge, MA.
- Gaussier, P. 1995. Special Issue on Animat Approach to Control Autonomous Robots interacting with an unknown world. *Robotics and Autonomous Systems*, 16.

- Goldberg, D. E. 1992. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, Redwood City, CA.
- Harvey, I. 1992. Species Adaptation Genetic Algorithms: A basis for a continuing SAGA. In F. J. Varela and P. Bourguine, editors, *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pages 346–354. MIT Press-Bradford Books, Cambridge, MA.
- Harvey, I., P. Husbands, and D. Cliff. 1993. Issues in Evolutionary Robotics. In J. Meyer, H. L. Roitblat, and S. W. Wilson, editors, *From Animals to Animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pages 364–373. MIT Press-Bradford Books, Cambridge, MA.
- Hertz, J., A. Krogh, and R. G. Palmer. 1991. *Introduction to the theory of neural computation*. Addison-Wesley, Redwood City, CA.
- Huxley, T. H. 1874. On the hypothesis that animals are automata. *Nature*, 10:362, 1874.
- Kitano, H. 1990. Designing neural networks using genetic algorithms with graph generation system. *Complex Systems*, 4: 461–476.
- Koza, J. R. 1992. *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, Cambridge, MA.
- Langton, C. G. 1990. Artificial life. In C.G. Langton, editor, *Artificial Life*. Addison-Wesley: series of the Santa Fe Institute Studies in the Sciences of Complexities, Redwood City, CA.
- Lee, W-P., J. Hallam, and H. Lund. 1996. A hybrid GP/GA approach for co-evolving controllers and robot bodies to achieve fitness-specified tasks. In *Proceedings of IEEE 3rd International Conference on Evolutionary Computation*. IEEE Press.
- Mataric, M. and D. Cliff, 1996. Challenges in Evolving Controllers for Physical Robots. *Robotics and Autonomous Systems*, 19(1): 67–83.
- McFarland, D. J. 1992. Autonomy and self-sufficiency in robots. AI-Memo 92-03, Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium.
- McFarland, D. J. and T. Boesser, 1993. *Intelligent Behavior in Animals and Robots*. MIT Press/Bradford Books, Cambridge, MA.
- Miglino, O., H. H. Lund, and S. Nolfi. 1996. Evolving Mobile Robots in Simulated and Real Environments. *Artificial Life*, 2: 417–434.
- Mondada, F. and D. Floreano. 1995. Evolution of neural control structures: some experiments on mobile robots. *Robotics and Autonomous Systems*, 16: 183–195.
- Mondada, F., E. Franzini, and P. Ienne. 1993. Mobile robot miniaturization: A tool for investigation in control algorithms. In T. Yoshikawa and F. Miyazaki, editors, *Proceedings of the Third International Symposium on Experimental Robotics*, pages 501–513, Tokyo, Springer Verlag.
- Montana, D. and L. Davis. 1989. Training feed forward neural networks using genetic algorithms. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 529–538, San Mateo, CA, Morgan Kaufmann.
- Nolfi, S., D. Floreano, O. Miglino, and F. Mondada. How to evolve autonomous robots: Different approaches in evolutionary robotics. In R. Brooks and P. Maes, editors, *Proceedings of the Fourth Workshop on Artificial Life*, pages 190–197, Boston, MA, MIT Press.

- Parisi, D., F. Cecconi, and S. Nolfi. 1990. Econets: Neural networks that learn in an environment. *Network*, 1: 149–168.
- Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press, Oxford.
- Pfeifer, R. and P. F. M. J. Verschure. 1993. Designing efficiently navigating non-goal directed robots. In J. Meyer, H. L. Roitblat, and S. W. Wilson, editors, *From Animals to Animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. MIT Press-Bradford Books, Cambridge, MA.
- Pomerleau, D. A. 1993. *Neural Network Perception for Mobile Robot Guidance*. Kluwer Academic Publishing, Boston.
- Rudnick, M. 1990. A Bibliography of the Intersection of Genetic Search and Artificial Neural Networks. Technical Report CS/E 90-001, Department of Computer Science and Engineering, Oregon Graduate Center.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning Representations by Back-Propagation of Errors. *Nature*, 323: 533–536.
- Steels, L. 1993. Building agents out of autonomous behavior systems. In L. Steels and R. Brooks, editors, *The “artificial life” route to “artificial intelligence”*. *Building situated embodied agents*, pages 102–137. Lawrence Erlbaum, New Haven.
- Yamauchi, B. and R. D. Beer. 1995. Sequential behavior and learning in evolved dynamical neural networks. *Adaptive Behavior*, 2: 219–246.
- Yao, X. 1993. A review of evolutionary artificial neural networks. *International Journal of Intelligent Systems*, 4: 203–222.

# **Evolving robot consciousness**

## **The easy problems and the rest**

Inman Harvey  
University of Sussex

### **Introduction**

Car manufacturers need robots that reliably and mindlessly repeat sequences of actions in a well-organised environment. For many other purposes autonomous robots are needed that will behave appropriately in a disorganised environment, that will react adaptively when faced with circumstances that they have never faced before.

The design of autonomous robots has an intimate relationship with the study of autonomous animals and humans robots provide a convenient puppet show for illustrating current myths about cognition. Like it or not, any approach to the design of autonomous robots is underpinned by some philosophical position in the designer. Whereas a philosophical position normally has to survive in debate, in a project of building situated robots one's philosophical position affects design decisions and is then tested in the real world — “doing philosophy of mind with a screwdriver”.

In this chapter I shall first follow other authors in distinguishing various uses of the word ‘consciousness’. Using Chalmers’ characterisation of ‘the easy problems of consciousness’ (Chalmers 1995) I shall show how the evolutionary approach to robotics handles them. But then the main focus of the paper will be what Chalmers calls ‘the hard problem’ — which I will suggest is an easy non-problem.

### **The easy problems and the hard problem**

There is currently a fashion for asserting that ‘consciousness’ has become a respectable topic for scientists, as well as philosophers, to discuss. A number of

scientists, many of them eminent in their own fields within which there is a general consensus on the usage of technical terms, have blithely assumed that there is a similar consensus in discussion of consciousness. Oblivious of the multiple meanings available for the word 'consciousness', they frequently talk past each other and their audience.

Philosophers have for the most part been aware of this potential for confusion (though that does not mean that they have always avoided it). Chalmers (1995) warns of these multiple meanings, and proposes a basic distinction between those types of consciousness that offer (relatively) 'easy' problems for the scientist, and in contrast the 'hard one'. I will be largely agreeing with his analysis of the easy problems, and will tackle them first.

Chalmers defines the easy problems as "those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms". He chooses a different list from mine, but he covers the same ground. My broad categories I will characterise in terms of degrees of consciousness while driving my car home from work past a particular bend in the road:

1. consciousness\_1: though I have no recollection of the journey, I got home safely, so I cannot have fallen asleep or blacked out.
2. consciousness\_2: though I have no recollection of the roadside advertisement recently erected at the bend, later that evening I choose the new brand of beer that was advertised there, so it did indeed affect my later behaviour.
3. consciousness\_3: I notice the advertisement, and on arriving home I can recollect and describe it.

Here I have described these scenarios in the first person, but we can check whether somebody else is conscious\_1–3 by simple tests:

1. Did they react to the environment?
2. Was their later behaviour changed?
3. Can they report verbally on what they experienced?

All three of these classes of consciousness are 'easy' in Chalmers' sense: in principle we expect no mystery or magic in the underlying neural mechanisms in animals or humans. The complexity and the details may be difficult, and Chalmers suggests that it might take a century or two of work to uncover them, but conceptually we have no difficulties. The first two types of consciousness we attribute to non-human animals, and we can and do currently make robots

that exhibit them. Indeed we have even simpler mechanisms. When we test a newly installed burglar alarm for its sensitivity to an intruder's movement, we are testing for consciousness<sub>1</sub>. A soft-drinks vending machine that accepts a sequence of coins and alters its internal state in doing so passes the test for consciousness<sub>2</sub>.

Consciousness<sub>3</sub> in contrast seems currently constrained to humans, on any definition of linguistic competence that excludes a trained parrot or a telephone answering machine. Nevertheless, as long as one takes the criterion for such consciousness<sub>3</sub> in a third party to be the issuing of appropriate words in an appropriately wide range of contexts (in continued interaction with other language-users) then I agree with Chalmers. I go along with the basic credo of a cognitive scientist that ultimately we will be able to demonstrate underlying mechanisms that generate such behaviours.

Where I differ from most of my colleagues is in my expectation that we will never *comprehend* how such mechanisms operate *as a whole*, even when we can *create* or *display* them, and comprehend any small part of them. This limitation is not because of any deep mystery, but simply due to the limitations of us poor humans in understanding complex systems; below I will discuss how an evolutionary approach allows emulation without comprehension.

This list does not yet include what I will here call consciousness\* and Chalmers (1995) characterises as:

The really hard problem of consciousness is the problem of *experience*. When we think and perceive, there is a whirl of information-processing, but there is also a subjective aspect.

## Zombies, computation and dynamical systems

That philosophical favourite, the zombie, can pass all the tests for consciousness<sub>1–3</sub>, yet lacks 'phenomenal consciousness' or 'qualia', or 'conscious experience' — my consciousness\*. It can react to red traffic lights, even utter the words that describe yesterday's red sunset, yet it does not *experience* the sensation of red that I have, that I assume you have.

I cannot distinguish a zombie from you by its behaviour; this has the oft-forgotten corollary that I cannot distinguish you from a zombie by *your* behaviour. This does not stop me from in practice treating all humans who display signs of consciousness<sub>1–3</sub> as being more than zombies, as having consciousness\*. I also find myself doing likewise with a fair number of animals,

and even the occasional machine when I am being slapdash: “that printer always chooses the day of a deadline to break down, it seems to enjoy being awkward”. One goal of Artificial Intelligence (AI) is to produce machines that emulate human performance of all kinds; not just the intelligence of humans, but ultimately the consciousness also. This could be broken down naively into two tasks:

1. Produce a zombie machine with the right behaviours — the easy problems.
2. Add the extra ingredient that gives it consciousness\*, that makes it more than a zombie — the hard problem.

The first task is that of the roboticist. In the past AI practitioners might have said AI-oriented computer scientist rather than roboticist, given the prevailing fashion of the 1960s-80s of equating cognition with computation. Many such as Penrose (1989), whose knowledge of AI is secondhand, do not appreciate that for some time the new thrust in AI has been towards situated, embodied cognition, and a recognition that formal computation theory relates solely to the constraints and possibilities of machines (or people) carrying out algorithmic procedures. For those who do not accept the computational perspective on cognition, the worries advanced by Penrose are meaningless and irrelevant.

The astronomer, and her computer, perform computational algorithms in order to predict the next eclipse of the moon; the sun, moon and earth do not carry out such procedures as they drift through space. The cook follows the algorithm (recipe) for mixing a cake, but the ingredients do not follow an algorithm as they rise in the oven. Likewise if I was capable of writing a computer program that predicted the actions of a small creature, this does not mean that the creature itself, or its neurons or its brain, was consulting some equivalent program in ‘deciding what to do’.

Formal computations are to do with solving problems such as ‘when is the eclipse?’ But this is an astronomer’s problem, not a problem that the solar system faces and has to solve. Likewise, predicting the next movement of a creature is an animal behaviourist’s problem, not one that the creature faces. However, the rise of computer power in solving problems naturally, though regrettably, led AI to the view that cognition equalled the solving of problems, the calculation of appropriate outputs for a given set of inputs. The brain, on this view, was surely some kind of computer. What was the problem that the neural program had to solve? — the inputs must be sensory, but what were the outputs?

Whereas a roboticist would talk in terms of motor outputs, the more cerebral academics of the infant AI community tended to think of plans, or representations, as the proper outputs to study. They treated the brain as the manager who does not get his own hands dirty, but rather issues commands based on high-level analysis and calculated strategy. The manager sits in his command post receiving a multitude of possibly garbled messages from a myriad sensors and tries to work out what is going on. Proponents of this view tend not to admit explicitly, indeed they often deny vehemently that they think in terms of a homunculus in some inner chamber of the brain, but they have inherited a Cartesian split between mind and brain and in the final analysis they rely on such a metaphor.

An alternative view has gained favour in the last decade, though its origins date back at least to the early cybernetics movement. One version of this is the Dynamical Systems view of cognition:

... animals are endowed with nervous systems whose dynamics are such that, when coupled with the dynamics of their bodies and environments, these animals can engage in the patterns of behavior necessary for their survival. (Beer & Gallagher 1992: 91)

At this stage we downgrade the significance of *intelligence* for AI in favour of the concept of *adaptive behaviour*. Intelligence is now just one form of adaptive behaviour amongst many; the ability to reason logically about chess problems may be adaptive in particular refined circles, but the ability to cross the road safely is more widely adaptive. We should note the traditional priorities of AI: the computationalists' emphasis on reasoning led them to assume that everyday behaviour of sensorimotor coordination must be built on top of a reasoning system. Sensors and motors, in their view, are 'merely' tools for information-gathering and plan-execution on behalf of the central executive where the real work is done. Many proponents of an alternative view, including myself, would want to turn this on its head: logical reasoning is built on top of linguistic behaviour, which is built on prior sensorimotor abilities. These prior abilities are the fruit of billions of years of evolution, and language has only been around for the last few tens of thousands of years.

A dynamical system is formally any system with a finite number of state variables that can change over time; the rate of change of any one such variable depends on the current values of any or all of the variables in a regular fashion. These regularities are typically summed up in a set of differential equations. A Watt governor for a steam engine is a paradigmatic dynamical system (van

Gelder 1995), and we can treat the nervous system plus body of a creature (or robot) as one also. The behaviour of a dynamical system such as the governor depends also on the current value of its external inputs (from the steam engine) that enter the relevant differential equations as parameters. In a complementary way, the output of the governor acts as a parameter on the equations that describe the steam engine itself as a dynamical system. One thing that is very rapidly learnt from hands-on experience is that two such independent dynamical systems, when coupled together into (e.g.) steam-engine-plus-governor (treated now as a single dynamical system), often behave in a counterintuitive fashion not obviously related to the uncoupled behaviours.

Treating an agent — creature, human or robot — as a dynamical system coupled with its environment through sensors and motors, inputs and outputs, leads to a metaphor of agents being *perturbed* in their dynamics through this coupling, in contrast to the former picture of such agents *computing* appropriate outputs from their inputs. The view of cognition entailed by this attitude fits in with Varela's characterisation of cognition as *embodied action*.

By using the term *embodied* we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological and cultural context. By using the term *action* we mean to emphasize once again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. Indeed, the two are not merely contingently linked in individuals; they have also evolved together. (Varela et al., 1991: 172–173)

## Evolutionary robotics and behaviourism

Moving from natural agents to artificial robots, the design problem that a robot builder faces is now one of creating the internal dynamics of the robot, and the dynamics of its coupling, its sensorimotor interactions with its environment, such that the robot exhibits the desired behaviour in the right context. Designing such dynamical systems presents problems unfamiliar to those who are used to the computational approach to cognition.

A primary difference is that dynamics involves time, real time. Whereas a computation of an output from an input is the same computation whether it takes a second or a minute, the dynamics of a creature or robot has to be

matched in timescale to that of its environment. A second difference is that the traditional design heuristic of *divide and conquer* cannot be applied in the same way. It is not clear how the dynamics of a control system should be carved up into smaller tractable pieces; and the design of any one small component depends on an understanding of how it interacts in real time with the other components, such interaction possibly being mediated via the environment. This is true for behavioural decomposition of control systems (Brooks 1991) as well as functional decomposition. However, Brooks' subsumption architecture approach offers a different design heuristic: first build simple complete robots with behaviours simple enough to understand, and then incrementally add new behaviours of increasing complexity or variety, one at a time, that subsume the previous ones. Before the designer adds a new control system component in an attempt to generate a new behaviour, the robot is fully tested and debugged for its earlier behaviours; then the new component is added so as to keep to a comprehensible and tractable minimum its effects on earlier parts.

This approach is explicitly described as being inspired by natural evolution; but despite the design heuristics it seems that there is a practical limit to the complexity that a human designer can handle in this way. Natural Darwinian evolution has no such limits, hence the more recent moves towards the artificial evolution of robot control systems (Harvey et al. 1997) Evolutionary Robotics (ER).

In this work a genetic encoding is set up such that an artificial genotype, typically a string of 0s and 1s, specifies a control system for a robot. This is visualised and implemented as a dynamical system acting in real time; different genotypes will specify different control systems. A genotype may additionally specify characteristics of the robot 'body' and sensorimotor coupling with its environment. When we have settled on some particular encoding scheme, and we have some means of evaluating robots at the required task, we can apply artificial evolution to a population of genotypes over successive generations.

Typically the initial population consists of a number of randomly generated genotypes, corresponding to randomly designed control systems. These are instantiated in a real robot one at a time, and the robot behaviour that results when placed in a test environment is observed and evaluated. After the whole population has been scored, their scores can be compared; for an initial random population one can expect all the scores to be abysmal, but some (through chance) are less abysmal than others. A second generation can be derived from the first by preferentially selecting the genotypes of those with

higher scores, and generating offspring that inherit genetic material from their parents; recombination and mutation is used in producing the offspring population that replaces the parents. The cycle of instantiation, evaluation, selection and reproduction then continues repeatedly, each time from a new population that should have improved over the average performance of its ancestors. Whereas the introduction of new variety through mutation is blind and driven by chance, the operation of selection at each stage gives direction to this evolutionary process.

This evolutionary algorithm comes from the same family as Genetic Algorithms and Genetic Programming, which have been used with success on thousands of problems. The technique applied to robotics has been experimental and limited to date. It has been demonstrated successfully on simple navigation problems, recognition of targets, and the use of minimal vision or sonar sensing in uncertain real world environments (Harvey et al. 1997). One distinguishing feature of this approach using ‘blind’ evolution is that the resulting control system designs are largely opaque and incomprehensible to the human analyst. With some considerable effort simple control systems can be understood using the tools of dynamical systems theory (Husbands et al. 1995). However, it seems inevitable that, for the same reasons that it is difficult to design *complex* dynamical systems, it is also difficult to analyse them.

This is reflected in the methodology of ER that, once the framework has been established, concerns itself solely with the behaviour of robots: “if it walks like a duck and quacks like a duck, it is a duck”. For this reason we have sometimes been accused of being ‘the New Behaviourists’; but this emphasis on behaviour assumes that there are significant internal states — These are not significant in the sense of representational — internal states are mentioned here to differentiate evolved dynamical control systems (which typically have plenty of internal state) from those control systems restricted to feedforward input/output mappings (typical of ‘reactive robotics’) and in my view this emphasis on behaviour is compatible with the attribution of consciousness. A major conceptual advantage that ER has over classical AI approaches to robotics is that there is no longer a mystery about how one can ‘get a robot to have needs and wants’. In the classical version the insertion of a value function *robot\_avoid\_obstacle* often leaves people uncomfortable as to whether it is the robot or the programmer who has the desires. In contrast, generations of evolutionary selection that tends to eliminate robots that crash into the obstacle produces individual robots that do indeed avoid it; and here it seems much more natural to say that it is indeed the *robot* that has the desire.

## Back to consciousness

Natural evolution has produced creatures, including humans, through millennia of trials, selection, and heredity with variation. ER similarly requires a multitude of trials of real robots within the real world situations they are required to face up to. These trials are explicitly behavioural tests, and on the basis of these we have clearly already produced robots that exhibit consciousness\_1 and consciousness\_2. For linguistic abilities, consciousness\_3, I would agree with Chalmers that in principle this is a problem with no mystery, an 'easy problem', though I would expect more than his couple of centuries will be needed to crack it.

This reintroduction of the word 'consciousness' may have made the reader uneasy; but I am explicitly referring to the definitions of consciousness that can apply to a zombie or a machine. Now why do I attribute something extra to the humans I meet everyday, what is the magic ingredient consciousness\*?

If I cannot distinguish between you and a zombie by your behaviour, yet I treat you as something more, then the difference is in my attitude. Given a creature, a human or a robot in front of me, I can adopt a number of different stances (Dennett 1987). The mechanical stance is one I frequently adopt with machines, rarely with humans: with this perspective I treat the components as lifeless matter obeying physical laws, with not a trace of consciousness\*. From a different perspective, however, I normally treat other humans as aware, intentional creatures that are conscious\* like myself; I take this perspective occasionally also with machines, though the nature of machines that I come across generally means such a perspective is only short-lived, and soon reverts to a less personal one. As with the two perspectives of a Necker cube, it is impossible to hold both views simultaneously.

On a country drive when I notice some faltering in the power of my car's engine, my attention focuses on it as an object; in Heideggerian terms the car is no longer ready-to-hand, but becomes present-at-hand. If it has been temperamental recently, and I am used to its quirks, then I may well treat it not so much as an object but more like a person, and nurse it carefully with a soft touch on the accelerator. When it finally fails I open up the engine and take a mechanical stance, looking for broken wires or dripping fuel. Whichever of these three stances I take at any one time, like a view of a Necker cube it temporarily blots out any of the other possible perspectives.

## An attitude problem

So if one accepts that the creation of robots with consciousness<sub>1–3</sub> offers merely ‘easy’ problems (stretching the sense of ‘easy’ for consciousness<sub>3</sub>), the additional magic ingredient for consciousness\* is merely a change of attitude in us, the observers. Such a change of attitude cannot be achieved arbitrarily; the right conditions of complexity of behaviour, of similarity to humans, are required first. If an alternative perspective is much easier or more beguiling, then it will be difficult to shift away from. The more we can understand of a system from the mechanical perspective, the less likely we are to attribute agency, personhood, consciousness\* to it — which is why the ER approach that can produce comprehensible behaviour from incomprehensible mechanisms offers possibilities that the conventional design approach lacks.

Dennett in his response to Chalmers (Dennett 1996) makes a move that has some resemblance to the one I have made here. Where Chalmers (1995) suggests that the extra ingredient consciousness\* is a fundamental feature of the world, alongside mass, charge, and space-time, Dennett rightly pours scorn on this: consciousness\* is not some new entity over and above all these subsidiary phenomena of consciousness<sub>1–3</sub>. Dennett’s response will leave Chalmers and many others including myself unsatisfied, however — he seems to be explicitly denying the phenomena, our experience of visual sensations, the redness of the rose, the smell of coffee, and reducing everything to behaviour.

I sympathise with this dissatisfaction that Dennett does nothing to acknowledge or resolve. And I can put forward a perspective from which this problem is not actually *resolved*, but rather *dissolved* in Wittgensteinian fashion.

I take a Relativist perspective, that contrary to the naive popular view does not imply solipsism, or subjectivism, or an anything-goes attitude to science. The history of science shows a number of advances, now generally accepted, that stem from a relativist perspective that (surprisingly) is associated with an objective stance toward our role as observers. The Copernican revolution abandoned our privileged position at the centre of the universe, and took the imaginative leap of wondering how the solar system would look viewed from the Sun or another planet. Scientific objectivity requires theories to be general, to hold true independently of our particular idiosyncratic perspective, and the relativism of Copernicus extended the realm of the objective. Darwin placed humans amongst the other living creatures of the universe, to be treated on the same footing. With Special Relativity, Einstein carried the Copernican revolu-

tion further, by considering the viewpoints of observers travelling near to the speed of light, and insisting that scientific objectivity required that their perspectives were equally privileged to ours. Quantum physics again brings the observer explicitly into view. As for mathematics “I would even venture to say that the principle of mathematical induction is the relativity principle in number theory” (Von Foerster 1984). Cognitive science seems one of the last bastions to hold out against a Copernican, relativist revolution. Amongst the few to have been liberated were some of the early cyberneticists (Von Foerster 1984) and more recent philosophies that owe something to them (Maturana and Varela 1987).

Cognitive scientists must be careful above all not to confuse objects that are clear to them, that have an objective existence for them, with objects that have a meaningful existence for other agents. A roboticist learns very early on how difficult it is to make a robot recognise something that is crystal clear to us, such as an obstacle or a door. It makes sense for us to describe such an object as ‘existing for that robot’ if the physical, sensorimotor, coupling of the robot with that object results in robot behaviour that can be correlated with the presence of the object. By starting the previous sentence with “It makes sense for us to describe ...” I am acknowledging our own position here acting as scientists observing a world of cognitive agents such as robots or people; this objective stance means we place ourselves outside this world looking in as godlike creatures from outside. Our theories can be scientifically objective, which means that predictions should not be dependent on incidental factors such as the nationality or location or star-sign of the theorist; however we can only be objective about objects, not about our own subjectivity or consciousness\*.

Heinz Von Foerster explains why relativism does not lead to solipsism, and in doing so points to why we attribute to others the same consciousness\*, qualia, that we experience. Other humans exist for me, and I live in a society where other humans have comparable physical and social relationships — “if you prick us, do we not bleed? .. and if you wrong us, shall we not revenge?”. When I try to look at my own behaviour from an external, scientific position, I see remarkable similarities with other people’s behaviour, including their interactions with third parties. As a relativist I take the Copernican stance of refusing a privileged ‘objective’ position — yet clearly the solipsistic position is uniquely privileged. The absurdity of solipsism was brought out by Bertrand Russell’s solipsist correspondent, who thought it such a sensible attitude that she wondered why there were not more people who agreed with it!

If we reject solipsism, this entails that we attribute consciousness\* to others who behave in such a way that we take a personal, intentional stance towards them. This may still leave unresolved the worry of the person who asks: “but is the red that she experiences the same as the red that I experience, when we look at the same object?” Here I follow Wittgenstein in saying that this is a linguistic confusion, a mistaking of subjects for objects.

When I see a red sign, this red sign is an object that can be discussed scientifically. This is another way of saying that it exists for me, for you, and for other human observers of any nationality; though it does not exist for a bacterium or a mole. We construct these objects from our experience and through our acculturation as humans through education. It makes no sense to discuss (.. for *us humans* to discuss ..) the existence of objects in the absence of humans. And (in an attempt to forestall the predictable objections) this view does *not* imply that we can just posit the existence of any arbitrary thing as our whim takes us. Just as our capacity for language is phylogenetically built upon our sensorimotor capacities, so our objects, our scientific concepts, are built out of our experience. But our phenomenal experience itself cannot be an objective thing that can be discussed or compared with other things. It is primary, in the sense that it is only through having phenomenal experience that we can create things, objective things that are secondary.

One version of the conundrum that puzzles people goes as follows: We agree that 4 billion years ago there was a lifeless planet with no conscious\* beings, yet now there are; at some stage this magic ingredient consciousness\* appeared, as a product of lifeless matter — how can this be? But if we carefully and consistently make a note of the observers involved in this scenario, the problem dissolves. The ‘We’ that ‘agree’ refers to us from the scientifically literate community of the 21st century. For those of the mid-20th century the earth was only 2 billion years old, and who can now guess what current orthodoxy we will agree on in 50 years time? For us it is the case that 4 billion years ago there was lifeless rock, and there are now conscious\* beings; for the sake of argument we can posit a time *T* when, for us, the first conscious\* being appeared. The mystery arises only when we imagine ourselves being present at time *T-1* waiting for something — what? — to happen. But our assumption of no consciousness\* before time *T* makes our imagined scenario — us as conscious observers *then* — illegitimate.

## Summary

I have started off my argument by agreeing with Chalmers' distinction between the easy problems of consciousness and the rest — except where Chalmers sees the rest as a hard problem, I see it as a linguistic non-problem. The easy problems, in the context of robotics, are those of generating the desired behaviours of a zombie machine that emulates the behaviours we see in animals or other humans.

Evolutionary Robotics gives us a methodology explicitly based on such behavioural criteria. As practised at Sussex we adopt a Dynamical Systems approach to the 'stuff' from which robot control systems are built (Harvey et al., 1997). This means that behaviour is derived from the way an organism is coupled with its environment. However, theories based solely on behavioural criteria leave out what Chalmers calls the hard problem of consciousness.

Chalmers' attempt at a solution is to assert that consciousness\* must be a fundamental entity of the universe in the same way that mass or charge are. I agree with Dennett that there are no other entities over and above consciousness\_1–3, but this still leaves for most people a sense of dissatisfaction — isn't Dennett denying phenomenal experience, implying that it does not exist?

From a relativist phenomenological position I would assert that indeed we do (of course!) have phenomenal experience, but this is not a 'thing that exists' in the sense that matter and charge, indeed tables, and apples, exist. Phenomenal experience is primary, and through our experience we construct matter, charge, tables, apples as objects that allow us to make sense of the world. Though in many scientific disciplines one can get away without stating explicitly 'entity E exists from the perspective of those specific observers', as contrasted with 'entity E exists for us specific observers', the various Copernican revolutions have come about through rejecting (e.g.) absolute velocities in favour of relative velocities — 'velocities from the perspective of A or B'. If cognitive science follows the more basic sciences in accepting a relativist revolution, then the common philosophical puzzlements in relation to consciousness will just dissolve.

It follows that the creation of a robot that, for us, has the same forms of consciousness, even consciousness\*, as you or me, does not have any 'difficult' hurdles to cross, in Chalmers' sense of 'difficult' — to speak of a robot or person *having* consciousness\* is of course a potentially dangerous form of words, as it could be taken, misleadingly, to imply that consciousness\* is a 'thing'. The so-

called ‘easy’ hurdles, however, will no doubt need many centuries, indeed perhaps millennia, of hard work.

We will, in practice, only want to attribute consciousness\* to robots that we can see have their own concerns, and needs — so that objects exist-for-them. There are at least two possible reasons to suggest that an Evolutionary Robotics approach may be particularly appropriate to achieve this end. Firstly, it is easy to get needs and wants into such robots without explicitly programming them in, thus avoiding the GOFAI trap of the yawning chasm between an internal rule named by the programmer *robot\_avoid\_obstacle* and the robot actually *wanting* to avoid the obstacle. Secondly, the evolutionary approach will produce control systems that we cannot analyse — indeed, for me a major motivation for this method is that it allows us to produce systems more complex than our shallow understanding can cope with. It follows that a mechanistic understanding of such systems will not be available to us in practice, only in principle. Since that perspective on the Necker cube, that interpretation, is not available to us, we will in practice adopt the other natural interpretation at the behavioural level of description. It will be much easier for us to treat such robots as conscious\*.

## Acknowledgments

I thank the EPSRC and the University of Sussex for funding, and Shirley Kitts for philosophical orientation.

## References

- Beer, R. D. and Gallagher, J. C. 1992. Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1(1):91–122.
- Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence*, 47:139–159.
- Chalmers, D. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):2–19.
- Dennett, D. 1987. *The Intentional Stance*. MIT Press, Cambridge MA.
- Dennett, D. 1996. Facing backwards on the problem of consciousness. *Journal of Consciousness Studies*, 3(1):4–6.
- Foerster, H. V. 1984. *Observing Systems*. Intersystems Publications, Seaside, California 93955.
- Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. 1997. Evolutionary robotics: the Sussex approach. *Journal of Robotics and Autonomous Systems*, 20:205–224.

- Husbands, P., Harvey, I., and Cliff, D. 1995. Circle in the round: State space attractors for evolved sighted robots. *Journal of Robotics and Autonomous Systems*, 15:83–106.
- Maturana, H. and Varela, F. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala Press, Boston.
- Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.
- van Gelder, T. 1995. What might cognition be if not computation. *Journal of Philosophy* 92:345–381.
- Varela, F., Thompson, E., and Rosch, E. 1991. *The Embodied Mind*. MIT Press.



## I Epilogue



# The future with cloning

## On the possibility of serial immortality

Neil Tennant

The Ohio State University

### 1. Introduction

The cloning of human beings, which was once only a science-fictional possibility, now appears eminently feasible even if not imminently so — eminently, because being able to do it to sheep and monkeys really *does* mean being able to do it with human beings; but not imminently, because the ethical crisis over the prospect has already caused a storm in the public media, and is likely to result in strict legal prohibitions. And this would be a good thing. For I do believe that not even the science-fiction writers or the philosophers who conduct thought-experiments in this area have dealt with all the frightening scenarios that could arise from cloning, and the impact that these might have on our concepts of personal identity and our conceptions of lives worth living.

The philosopher, in his or her thought laboratory, conducts thought experiments — speculations about bizarre possibilities that stretch everyday concepts to the limit, with the intention of revealing what is essential, and what accidental, in the nature of things. I make no apology, therefore, for the wild character of the following speculations. They go way beyond even the ‘just so’ stories that have been told about our evolutionary *past*. Perhaps they should be described rather as ‘God forbid so’ stories, in so far as we care about, and might be able to influence, our evolutionary *future*.

My plan in this chapter is to look first at the possible consequences of having cloning as a live reproductive option. Then I shall describe a ‘cloning extension’ of a famous thought-experiment about ‘body-hopping’ due to Bernard Williams. This will put us in a position to contemplate even more frightening consequences than those afforded by cloning alone, or by body-hopping alone.

## 2. Cloning

The philosophical literature has dealt with ‘telecloning’ — the sort of thing that can happen if two copies of Scottie are beamed up. What I would like to do here, however, is consider the more straightforward, biological kind of cloning, such as that of Dolly the sheep. This is the kind that involves creating a new zygote (= fertilized egg) from the cellular material of an organism so as to create another (younger) organism — the clonal offspring — genetically identical to its clonal parent. We shall set aside as irrelevant any shortcomings of present cloning processes that might result in the clonal offspring being ever so slightly different, genetically, from the clonal parent. Slight differences might arise, for example, from the way parental mitochondrial DNA fails to be duplicated exactly in the new offspring. My own thought experiments will here be premised on the thoroughgoing genetic identity of clonal parent and clonal offspring.

## 3. Lives worth living

### 3.1. Straight sexual psyches at present

Evolutionary psychology is beginning to underscore just how much of our normal psychological baggage has been accumulated in the struggle for survival and reproduction.<sup>1</sup> Finding a mate at sexual maturity is the central project around which most of heterosexual human personality is buttressed. Our bodies and our characteristic male and female psychologies are (at least statistically) adapted to finding other bodies of the opposite sex with which to mix genes to create further offspring. Throughout our evolutionary past, there has been only one way to do this, namely, by having a sperm fertilize an egg and then having one or other parent (usually the egg-provider) harbour the fertilized egg until it has grown to the point where it can assume a spatially separated existence (albeit with further nurturing during a prolonged stage of dependency).

Heterosexual human beings are now ‘wired’ to behave in ways likely to conduce to conceptions of offspring — semi-copies of each parent, with half their genes coming from their father, and the other half from their mother. This basic genetic fact — that half of your offsprings’ genes have to come from someone else — is what so powerfully shapes each normal human being’s life-

projects. We cannot make high-fidelity copies of ourselves; we can only make semi-copies, genetically speaking. The evolutionarily stable strategies that our species has arrived at involve (according to the sociobiologists and evolutionary psychologists, as well as Grannie): suppressed oestrus; great male sexual jealousy and preference for nubile youth; female coyness, reticence and preference for men with status and control of resources; greater visual titillatability on the part of the male, and preference for sexual variety; elaborate courtship rituals; a largely monogamous mating system; characteristic rates of philandering; and intense emotional loyalties to children, especially on the part of their mother. Such is our conditioning by the constraint of semi-copying that we invest enormous time and psychic energy into evaluating potential mates, and fantasizing about how 'special' the resulting offspring might be, by virtue of the respective genetic contributions of their parents. We hope that the partners who attract us will 'breed true', and we discreetly check out their family backgrounds to try to determine whether they will.<sup>2</sup>

### 3.2. These immortal coils?

Imagine now what would happen if we no longer needed to find mates to make semi-copies of ourselves. Imagine, that is, that cloning technology were available, even if at a considerable price. This technology would offer the possibility of perfect genetic copying, without all the hassle involved — enjoyable though it may be — in finding a mate for semi-copying. There would be the other hassle of having to bear and raise the clonal offspring. That would set in train other, extremely interesting, selective forces. (In speaking of 'hassles' here I do not of course mean to imply that it is a fraught and unsatisfactory business. On the contrary; we are now wired to derive considerable satisfaction from these reproductive projects, and accordingly to be motivated to undertake them.) All it would take is some heritable variation in preferences for the old way versus the new way of creating offspring, and natural selection would go to work with a vengeance.

Think of a person as perfectly endowed as can be — a person with universally acknowledged good looks, a tremendous intellect, a wonderfully agreeable personality, a fine moral character, possessed moreover of a considerable fortune. Such persons are of course hard to find. But let us suppose that there were such, and that this person were female.

Why should this wonderful woman *risk* mixing her genes with those of a man, if she has access to cloning technology? She could 'conceive' and bear a

child all on her own and of her own. With the maternal resources to provide for her, this daughter would not need resources from a man.

What if our perfectly endowed person were male? Jane Austen, in *Pride and Prejudice* said that it is a truth universally acknowledged that such a man would be in need of a good wife. But *would* he, if the question to be answered concerned only his *inclusive genetic fitness* (not his personal happiness) in an environment where cloning technology is available?

Why should this wonderful man risk mixing his genes with those of a woman, if he has access to cloning technology? Or, rather, why shouldn't it be a much better strategy for him to go for a 'mixed' portfolio of genetic investment? Such a mixed strategy might involve considering bids from interested mates prepared to undertake the (bearing and) nurturing of his clonal offspring in return, say, for his occasionally allowing them to reproduce in the old-fashioned, gamete-uniting way. He just might have some takers with such a proposition.

Consider now the long-term impact on the human psyche (both male and female) if these new niches of reproductive opportunity were allowed to work long enough with their selective forces for the psychological dispositions that would make men (and women) exploit them to best advantage (from the genes' point of view). Who would win out eventually? — the straight copiers (the cloners) or the straight-sexers (the semi-copiers)? How would a distant descendant in such a regime of opportunities feel himself impelled to apportion his time and effort between, say, admiring and lusting after attractive women, and admiring and lusting after the most impressive cloning-incubators-cum-robotic-child-minders on the market? Remember that he might be around only because of the latter sort of lusting on some distant ancestor's part. All our speculation is premised on there being some heritable differences, however small, among human beings with regard to the proclivities I am imagining. It may be that they will arise only by random mutations; but it may also be that they are already within us, latent, waiting to be invoked and engaged by the advent of the new technology. Human beings are perverse and varied enough for this not to be beyond the bounds of possibility.

What would become of quiet parental pride in the achievements of one's offspring? Would clonal parents still be able to *admire* their clonal offspring for what they accomplished? Or would they feel that *they themselves*, in some more literal sense, deserved *all* the credit for their offsprings' accomplishments? Consider the mathematicians who proved the four-colour theorem, by (fa-

mously) deploying a computer to trudge through thousands of special cases. The special cases lent themselves to algorithmic decision. But this need not be the case in general. Suppose there were a mathematician with particular gifts and honed insights into a special corner of the mathematical universe, who knew that the only thing that stood between him and the Fields Medal was some very non-algorithmic and insightful ‘crunching of cases’, for which he himself had the native ability, but not the required *time*, before the age of forty. Suppose he accordingly cloned himself one hundred times and set the whole company of himself to work on the problem in parallel. Could he reasonably claim credit for the eventual theorem proved by that *ensemble* of ‘himself? My colleague Harvey Friedman, himself a very gifted mathematical logician and pianist, has mused out loud about the desirability of being able to make two copies of himself, and instruct one to work on certain foundational problems, and the other to concentrate on finding hitherto unrealized nuances in Chopin’s fourth Ballade.

It really is a mind-boggling spectrum of possibilities — one that threatens to completely unravel and re-stitch every thread in our gendered psychic make-ups. Sexual *frisson* and sexual yearning could become a thing of the past; or a highly seasonal thing; or an environmentally triggered thing, that could skip several generations, outside conditions permitting. Parental devotion to clonal offspring could become a frighteningly fanatical thing.

### 3.3. The war between the sexes

Since any clone is genetically identical to its parent, a more alarming prospect looms. There will be clones of males, and clones of females. And females will have the edge. For only a female can bear her own clone, whereas a male would still have to find a female to bear his. It wouldn’t be long before there were no longer any males around, given social and economic circumstances that rendered their testosterone-linked special traits less survival-enhancing. For the average woman prepared to bear a clone would be acting much more in her own genetic interests if she simply bore her *own* clone! The only inducement to bear a man’s clone might come in the initial stages of the new technology, while men still had a monopoly of the political power and economic resources that could coerce or entice women to do clonal grunt labour (surrogate motherhood) — perhaps, as suggested above, in return for some old-fashioned spawning from time to time, or even, if they could wangle it, for some opportu-

nities once in a while to *clone themselves* by means of their employer's machines. This in turn could set up a new selective force — for greater discernment, on the part of males, of those traits in females that are likely to be inherited by the females' own clones, and that would incline them to bear the men's clones for them in return for the odd sexual or clonal favours just mentioned.

### 3.4. An end to parent-offspring conflict and sibling rivalry?

Only one optimistic note can be sounded on this score: if the standard gene-selectional account of sibling rivalry and of parent-offspring conflict is correct, we could expect greater harmony within the (clonal) family. Ego's degree of genetic relationship with Parent would be 1, as it would also be with every Sib.

But then there would be another selective force to contend with — the one favouring viciousness towards recombinant siblings, and fierce loyalty to the fullest of full sibs, namely the clonal ones. Moreover, since females would have the innate upper hand in the cloning age (by not needing male gametes or male bodies for their own replication) it could turn out that the most ferocious competition would break out *among females* for access to, and control of, cloning technology. Perhaps it would be *females*, rather than males, who would eventually commit by far the greater proportion of homicide and infanticide!

Clonal technology would massively disrupt the present psychological equilibrium of our species. Initially it would be used within a gendered species still loaded with the dispositions to sexual behaviour inherited from the recombinant, dizygotic past. It is hard to say how many generations would be needed before the psychologies of men and women had been fundamentally shifted in adaptation to the new selective forces set up by the existence of cloning technology. But it is safe to say that our speculations as to what sorts of inclinations would assert themselves as selectively optimizing have only just begun. The family as we know it could disappear, and a newer, uglier war of the sexes could begin. People could even start paying for mutagens that could provide a modicum of genetic variability within clonal broods, at a carefully tailored level way below that of recombinant semi-copying.

I was cautious to anticipate greater harmony only within the *clonal* family. What of the ordinary family, though? — the kind with Mum and Dad and kids that mix the two? This social unit might become a thing of the past, or at

least of the poorest regions of the globe. Those who distrusted the lack of genetic variability induced by cloning, and who put their faith in hybrid vigour, might start a new religion: 'Blessed are the sleek and wild, for they shall inherit the earth.'

### 3.5. Altered appetites

Manufacturers of the new clonal technology might try to gain a competitive edge by exploiting ancient and still powerful aspects of sexual psychology before they became thoroughly outmoded — equipping their new products with pheromone emitters, perhaps, or even making them not shiny, but fleshily textured and shaped like human beings! There could be a fascinating interplay, down through generations of clones and their cloning technologies, between the appetitive structure of the clones' personalities and the perceptible features of their cloning technologies. Imagine the motivational shake-out that would be involved in switching from having to find and retain a mate to having to find and make most efficient use of a cloning machine! How might our current modes of sexual desire and receptivity be built upon, or rendered obsolete, by the new selection pressures? Would distant clonal descendants need to experience anything analogous to sexual desire in order to be motivated to clone themselves further?

Would there come a point at which manufacturers would be much better off just using their current product to clone away at themselves, rather than diverting their time and effort into the further R&D needed to make even better cloning aids for the wider market? Would they become even more industrially secretive, trying to reserve the ever-better prototypes for exclusive use by themselves and their own clonal offspring? Would this very incentive become so great that — Heaven forbid — it *destroyed the market as we might presently know it* for cloning technology? Could this possibly mean that one would not receive in the mail unsolicited catalogues from cloning technology companies?! When matters get so out of hand that even the *market* is disrupted, then we know we have a crisis of potentially unmanageable proportions. Disappearance of the family as we know it? — well, all right. Intensification of the war between the sexes? — well, we had suspected it was happening all along. But disruption of the *market*, for any one of its innovative products? — God forbid!

Indeed, why should we even think that the altered sensibilities of the future would include any analogue of sexual desire, involving arousal and intentional

focus on a body of the opposite sex? The relevant kinds of conscious experience (whether merely epiphenomenal, or genuinely causal), on the part of clones whose best genetic bet would be to carry on cloning themselves, might be radically different from those of people for whom sexual reproduction has been, and is, the only live option. New sorts of fantasizing would probably arise, involving intentional fixation on the technological methods and/or apparatus involved in cloning. Members of the opposite sex might become an *absolute irrelevance* — indeed, a competitive hindrance — to the business of replication. But, given the great likelihood that the technological apparatus involved in cloning would undergo great shifts in materials and design concepts, the new intentional focus for the clonally randy might have to be set at quite an abstract level. This in turn could set up selective forces for those aspects of a cognitive apparatus that could serve up the abstract foci and keep them suitably engaged with the appetitive drives involved in goal-directed planning. Just as there has been, according to sociobiology, natural selection for *self-deception*, so too now might there be natural selection for *self-seduction*, itself building on the earlier capacity for self-deception!

#### 4. Personal identity

##### 4.1. Body-swapping revisited

Bernard Williams, in his well-known paper ‘The Self and Its Future’, described a thought experiment in which two persons, A and B, are subjected to a ‘memory and personality’ switch. It was a cobbler-and-prince scenario, for a more egalitarian technological age. After the sophisticated procedure of brain scans and electronic infusions invoked in Williams’s make-believe, the A-body-person seems to have the memories and the personality of B, and vice versa.

A and B are told about the procedure they are about to undergo. Each is asked to choose which of the two post-operative persons — the A-body-person or the B-body-person — is to be tortured, and which is to be given a considerable financial reward. They are asked to make their individual choices in an entirely selfish way. Williams sets out to show how the pre-operative choices of A and B, and the post-operative behaviour of the A-body-person and the B-body-person, lead one to suppose that personal identity (so far as A and B are concerned) is bound up with the continuity of their respective memories

and personality rather than with the spatio-temporal continuity of their bodies. We should note, moreover, that in Williams's scenario, unlike Locke's of the cobbler and prince, there is good reason to believe that the continuity of A's (or B's) memories and personality is *causally* underwritten. The psychological entity or structure is continuously embodied in, and sustained by, physical things and processes: first a brain, then (via a recording device) a floppy disk, then (via an infusing and over-writing device) another brain. The post-operative A-body-person and B-body-person do not just wake up one morning, like the cobbler and the prince, to find themselves the subjects of a miraculous body-swap or mental migration. Rather, the tokenings of their psychological (=person?) types form continuous spatio-temporal worms that just happen to have funny segments within them — namely, their sojourns 'on' a floppy disk.

It was an interesting part of the thought experiment that Williams entered the proviso that A and B should be sufficiently similar in physical appearance for it to be possible for an acquaintance of A to 'see in' the post-operative B-body-person the personal characteristics of A. This would be highly unlikely if, say, A were a man and B were a woman; or if A were a blond Scandinavian giant and B were a short dark New Guinean. Williams asked his reader to acquiesce in the assumption of close enough physical resemblance to rule out such cases; and the reader could readily concur.

Williams did not think of improving his thought experiment at the time by postulating that A and B were identical twins who had been separated at birth and re-united only at the time of the experiment. That would have made it as easy as possible for an acquaintance of A to 'see in' the post-operative B-body-person the personal characteristics of A. But precisely that strength would also have entailed a corresponding weakness. For, in the case of identical twins — even ones who have lived apart all their lives, and accordingly have been exposed to very different influences moulding their characters, and have acquired very different memories — acquaintances of each could be inclined to confuse him with his twin, even without the memories-and-personality-swap procedure having been performed. Such confusion stems from the exact physical likeness, and (usually) exact similarity of gestures and facial expressions, which are exactly similar because they are so strongly genetically controlled. The conclusion that the B-body-person might be taken for A is therefore much less arresting, in the case of twins, than it would be in the case of non-twins who were nevertheless not so dissimilar in physical appearance as to make it impossible to secure that 'take'.

Both the twin-involving version, and the original non-twin-involving ver-

sion, of Williams's thought experiment share the feature that the persons A and B are contemporaries — of exactly, or roughly, the same age, respectively. The transfer of the person of A to the body of B (and vice versa) is an arresting possibility; but each could look forward to normal lives subsequently. And normally, lives end in death. This is the point at which my own thought experiment will make a difference.

#### 4.2. Rip, Rip\* and Rip\*\*

##### 4.2.1. *Rip*

Recall the Rip van Winkel story. Rip falls asleep under a tree, and stays asleep for 25 years. Of course, he ages while asleep. His features grow more wrinkled and he loses some hair. When he awakes his memories cover only that stretch of his life up to the point at which he fell asleep. Of the intervening 25 years he knows nothing — except that he was asleep during that time under that tree. He becomes a figure of fun among youngsters for being 'out of it' — for not having up-to-date memories of the recent past. He is also an object of concern for his friends, who have to help him adjust to all that has happened in the past 25 years. But he is still taken for Rip van Winkel; he is still the same person. He just has to learn to live with a rather large gap in his memories — a gap that he seeks to fill with as many first-hand accounts as he can from all his surviving friends. Let us call him the 'grown-old-Rip'.

##### 4.2.2. *The stayed-young-but-caught-up-Rip*

Consider now a variant of the Rip van Winkel story. Rip falls asleep under a tree, and stays asleep for 25 years. But while he sleeps he does not age. (We can suppose that he is cryonically suspended — that is, put into such a deep freeze that all cellular activities and processes, including those involved in aging, cease.) When Rip wakes up he is convinced that he has been asleep for only his usual short nap. It is disconcerting and disorientating for him to learn from those around him that he has been 'frozen in time', as it were, and that his memories are so far behind. He does not feel, physically, as though he could have missed out on 25 years of happenings around him; but in fact he has.<sup>3</sup> He has to learn to adapt to the strange and the new. He has to pick up from friends information about all that he had been missing. But they (the surviving ones) welcome him back as the Rip they know. It is, however, somewhat disconcerting for them that their old but young-looking friend has not aged along with them. They concur with his recollection of pre-sleep times. They can confirm his whereabouts and the

occurrence of various significant events in his earlier life. Gradually their stories help him ‘catch up’ and connect him again with the present.

Rip is the same person as before; he just happens to be in a rather strange predicament. His perseverance in his catch-up education, however, is astounding. He attends so avidly to his friends’ accounts of the past 25 years that in due course it’s as though he had lived through those years with them. By the apparent age of 50, 25 years after his thawing, he is like someone who has lived for 75 years. Of course, he has enjoyed only 50 years of living outside his cryonic suspension — 25 years before it, and 25 years after it. The remaining 25 years in the middle — the years of his cryonic suspension — he knows of only vicariously. Still, he has learned all sorts of ways in which he can cover up his lack of any direct experience of what transpired during those 25 years. At the chronological age of 75 but the cellular age of 50, he seems to have ‘packed in a lot of living’. Let us call him the ‘stayed-young-but-caught-up-Rip’.

#### 4.2.3. *Rip\**

I shall now tell you a story about the ‘for-ever-stay-young-and-never-say-die-Rip’, or *Rip\** for short. *Rip\** has a horror of growing old. He is about to turn 50. But he wants to feel once again as though he were 25 — full of vim and vigour and creative juices. He has no access to cryonics. Instead, he has access to cloning technology. *Rip\** creates a clonal offspring of himself. He is a very rich man, and able to induce a surrogate mother to bear his clone for him and bring it up. He even fathers it for a very short while.

#### 4.2.4. *Rip\*\**

*Rip\** calls this clonal offspring of his *Rip\*\**. He arranges for Bernard Williams’s memory-and-personality-recording device to take a read-out from his own (*Rip\**’s) brain, and store those memories and that personality on a floppy disk. He induces Bernard Williams to enter a contract whereby, at the age of 25, *Rip\*\** will be subjected to Williams’s memories-and-personality-infusion procedure using his (*Rip\**’s) stored memories and personality. Upon signing the contract, *Rip\** calls in Jack Kevorkian and gets done with his 50 year-old body. *Rip\*\** is only just out of diapers. The epitaph on his father’s tombstone reads ‘Gone, but not for good’.

On *Rip\*\**’s 25th birthday Bernard Williams duly carries out his contractual obligations. *Rip\**’s 50 years’ worth of memories and personality are infused into *Rip\*\**’s brain, thereby wiping out, or over-writing, all of *Rip\*\**’s own memories. *Rip\*\**’s personality traits, however, are much like those of *Rip\**,

since Rip\*\* is a very exact chip off the old block. He laughs in the same way, abhors Vegemite, is attracted to leggy blondes, loves reggae, and is a weekend cross-dresser, just like Rip\*. After the procedure, the Rip\*\*-body-person wakes up, his mind full of 50 years' worth of Rip\*'s memories, convinced that he is Rip\*. But Rip\*'s body died 25 years earlier; and of those intervening 25 years the Rip\*\*-body-person appears to have no memories. The continuous stretch of 50 years that he seems to remember began 75 years earlier. Still, he perseveres. Like the 'stayed-young-but-caught-up-Rip', the Rip\*\*-body-person assiduously absorbs first-hand accounts of those 'missing' and most recent 25 years, and in due course 'catches up'.

Now his 'father' (Rip\*) had a friend, Rip#, born on the same day, who did have access to cryonics.<sup>4</sup> At the age of 25, 25 years before Rip\* called in Kevorkian, Rip# had himself cryonically suspended for 50 years.<sup>5</sup> At the time of Rip\*\*'s procedure 50 years later, Rip# was awoken by thawing. He then attended the 'catch-up' sessions with the Rip\*\*-body-person. The 'stayed-young-but-caught-up-(on 50 years)-Rip#' mistakenly (?) came to believe of the 'stayed-young-but-caught-up-(on 25 years)-Rip\*\*-body-person' that he was none other than the 'stayed-young-but-caught-up-(on 50 years)-Rip\*'. To the 'stayed-young-but-caught-up-Rip#', his old friend Rip\* appeared to have had the gumption (and, as was well known by all his friends, he certainly had the ready cash) to be cryonically suspended for the same period of time that he (Rip#) himself had dared.

The question is: can we say that Rip\* discovered a recipe for serial immortality?

#### 4.3. Death, the giver of meaning

In another classic paper, 'The Makropoulos case: reflections on the tedium of immortality', Bernard Williams makes a telling case for the tedium and despair of the life of Elina Makropoulos, the Čapekian character who miraculously lives for 300 years at the apparent age of 42. Arresting the natural cadenza of a life, which normally has built into it a program of graceful waning, and keeping it on some perpetual plateau that for normal human beings would be a transitory stage of their existence, does not, contrary to initial and unreflective impressions, have much to be said for it. The satisfactions of a life well led appear to depend rather heavily on its being a life well tapered. Projects should be undertaken in the fullness of their time (for the person concerned), and one attains grace and happiness by making the adjustments called for by normal

aging, which brings with it various normal changes in bodily appetites, career ambitions, mental strivings and creative inspirations.

It is hard to fathom just how much of the normal Gaussian curve of maturation and decay, and the appropriateness of various undertakings within it, depend on our being dizygotic, sexually reproducing, mortal organisms with our particular evolutionary history.

## 5. Combining the two thought experiments

We have imaginatively sketched some disquieting consequences (given present evolutionarily conferred sensibilities) of having cloning as a live option. And we have looked at a cloning version of Williams's thought experiment about body-hopping. It is time now to combine the ingredients of these two thought experiments, to see what further frightening possibilities might suggest themselves.

First, parental selection of the best clonal offspring for eventual parental re-habitation, *à la Rip*\*\* above, could become the funeral and resurrection of choice.

But, if accumulated experience and wisdom really is survival-enhancing, one might expect there to be competition among clonal offspring for the privilege of being over-written by the paternal memories — unless their much more recently acquired knowledge is much more valuable than their father's. On the other hand, there might not be much *genetic* sense in vying with exact genetic copies of oneself for the privilege of letting the old block settle into one's own particular chip. Perhaps the whole clonal brood would rather form a coalition in order to preserve Dad in much more distributed form, by *each* being (willingly) over-written with his accumulated smarts! Imagine the social ceremonies that would attend such an evolutionary development.

We conjectured above that parental devotion to clonal offspring could become fanatical — indeed, in some sense literally self-possessive. But that would be the case only if human organisms continued to live normal lives that end in death and do not involve body-hopping.

With Bernard Williams's technology in hand, however, things might turn out drastically different. There might well be strong selection, in such situations, for *terrible parental neglect* of the usual emotional and cognitive nurturing of clonal offspring, if not neglect of their straightforward need for protection against physical harm. For, if one's own memories and personality

are one day going to be infused into these clonal receptacles, thereby overwriting everything in their own poor brains — including their memories of a wretched childhood, and the delinquent tendencies that result — would it *matter*, in the long term, to the newly selected parental sensibilities, whether one's clonal offspring were emotionally stable and well-nurtured in the interim? Perhaps they could be kept relatively feral and untended, until needed for the ritual of personality infusion. It would then be childhood, not life itself, that was nasty, brutish and short. Of course, given our present naturally-selected sensibilities, we experience outrage at the thought of treating youngsters so callously. But what we have to inquire after here is how those very sensibilities would be *re-shaped* by natural selection in the sort of environment, with its replication opportunities, that we are imagining.

Natural selection cannot be relied upon *a priori* to continue to sustain current ethical sensibilities. A future ethics will *a fortiori* be the ethics of those who won the struggle for survival and reproduction, in an era re-configured by cloning technology.

Such changed attitudes to child-rearing, although revolting to present sensibilities, might initially give cloning body-hoppers the edge, until their parental neglect reached the point where it was counter-productive. Their neglected, rebellious and unappreciated clonal offspring might start looking elsewhere than their immediate ancestor for memory-and-skill infusion, and be willing to consider bids from those who are relatively unsuccessful at cloning future receptacles for their own serial (and eventually parallel!) immortality. The latter might be willing to risk infusing their own memories and personality into a genetically unrelated receptacle. There would be Fagans in search of Artful Dodgers for back-street personality-and-memory infusions. Much might depend on the relative accessibility of Bernard Williams's technology as compared to that of the Hibernian cloning conglomerates.

## 6. Conclusion

Evolutionary psychology, reproductive technology, theoretical economics and good old philosophical thought experimentation have hardly begun to join forces on this front. The possibilities are fascinating, bizarre and frightful. They need to be thought through, even if only for their morbid or possibly anti-thanatological entertainment value.

## Notes

1. Good sources with which to start are D. Symons (1979) and M. Ridley (1993).
2. But we have to remember that sexual re-combination is not so straightforward. When some famous actress pestered Bernard Shaw to marry her, by asking him to imagine children possessed of her looks and his intellect, his wise riposte was that she should consider what they would be like with his looks and her intellect.
3. Interestingly enough, this Rip will be in the same predicament as the space-traveller who accelerates to a great velocity on a trip away from Earth, before eventually returning. According to relativity theory, he will come back to find that all his friends are much older than he appears to be. Now, if we assume that during his (short) flight he was actually *asleep*, then we have the following predicament: after what *really was* (for him) a short nap, he awakes (upon his return to Earth) to find all his friends apparently much older!
4. Rip\*\* and his 'father' Rip\* enjoy a degree of relatedness equal to one, as the sociobiologists would say (rather than one half).
5. The reader who thinks that no one in their right mind would ever wish to have this done to them should read Dennett's book *Darwin's Dangerous Idea*. There Dennett confesses that he might choose cryonic suspension if he could thereby be assured of the prospect of making the acquaintance of aliens known to be on their way to Earth, but not due to arrive for a good few more decades!

## References

- Dennett, D. 1995. *Darwin's Dangerous Idea*. New York: Simon & Schuster.
- Ridley, M. 1993. *The Red Queen*. Penguin Books.
- Symons, D. 1979. *The Evolution of Human Sexuality*. New York: Oxford University Press.
- Williams, B. 1970. "The self and its future". *Philosophical Review*, 79, 161-180.
- Williams, B. 1973. "The Makropoulos case: reflections on the tedium of immortality". In B. Williams, *Problems of the Self*, Cambridge: Cambridge University Press, 82-100.



# Subject index

## A

A-consciousness 111–112, 131, n. 2  
access consciousness 111  
accidental associations 146  
acquisition of representations 106–107  
adaptation xv, 185–187  
adaptation(ist) explanations 24–25, 30  
adaptation level 179  
adaptive  
    advantages 21, 27  
    behavior vs. intelligence 209  
    benefits of consciousness xiv, 182, 184–185  
    explanations 4  
    mechanisms xvii  
    value of consciousness xiv, 182, 184–185  
agents with internal dynamics 164  
agents  
    dynamical 164  
    pure sensory-motor 163  
    sensory-motor 163  
algorithms xvii  
    conscious 183  
    genetic 191, 212  
animals as automata 201  
artificial consciousness 200–201  
artificial evolution 211–212  
    of autonomous robots 187–188  
    of neurocontrollers 187  
artificial neural nets 134, n. 16  
    selection xiii, xviii  
automata 181  
automatic behavior 196  
autonomous  
    animals xvii  
    controllers 193  
    humans xvii

    robots xvii, 205  
    robots, artificial evolution of 187–188  
autonomy xvii, 185  
awareness xiii, 43, 181–183  
    with articulation xiii  
    first order 45  
    second order 45

## B

backward masking 116–117  
behavior  
    based robotics 165  
    automatic 195  
    home-seeking 196  
    not thinking, basis of conscious activity 200  
behavior-based robotics xvii  
behavioral plasticity xiii  
biological intelligence 200  
Blind Watchmaker 6, 8, 10, 15  
blindsight 45, 57–60  
bodies 144, 184  
“body-hopping” 223  
BRUTUS 123–126, 130, 135, n. 23

## C

Cartesian  
    dualism 156, n. 1  
    Matter 146  
    physics 156, n. 2  
causal grain xv  
    levels 84, n. 1  
    of phenomenal consciousness 64–67  
    project 65–67  
causal  
    mechanisms 3  
    role functions 34  
causal-role epiphenomenalism 34–35

- causality 139
  - causation 140, 157, n. 6
  - central state materialism 63
  - Church-Turing thesis 9
  - classic account of consciousness 89–91
    - interpreter argument against 101–106
    - naive argument against 91–101
    - representational argument against 101
  - cloning xviii, 223–224
    - serial 230–235
  - “cognition” xiv
  - cognition, myths about xvii
  - cognitive
    - functions 11
    - maps 168
    - versatility xiii
  - colonial families 228–229
  - communication xiii
  - computational equivalence 9
  - computer algorithms 181
  - concrete operational stage 165
  - connection weights 167
  - connectionism 108–109, 164
  - conscious algorithms 183
    - inessentialism 31–32, 38–39, n. 10
    - vs. unconscious behavior 3
  - consciousness xiii, xv, xvii, 156, 163–164
    - access 111
    - adaptive value of xiv
    - artificial 200–201
    - and creativity xvi, 111
    - and taxonomization 32
    - as a “spandrel” 5
    - as adaptive 21
    - as propositional 89–90
    - as up-dating one’s position in the world 183
    - causal grain xv
    - its causal role xv
    - its explainability xv
    - its ontological status xv
    - monitoring 111
    - organismic 4
    - phenomenal 111
    - self- 111
    - “survival value” of 3
    - symbolic processing model of 89–90
    - thermostatic 4
    - varieties of 22–24, 206–210, 213–218
  - consequences
    - differential 4
    - functional 4
  - constantly-changing environments 178
  - copiers vs. semi-copiers 226
  - “copying errors” xviii
  - creature consciousness xv
  - cyronic suspension 237, n. 5
- D**
- Darwinian-Indistinguishability xiv
  - decomposition, behavioral 211
    - functional 211
  - defined functions 109, n. 2
  - differential consequences 4
  - dissolving the hard problem 214–216
  - dreams as exaptations 25–26
  - dualism 5, 139, 144, 148, 154–156
  - dualistic theories xvi
  - dynamic agents 166
  - dynamical
    - agents 164
    - controllers 168
    - systems 209–210, 217
- E**
- E. coli* bacteria xix
  - ecological explanations 27
  - embodied
    - action 210
    - systems 164–166
  - embodiment xvii, 183–184, 210
  - emergence 142, 157, n. 5
  - emergent properties 156–157, n. 3
  - emotionless” AI 132, n. 7
  - empiricism 10, 147
  - environmentally-relative traits xviii
  - epigenesis 136, n. 26
  - epiphenomenalism xv, xvi, 21, 141–143

causal-role 34–35  
 etiological 34, 36–37  
 limited-exception metaphysical 39,  
   n. 14  
 strict metaphysical 35–36  
 epistemically live theoretical options 72,  
   83–84  
 etiological  
   accounts 37–38, n. 3  
   epiphenomenalism 26, 33–34, 36–37  
   functions 24–25  
 evolutionary robotics 166  
 evolution xiii  
   artificial 211–212  
   and creativity 121–131  
   as a branching tree structure xiv  
   as a set of causal mechanisms xiii  
   as a set of historical explanations  
     xiii–xiv  
   of phenomenal world xvii  
 evolutionary  
   optimization 188–189  
   psychology 223–224  
   robotics 187–188  
 evolvability  
   of A-consciousness 115  
   of M-consciousness 114  
   of P-consciousness 115–118, 121–  
     131  
   of S-consciousness 114–115  
 evolvable  
   language of thought interpreters  
     103–106  
   programs in the brain 98–100  
   software in the brain 100–101  
 evolved robots 172, 175–176  
 evolving  
   individuals 179  
   software 93–95  
   symbolic representations 107–109  
 exaptations 25–26  
 experience 146, 207–208  
 explanations 149–151

F  
 free will 5  
 feedback 59  
 “feeling” 143  
 feeling pain 4  
 feelings 157, n. 8  
 first order awareness 45  
 fitness  
   enhancing 21  
   of programs 95–96  
   robotic 191–200  
 formal operational stage 165  
 forward engineering 6  
 function of awareness 43–44  
 functional  
   consequences 4  
   explanations 5  
 functions  
   of metaconsciousness 46–48  
   defined 109, n. 2  
   primitive 109, n. 2

G  
 genetic  
   algorithms 191, 212  
   drift xiii  
   engineering xiii  
   fitness, inclusive 226  
   mutation xiii  
   programming 89  
 genetically-relative traits xviii  
 God 140, 152–153, 158, n. 12  
 grain project 65–67  
 grand unified theory of everything 8  
 group selection xiii

H  
 hard problem, dissolving the 214–216  
 health inessentialism 31–32  
 heritability 27–28  
 higher-order consciousness 89  
 home-seeking behavior 195–196  
 homunculi 187–188  
 human cognition, stages of 165

**I**

ideal adaptationist explanations 27–29  
immortality, serial 234–236  
incest 13  
inclusive  
    fitness calculations 13  
    genetic fitness 226  
increasingly complex levels of  
    situatedness 182  
increasingly-rich environments 178  
indistinguishability 10  
information 27–29  
innate mental language xvi  
Input/Output equivalence 9  
intelligence vs. adaptive behavior 209  
intelligence, biological 201  
“intentional stance” 15, n. 1  
internal dynamics xvii, 168, 173, 177  
    agents with 164  
internal representations xvii, 163, 166,  
    170, 173, 175, 177  
interpreter argument against classic  
    account of consciousness 101–106  
interpreters 91–93  
    as evolutionary products 102  
    needed 97  
introspection 89  
inverted  
    earth 49–50  
    heonic spectra 56–57  
    loudness 54  
    luminance 51–52  
    pitch 52–54  
    qualia 68  
    spectra xv, 48–49, 50–51

**K**

Khepera robots 166–167, 170, 173,  
    189–191, 196

**L**

language  
    as acquired xvi  
    as innate xvi

    of thought xvi, 89  
    of thought data structure 100  
    of thought interpreters 102–103  
lawlike generalizations 140  
learning 186–187  
    automata 186–187  
    autonomous agents 186–187  
    by imitation 166  
    mechanisms 186–187  
levels of causal grain 84, n. 1  
*life-world* 144  
lifetime learning 178–179  
limited-exception metaphysical epiphe-  
    nomenalism 39, n. 14  
“literalism” 150  
Lockean substances 157–158, n. 9  
    World 146  
logical (or conceptual) supervenience  
    68–70  
logical possibility, arguments for 120  
logically necessary traits 29–30

**M**

M-consciousness 111, 113  
machine code 91–92  
macromorphology 10  
materialism xvi, 139, 141  
materialism/physicalism 63  
mathematics 147–148, 155  
matter 148, 154–156  
    vs. mind xvi  
mental states 14  
mentalistic interpretations 4  
metaconsciousness xv, 44–48, 58–61  
    functions of 46–48  
metaphysically necessary traits 29–30  
micromorphology 10  
mind xix, 148–149, 154–156  
mind/body problem 5, 14–15  
minds 141, 144  
missing qualia xv  
mobile  
    robotics 164–165  
    robots 170

MONA-LISA 128–131, 136, n. 29  
 monitoring consciousness 111  
 moral realism 143  
 morphogenesis 136, n. 26  
 morphology 10

## N

naive argument against classic account of  
   consciousness 91–101  
 naturalism 63  
 natural selection xiii, xviii, 24, 142  
   in the broad sense xiii  
   in the narrow sense xiii  
 neural  
   nets 134, n. 16  
   networks 167, 170–171, 189–190  
 neurocontrollers, artificial evolution of  
   187  
 neuroscience 63  
 nomological (or natural) supervenience  
   68–70  
 nomologically necessary traits 29–30  
 non-generalizability 97  
 non-parameterized evolutionary  
   approach 126–130

## O

observations 147  
 Occam's Razor 16, n. 2  
 occasionalism 141  
 ontology of consciousness 63  
 organismic consciousness 4

## P

P-consciousness 111–112, 134, n. 17,  
   135, n. 20  
 pain as adaptive 21  
 panpsychic alternative 85, n. 6  
 panpsychism 139  
 parameterized logicist approach 123–126  
 parent-offspring conflict 228–229  
 perception 44  
 perceptual awareness 89  
 perfect endowment 225–226

personal identity 230–235  
 phenomenal consciousness 64, 111, 207–  
   208  
   “causal grain” of 64–67  
 phenomenal properties xiii  
 phenomenal states xv  
   as functional-representational 72–  
   78, 84, n. 3  
   as non-cognitively explainable 81–83  
   as ontologically *sui generis* 78–81  
 phenomenological  
   experience 217–218  
   unity of experience 21  
 philosophers' zombies 118–121  
 physical possibility, arguments for 120  
   states 14  
 planetary motion 157, n. 4  
 planets 152–153  
 Platonism 155–156, 158, n. 16  
 pleiotropic effects xviii  
 polygenic effects xviii  
 pre-operational stage 165  
 “primary consciousness” 183  
 primary qualities 145  
 primitive  
   command set 96  
   functions 109, n. 2  
 principles of evolution xiii  
 problem of  
   causal efficacy 70  
   explainability 70–71  
 propositional thought xv  
 protection from errors 97–98  
 proto-consciousness 181, 183, 200  
 “proto-phenomenal” properties 85, n. 5  
 proximal mechanism 13  
 psychophysical identity theory 85, n. 4  
 pure sensory-motor  
   agents 163  
   controllers 175  
   neural controllers 171  
   robots 167–168  
   systems 166  
 “puppet shows” xvii

**Q**

qualia xiii, xv–xvi, 43, 48, 143, 146, 158,  
n. 10, 207–208

relational account of 60

qualities

primary 145

secondary 145

**R**

random drift 24

rape 12–13

re-entrant connections 187–188

reduced-regularity environments 177–  
178

“reduction” 14

reductive explanations 69

reductivism 141–144

“reentrant” processing 59

reflection 89

reinforcement 186–187

relational account of qualia 60

“representation” 163

representation argument against classic  
account of consciousness 101, 106–109

reproduction

asexual 225–227

sexual 224–225

reverse engineering 6

robot consciousness xvii, 181

robotic

capacities 11

fitness 191–200

robotics

behavior-based xvii, 165

evolutionary 166

mobile 164–165

robots xvii

rotational errors 169–170, 171

“rule of thumb” xviii

**S**

S-consciousness 111–113, 132, n. 3

scientific underdetermination 8

second order awareness 45

secondary

properties 158, n. 10

qualities 145–146, 151

selected effect accounts of function 38, n.  
4

self-awareness xiii, 43

with articulation xiii

self-consciousness 111, 182–183

“self-hood” 158, n. 14

self-sufficiency 185

sensation 89

sensory patterns 175

sensory-motor

agents 163

coordination xvii, 177

stage 165

serial

cloning 230–235

immortality 234–236

sexual

reproduction xiii, xviii

selection xiii, xviii

sibling rivalry 228–229

situatedness, increasingly complex levels  
of 182

spandrels 25

stages of human cognition 165

stars 152

strict metaphysical epiphenomenalism  
35–36

Strong Equivalence 9

subjective

experience xiii

representation 43

superordinate traits 32, 39, n. 13

supervenience 63, 68–70, 78–81

“survival value” of consciousness 3

symbol grounding problem 9–10

symbolic processing model of conscious-  
ness 89–90

synaptic weights 198–199

synesthesia 54–56

## T

theoretical options, epistemically live 72,  
 83–84  
 theories of consciousness vs. theories of  
 mind xix  
 theories, underdetermination of 8  
 “theory of mind” 15, n. 1  
 thermostatic consciousness 4  
 “thinking” 143  
 “thinking things” 142–143, 149  
 “thisness” 64  
 total total Turing test (TTTT) 7–8  
 total Turing test (TTT) 7–8  
 traits  
     environmentally-relative xviii  
     genetically-relative xviii  
 transesthesia 54–56  
 TT (Turing’s test) xiv, 6  
 TTT (total Turing test) 7–8  
 TTTT (total total Turing test) 7–8  
 Turing equivalence 9  
 Turing indistinguishability xiv, 6–8, 15  
 Turing’s test (TT) xiv, 6–8, 15, n. 1

## U

*Umwelt* 144  
 unconscious 12  
 unconscious vs. conscious behavior 3  
 underdetermination  
     of theories 8  
     scientific 8  
 unity of consciousness 37, n. 2

## V

varieties of  
     consciousness xiii, 22–24, 43, 89, 111  
     epiphenomenalism 33–37  
 vegetative functions 11  
 Wernicke’s  
     aphasia 147  
     area 146  
 “world without qualities” 147

## Z

zombie possibilities xv  
 zombies xiv, 12–14, 49, 57–58, 67–68,  
 71, 76–77, 118–121, 132, n. 8, 133, n.  
 12, 207–210, 213, 217  
     philosophers’ 118–121



# Name index

## A

Alcock, J. E. 5, 16  
Amundson, R. 25, 34, 38–39  
Anscombe, G. E. M. 157–158  
Aristotle 139, 149, 154  
Armstrong, D. 45, 58, 61  
Asada, M. 179  
Ashcraft, M. 117, 137  
Atmar, W. 188, 201, 226  
Averback, E. 137

## B

Baker, R. 16  
Bakker, P. 166, 179  
Baluja, S. 188, 201  
Barfield, O. 151–154, 158  
Barkow, J. 12, 16  
Barto, A. 186, 202  
Bechtel, W. 67, 86  
Beer, R. 166, 179, 187, 202, 204, 209, 218  
Berkeley, G. 146, 159  
Bieri, P. 64, 85  
Blake, W. 147, 159  
Block, N. 44, 49, 61, 111–113, 119, 132, 134, 137  
Boakes, R. 201–202  
Boden, M. 137  
Boesser, T. 203  
Brandon, R. 24, 27, 29–30, 37–39  
Bringsjord, S. vii, ix, xvi, 111–112, 119, 122–124, 131, 133–135, 137  
Broca, P. 28  
Brooks, R. A. 165, 179, 218  
Byrne, R. 28, 39

## C

Cangelosi, A. 188, 202  
Carruthers, P. 45, 61  
Catania, A. C. 4, 16

Ceccnoni, F. 204

Chalmers, D. 39, 60–61, 66, 68–70, 78–82, 85, 119–120, 133, 137, 179, 205–207, 213–214, 217–218

Cheneval, Y. 190, 202

Cheng, K. 169, 179

Chesterton, G. K. 140, 142, 148, 159

Chopin 227

Church, A. 9, 16

Churchland, P. M. 22, 33, 35, 40

Churchland, P. S. 22, 40

Clark, A. 67, 85, 90, 104, 108–109, 202

Clark, S. R. L. vii, ix, xvi, 139, 149, 154, 157–159

Clarke, A. 17

Cliff, D. 166, 179, 187, 202–203, 218–219

Cole, D. vii, ix, xv, 43, 51, 61, 132, 137

Cooper, L. 122, 137

Copernicus, N. 214

Coriell, A. S. 137

Cosmides, L. 16

Crick, F. 22, 37, 40

Cummins, R. 34, 40

## D

Darwin, C. xiii, xvi, 14, 111

David, W. 137

Davidoff, J. 67, 86

Davies, P. 40, 143, 159

Davis, L. 203

Davis, W. 132

Dawkins, R. 4, 12, 16

Dekker, J. C. 126, 137

Democritus 158

Dennett, D. 5–6, 15–16, 30–33, 39–40, 106, 109, 119, 133, 137, 182, 202, 213–214, 217–218, 237

Descartes, R. 44, 139–140, 144, 148–149, 151, 156, 158–159

Donagan, A. 156, 159  
Dorigo, M. 186, 188, 202  
Dretske, F. 59, 61, 66, 77, 86, 119, 138  
Dunbar, R. 143, 159

## E

Earman, J. 133, 138  
Edelman, G. 22, 33, 40, 59, 61, 100, 106,  
109, 187, 202  
Einstein, A. 214  
Elman, J. 90, 110, 180  
Empiricus, S. 158

## F

Fagan 236  
Fehr, C. 37  
Fernald, R. D. 9, 16  
Ferrucci, D. 111, 122, 124, 135, 137  
Fetzer, J. H. ix, xix  
Fink, B. 38, 40  
Fjelde, R. 138  
Flanagan, O. vii, ix, xv, 21, 25, 29–32,  
38–41, 61, 67, 86, 119, 134, 138  
Flordano, S. 202  
Floreano, D. vii, ix, xvii, 180–181, 191,  
195–198, 201–204  
Fodor, J. 47, 90, 99, 110, 170, 179  
Foelber, R. 132, 137  
Foerster, H. 215, 218  
Fox Keller, E. 21, 40  
Franzi, E. 179, 201, 203  
Friedman, H. 227

## G

Galileo, G. 140, 153, 155, 159  
Gallagher, J. 209, 218  
Gallistel, C. R. 169, 171, 179  
Garon, J. 93, 110  
Garson, J. W. vii, ix, xvi, 89, 104, 110  
Gaussier, P. 186, 203  
Glanvill, J. 139  
Godfrey-Smith, P. 38, 40  
Goldberg, D. 188, 191, 203  
Gopnik, A. 16

Gould, S. J. 5, 16, 25, 40  
Graham, G. vii, x, xv, 37, 63, 84, 86  
Grantham, T. 37–38  
Grene, M. 156, 159  
Griffiths, P. 38, 40  
Guignard, A. 201  
Guzeldere, G. 44, 47, 61

## H

Hallam, J. H. 166–168, 179  
Hamilton, W. 44  
Hardcastle, V. 37  
Harnad, S. vii, x, xiv, 3–4, 6–7, 9–11, 14,  
16–18, 118–119, 138  
Hart, E. 124  
Harvey, I. viii, x, xvii, 179, 187–188, 202–  
203, 205, 211, 217–219  
Heemskerk, J. 165, 180  
Heracleitus 158  
Hertz, J. 186, 203  
Hinton, G. 204  
Hitler, Adolf xviii  
Hofstadter, D. 127–129, 135, 138  
Hooykaas, R. 156, 159  
Horgan, T. vii, x, xv, 37, 63, 68, 70, 84,  
86, 104, 110  
Hubel, D. 67, 86  
Husbands, P. 179, 202–203, 212, 219  
Huxley, T. 201, 203

## I

Ibsen, H. 135  
Ienne, P. 179, 203

## J

Jackson, F. 64, 78, 86  
Jakobi, N. 218  
James, W. 60–61, 145, 156, 159

## K

Kafka, F. 119–120  
Kant, I. 149  
Kenny, A. 157  
Kevorkian, J. 233–234

Kim, J. 68, 86  
 Kitano, H. 188, 203  
 Kitcher, P. 38, 40  
 Kitts, S. 218  
 Knuth, D. 127, 138  
 Koch, C. 22, 37, 40  
 Koenig, O. 47, 61  
 Kohler, I. 51  
 Kosslyn, S. 47, 61  
 Koyre, A. 159  
 Koza, J. R. 94–98, 107, 109–110, 203  
 Krogh, A. 203  
 Kugel, P. 126, 138  
 Kuniyoshi, Y. 166, 179

## L

Langton, C. 187, 203  
 Lauder, G. 25, 34, 38–39  
 Lee, W-P. 188–189, 203  
 Leibniz, G. 140, 159  
 Lewis, C. S. 157, 159  
 Lewis, D. 85–86  
 Lewontin, R. 25, 40  
 Libet, B. 5, 17, 157  
 Lincoln, A. 128–129  
 Livingstone, M. 67, 86  
 Lloyd, D. 89, 110  
 Lloyd, E. 21, 40  
 Locke, J. 146  
 Lofgren, I. 136, 138  
 Lund, H. H. 166–168, 179, 203  
 Lutz, R. 17  
 Lycan, W. 38, 40, 47, 62, 66, 77, 86

## M

Macdonald, C. 63, 86  
 Mackie, J. L. 143, 159  
 Makropoulos, E. 234  
 Margules, J. 169, 171, 179  
 Marr, J. 136, 138  
 Mataric, M. 187–188, 203  
 Maturana, H. 219  
 McFarland, D. 185–186, 203  
 McGinn, C. 60, 62, 81, 83, 86

McLaughlin, B. 79, 86  
 Metzinger, T. 61  
 Meyer, Y. 128, 138  
 Miglino, O. vii, x, xvii, 163, 166, 179–  
 180, 188, 203–204  
 Millikan, R. 37–38, 40–41  
 Mondada, F. 166, 179–180, 189–190, 201  
 Montana, D. 191  
 Mulhauser, G. v, 109

## N

Nafasi, K. 179  
 Nagel, T. 14, 17, 22, 41, 60, 62, 81, 86,  
 157, 159  
 Nahmias, E. 28, 41  
 Neader, K. 34, 37  
 Neander, K. 24, 41  
 Nijhout, H. 33, 41  
 Nilsson, N. 165, 179  
 Noel, R. vii, x, xvi, 111, 128, 136  
 Nolfi, S. vii, x, xvii, 163, 166, 172, 179–  
 180, 188, 202–204  
 Norvig, P. 115, 131, 138  
 Nozick, R. 46, 62

## O

Occam 8, 16

## P

Palmer, R. 203  
 Parisi, D. 180, 184, 202, 204  
 Penrose, R. 60, 62, 143, 148, 153, 155,  
 157, 159, 200, 204, 208, 219  
 Perry, J. 112, 138  
 Pfeifer, R. 165, 180, 204  
 Piaget, J. 164–165, 180  
 Pinker, S. 47, 62, 147, 159  
 Place, U. T. 63, 86  
 Plato xvi, 111, 155, 158  
 Plotinus 155, 159  
 Plum, F. 132, 138  
 Polger, T. vii, x, xv, 21, 29–30, 32, 39–41,  
 119, 134, 138  
 Pollock, J. 114, 132, 138

Pomerleau, D. 186, 204  
Posner, J. B. 132, 138  
Post, E. 126, 138  
Povienlli, D. 132, 138  
Premack, D. 15, 17  
Pylyshyn, Z. 9, 17, 47, 90, 99

## R

Ramsey, F. 152, 159  
Ramsey, W. 93, 110  
Rand, B. 159  
Ridley, M. 237  
Rogers, Professor 124  
Rosch, E. 219  
Rosenthal, D. 44, 47, 62, 113, 116, 119, 138  
Rudnick, M. 188, 204  
Rumelhard, D. 186, 204  
Russell, B. 215  
Russell, S. 115, 131, 138

## S

Scheier, C. 165, 180  
Schnepf, U. 202  
Schutz, A. 144, 159  
Searle, J. 118, 122, 133, 138  
Sharkey, N. E. 165, 180  
Shaw, B. 237  
Shepard, R. 122, 137  
Shields, W. M. 12, 17  
Shoemaker, S. 134, 138  
Smart, J. J. C. 63, 80–81, 86  
Sober, E. 38, 41  
Socrates 158  
Sorley, W. R. 44, 62  
Sperling, G. 117, 138  
Steels, L. 187, 204  
Steiner, R. 152, 159–160  
Steklis, H. D. 18  
Stells, L. 185  
Stich, S. 93, 110  
Stratton, G. 51  
Striver, D. 124

Sutton, R. 202  
Symons, D. 237

## T

Tani, J. 166, 180  
Taylor, C. 154, 160, 179  
Teer, Professor 124  
Tennant, N. viii, x, xviii, 223  
Thompson, A. 218  
Thompson, E. 219  
Thorndyke, P. W. 125, 138  
Thornhill, N. 12, 18  
Thornhill, R. 12, 18  
Tienson, J. 85, 104, 110  
Tolstoy, L. N. 132  
Tononi, G. 59, 61  
Tooby, J. 16  
Turing, A. xiv, 6, 9, 14, 18  
Tye, M. 66, 72–78, 81, 86

## V

van den Berghe, P. L. 13, 18  
van Gelder, T. 109–110, 210, 219  
van Winkel, R. 232–233  
Varela, F. 210, 215, 219  
Verschure, P. F. 204  
Von Foerster, H. 215, 218  
von Uexkuell, J. 144, 160  
Vrba, E. 25, 40

## W

Watkins, C. J. 202  
Weiskrantz, I. 62  
Wernicke, C. 28, 146–147  
Wigner, E. 143  
Wiley, B. 139, 160  
Williams, B. 230–237  
Williams, R. J. 204  
Wills, C. 28, 41  
Wittgenstein, L. xviii, 214, 216  
Woodruff, G. 15, 17  
Wright, L. 37–38, 41

**Y**

Yamauchi, B. 187, 204

Yao, X. 188, 204

**Z**

Zawidzki, T. 67, 86

Zenzen, M. 137

In the series ADVANCES IN CONSCIOUSNESS RESEARCH (AiCR) the following titles have been published thus far or are scheduled for publication:

1. GLOBUS, Gordon G.: *The Postmodern Brain*. 1995.
2. ELLIS, Ralph D.: *Questioning Consciousness. The interplay of imagery, cognition, and emotion in the human brain*. 1995.
3. JIBU, Mari and Kunio YASUE: *Quantum Brain Dynamics and Consciousness. An introduction*. 1995.
4. HARDCASTLE, Valerie Gray: *Locating Consciousness*. 1995.
5. STUBENBERG, Leopold: *Consciousness and Qualia*. 1998.
6. GENNARO, Rocco J.: *Consciousness and Self-Consciousness. A defense of the higher-order thought theory of consciousness*. 1996.
7. MAC CORMAC, Earl and Maxim I. STAMENOV (eds): *Fractals of Brain, Fractals of Mind. In search of a symmetry bond*. 1996.
8. GROSSENBACHER, Peter G. (ed.): *Finding Consciousness in the Brain. A neurocognitive approach*. 2001.
9. Ó NUALLÁIN, Seán, Paul MC KEVITT and Eoghan MAC AOGÁIN (eds): *Two Sciences of Mind. Readings in cognitive science and consciousness*. 1997.
10. NEWTON, Natika: *Foundations of Understanding*. 1996.
11. PYLKKÖ, Pauli: *The Aconceptual Mind. Heideggerian themes in holistic naturalism*. 1998.
12. STAMENOV, Maxim I. (ed.): *Language Structure, Discourse and the Access to Consciousness*. 1997.
13. VELMANS, Max (ed.): *Investigating Phenomenal Consciousness. Methodologies and Maps*. 2000.
14. SHEETS-JOHNSTONE, Maxine: *The Primacy of Movement*. 1999.
15. CHALLIS, Bradford H. and Boris M. VELICHKOVSKY (eds.): *Stratification in Cognition and Consciousness*. 1999.
16. ELLIS, Ralph D. and Natika NEWTON (eds.): *The Caldron of Consciousness. Motivation, affect and self-organization – An anthology*. 2000.
17. HUTTO, Daniel D.: *The Presence of Mind*. 1999.
18. PALMER, Gary B. and Debra J. OCCHI (eds.): *Languages of Sentiment. Cultural constructions of emotional substrates*. 1999.
19. DAUTENHAHN, Kerstin (ed.): *Human Cognition and Social Agent Technology*. 2000.
20. KUNZENDORF, Robert G. and Benjamin WALLACE (eds.): *Individual Differences in Conscious Experience*. 2000.
21. HUTTO, Daniel D.: *Beyond Physicalism*. 2000.
22. ROSSETTI, Yves and Antti REVONSUO (eds.): *Beyond Dissociation. Interaction between dissociated implicit and explicit processing*. 2000.
23. ZAHAVI, Dan (ed.): *Exploring the Self. Philosophical and psychopathological perspectives on self-experience*. 2000.
24. ROVEE-COLLIER, Carolyn, Harlene HAYNE and Michael COLOMBO: *The Development of Implicit and Explicit Memory*. 2000.
25. BACHMANN, Talis: *Microgenetic Approach to the Conscious Mind*. 2000.
26. Ó NUALLÁIN, Seán (ed.): *Spatial Cognition. Selected papers from Mind III, Annual Conference of the Cognitive Science Society of Ireland, 1998*. 2000.
27. McMILLAN, John and Grant R. GILLET: *Consciousness and Intentionality*. 2001.

28. ZACHAR, Peter: *Psychological Concepts and Biological Psychiatry. A philosophical analysis.* 2000.
29. VAN LOOCKE, Philip (ed.): *The Physical Nature of Consciousness.* 2001.
30. BROOK, Andrew and Richard C. DeVIDI (eds.): *Self-reference and Self-awareness.* 2001.
31. RAKOVER, Sam S. and Baruch CAHLON: *Face Recognition. Cognitive and computational processes.* 2001.
32. VITIELLO, Giuseppe: *My Double Unveiled. The dissipative quantum model of the brain.* 2001.
33. YASUE, Kunio, Mari JIBU and Tarcisio DELLA SENTA (eds.): *No Matter, Never Mind. Proceedings of Toward a Science of Consciousness: Fundamental Approaches, Tokyo, 1999.* 2002.
34. FETZER, James H.(ed.): *Consciousness Evolving.* 2002.
35. Mc KEVITT, Paul, Seán Ó NUALLÁIN and Conn Mulvihill (eds.): *Language, Vision, and Music. Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999.* n.y.p.
36. PERRY, Elaine, Heather ASHTON and Allan YOUNG (eds.): *Neurochemistry of Consciousness. Neurotransmitters in mind.* 2001.
37. PYLKKÄNEN, Paavo and Tere VADÉN (eds.): *Dimensions of Conscious Experience.* 2001.
38. SALZARULO, Piero and Gianluca FICCA (eds.): *Awakening and Sleep-Wake Cycle Across Development.* n.y.p.
39. BARTSCH, Renate: *Consciousness Emerging. The dynamics of perception, imagination, action, memory, thought, and language.* 2002.
40. MANDLER, George: *Consciousness Recovered. Psychological functions and origins of conscious thought.* n.y.p.
41. ALBERTAZZI, Liliana (ed.): *Unfolding Perceptual Continua.* n.y.p.
42. STAMENOV, Maxim I. and Vittorio GALLESE (eds.): *Mirror Neurons and the Evolution of Brain and Language.* n.y.p.
43. DEPRAZ, Natalie, Francisco VARELA and Pierre VERMERSCH.: *On Becoming Aware. On Becoming Aware.* n.y.p.
44. MOORE, Simon and Mike OAKSFORD (eds.): *Emotional Cognition. From brain to behaviour.* n.y.p.