Embodiment and the Inner Life

cognition and consciousness in the space of possible minds

Murray Shanahan

Embodiment and the Inner Life

This page intentionally left blank

Embodiment and the Inner Life

Cognition and Consciousness in the Space of Possible Minds

Murray Shanahan
Professor of Cognitive Robotics
Department of Computing
Imperial College London, UK

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford ox2 6dp
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2010

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Glyph International, Bangalore, India
Printed in Great Britain
on acid-free paper by
MPG Biddles Ltd, Bodmin, Cornwall and King's Lynn, Norfolk

ISBN 978–0–19–922655–9

1 3 5 7 9 10 8 6 4 2

‘We move from silence into silence, and there is a brief stir between, every person’s attempt to make a meaning of life and time’.
George Mackay Brown, *For the Islands I Sing*

This page intentionally left blank

Preface

This text is the outcome of over three decades of reflection on the mind and the brain. Hardly a day has gone by since my early teens when I haven't thought about what makes humans and animals tick, how our brains work, and how it might be possible to build artefacts that behave like us. Yet the disciplines that address these questions are surprisingly immature even today. So the present work is highly speculative, and the critical reader will find plenty to gripe at. So I must appeal for a certain degree of charity and open-mindedness. An overarching theory of cognition and the brain that can be shot down is better than no overarching theory at all—as long as it is well-informed, up-to-date, and carefully thought through. And of course, it is always possible that aspects of the theory are right.

I have debts of gratitude to many people. Academics from a variety of fields have been kind enough to respond to my unsolicited emails prodding them for details or clarifications of their work. To attempt to thank each one individually would be to risk embarrassing omissions. So I hereby offer my generic thanks. Certain colleagues I must single out, however. Bernie Baars has been unfailingly generous with his time, his insight, and his support. His influence permeates my thinking, as will become obvious. Igor Aleksander has been similarly magnanimous, offering me valuable advice and help during a difficult intellectual transition, as I set aside much of the work on which I had built my academic reputation while my amateur's interest in consciousness began to infuse (some might say infect) my professional life. My thanks also to many other people with whom I have had useful, and often inspiring, discussions on the subject matter of the book, particularly Ron Chrisley, Nicky Clayton, Owen Holland, Adrian Moore, Richard Newcombe, Anil Seth, and Aaron Sloman. I apologize to any friends I have forgotten. Finally, my thanks to Imperial College for giving me the freedom to pursue my somewhat unconventional interests.

To my wife and children I must apologize. During the years of my writing this book they have been forced to put up with an especially distracted husband/father, and all in all they have coped rather well. Still, I won't deny that without them I would have finished the book more quickly. But my life would have been so much impoverished. Lastly, on the advice of my sensible friends I will resist the temptation to thank Tooty, our cat. But his faithful presence

throughout this project, usually in my rocking chair, is hereby duly recorded. He is looking at me now as I write, disapprovingly. I know what that look means. It says: 'You must not allow anthropomorphism to cloud your scientific judgement'.

Murray Shanahan
North Norfolk and South Kensington
December 2009

Contents

Introduction

Why does death concern us? Other animals are naturally fearful in the face of danger. But humans have a special relationship with death. Irrespective of religious inclination, most of us are, at some time in our lives, exercised by the possibility of our own extinction, a possibility that other animals cannot fully entertain. Unlike the rest of the animal kingdom, we can imagine such an event. We can reflect on the possibility, and we can acquire beliefs about it. But what exactly is it that death threatens to extinguish? A common sort of answer would be that it threatens to extinguish consciousness. As in a dreamless sleep, in death (perhaps) I will experience nothing. Or rather, as death unlike sleep is (allegedly) a permanent condition, there will no longer be a consciously experiencing 'I'.

Here's another taxing question. Why does the joy or the suffering of our fellow creatures matter? If an inanimate object such as a brick or a television is abused in some way or other, we may feel distaste or displeasure, but we do not feel compassion. Yet when a fellow creature squeals in pain, we feel sorry for it. Or if we are indifferent, our indifference itself is noteworthy in a way that it is not in the context of a shattered brick. But what exactly distinguishes the brick from the animal here? A common sort of answer would be that an animal possesses consciousness whereas a brick does not. An animal's life is a succession of conscious experiences, and it is the character of those experiences—good or bad—that matters. The story of a brick, on the other hand, makes no mention of such things.

We need not endorse the precise wording of these answers—any choice of words would be open to philosophical critique—in order to accept their general tone. Consciousness is not a philosopher's invention. It is central to our humanity. It is what we care about most. The elemental facts of birth and death, of suffering and joy, serve to dramatize the concept of consciousness to the point of undeniability, if not to bring it within the sphere of definability. The aim of the present book—or rather the aim of the larger intellectual project of which the book forms a small part—is to account for this fundamental of the human predicament in scientific terms. That the Universe contains conscious creatures is a curiosity and a wonder, and the scientific understanding we seek must explain this wonder, must explain how it is possible. The challenge is not so much to explain how this came about, historically speaking, but rather to

explain what it is that allows conscious creatures to stand out, in the way they do, against the backdrop of the physical Universe.

Here is another curiosity, another wonder in need of a scientific account. Not only does our consciousness possess an ever-changing character, joyful or wretched in turns, it also has content. Our conscious thoughts are *about* things. The things they can be about include not only the furniture of the everyday world, but also abstractions such as numbers, electrons, musical genres, and political beliefs. Most remarkably, our conscious thoughts can be about our own consciousness, its character, its content, its temporal limits, and its place in the Universe. It is this capacity for reflection that so exquisitely and poignantly shapes the human predicament.

Now, it will be obvious to anyone with the slightest education in philosophy that the earlier paragraphs allude to two of its deepest and longest standing difficulties, namely the mind–body problem and the problem of intentionality. But the opening chapter of the present book belongs to the relatively youthful tradition of trying, not to solve such problems, but to dissolve them. According to this way of thinking, which was inaugurated by Wittgenstein, a fog of metaphysical confusion is liable to obscure a proper view of mind. We cannot hope to find the desired scientific explanations amid this fog, forever wandering from one ism to another (dualism, materialism, cognitivism, …). Yet to emerge from the fog is to return to a kind of prelapsarian state, a state before philosophy, which is also a state after philosophy. This post-reflective condition is hard to attain, and requires sustained, personal confrontation with philosophy itself, working through its founding concepts—language, meaning, truth, mind, experience, and so on.

Although the book is best seen as a whole, it's quite possible for one reader to accept the philosophy and reject the science, while another reader rejects the philosophy and accepts the science. (The possibility of rejecting both goes without saying.) Not everyone with a scientific interest in the mind is so exercised by philosophy as to become mired in metaphysics. Those who are not so mired—and indeed those who are but find a continentally inflected approach unsatisfactory or pretentious—may safely ignore the exclusively philosophical portion of the book (Chapter 1), and proceed directly to the material that takes for granted a license to investigate consciousness and cognition using the scientific method (Chapter 2 onwards). In an important sense, those who are untroubled by philosophy's largest questions, and who are able to do science without those questions sneaking up behind them, are privileged. They are in much the same position as the post-reflective philosopher. Both characters maintain a kind of silence in the presence of philosophy's demons, and each is duly empowered to make progress on the scientific front.

This is not to say progress can be made without an appropriate conceptual framework. According to the argument of the book, the right conceptual framework must be built around the fact of our embodiment (Chapter 2). Cognition has arisen because it beneficially modulates a creature's behaviour. That is to say, it helps the creature to survive, to thrive, and to procreate, and to do this it intervenes in the sensorimotor loop by means of which the creature interacts with its physical and social environments. Cognition's trick is to open out the space of what the environment affords the creature beyond what is immediately apparent. For a cognitively well-endowed creature, that space of possibilities—the space of potential affordances—is combinatorially structured, which is to say the possibilities for action can be assembled and re-assembled in arbitrarily many ways. Moreover, the space of potential affordances is open-ended. The enculturated creature, especially, can invent. The open-endedness of the space of possible affordances and its combinatorial structure commend its exploration, and this is cognition's forte.

Now, what of consciousness? Another feature of the right conceptual framework, according to this book, is a contrastive approach to consciousness. Rather than consciousness 'in itself' (whatever that might mean), the initial object of study should be the contrast between two conditions, the conscious and the unconscious, in circumstances as closely matched as possible. This conscious/unconscious distinction is (so very nearly) within empirical grasp. However, a number of strange temporal phenomena and near paradoxes (such as colour phi and the Sperling effect) make it hard to design experimental paradigms in which the distinction is incontrovertible. Chapter 3, in lieu of fully operationalizing the distinction, resorts to a science fiction scenario in which an idealized contrastive data set is constructed for a human subject in an ecological setting. The proposed experimental paradigm is not practically realizable, but it serves as a yardstick.

With an empirically meaningful target in view, albeit an idealized one, we can advance towards a 'theory of consciousness'. The primary influence on the ideas presented in Chapter 4 is Baars's notion of a global workspace architecture, which comprises a set of parallel specialist processes that compete and co-operate in a struggle for system-wide influence. According to the theory, the distinction between the unconscious and conscious conditions aligns with the difference between localized processing and processing mediated by the global workspace. In the conscious condition, so the argument goes, a kind of cognitive and behavioural integration is achieved, thanks to the global workspace, whereby the whole person (or animal), including the full resources of their brain, is brought to bear on the ongoing situation. A corollary of this is that the full space of possible affordances, open-ended and combinatorially structured, becomes amenable to exploration.

The burden of the work from there on is to give substance to this thesis in various ways. Details of the architecture need to be filled in, and relevant notions of process and computation have to be spelled out. The concept of 'integration' is clarified in terms of influence and information—two concepts that themselves stand in need of elucidation. But the real challenge is to map the architecture onto the brain. In the spirit of a commitment to understand the space of possible minds, not merely the human, or even the biological, this proceeds in two stages. The first, and most important, stage maps the abstract global workspace architecture into a more concrete, but biology-free, description in terms of connectivity and dynamics. The second stage is to map the connectivity and dynamics onto real biology. For didactic purposes, the two stages are sometimes interleaved, but they are conceptually distinct.

As far as connectivity is concerned, a network having small-world properties, hierarchically modular organization, and a prominent connective core is proposed as a plausible substrate for the requisite communications infrastructure. According to the proposal, the connective core is the counterpart to the global workspace. In the conscious condition, influence and information funnel into the connective core from the whole network, and fan out from it to the whole network. In a sense, the connective core is a bottleneck that enforces a form of serial processing. But the serial procession of broadcast states it produces is the result of sifting and blending the results of massively many parallel processes working behind the scenes.

To complement these ideas from the theory of networks, dynamical systems theorists have furnished the mathematicians, physicists, and computer scientists of the early 21st century with a splendid collection of conceptual exotica—metastability, chaos, chaotic itinerancy, self-organized criticality, complexity, and the balance of integration and segregation. Many researchers have articulated the intuition that one or more of these concepts is central to understanding cognition, consciousness, or the brain, providing more or less evidence or argument to back up their claims. But to date, no one has convincingly blended any subset of these ingredients into a satisfying whole, a framework with real explanatory potential. It would be rash to suggest the present material (Chapter 5) achieves this goal. However, taken together, Chapters 4 and 5 do offer a joined-up theory in which many of these elements have a vital role, affirming the intuitions of their respective advocates.

At the heart of the theory is an account of the way coalitions of coupled processes come together and break apart. Each coalition is an attractor of the system, and the system wanders from one attractor to the next, lingering awhile in each one. An important aspect of this account is its treatment of the contrast between repertoires of coalitions that are limited to the tried-and-tested and

repertoires of coalitions that are open-ended. The central claim is that the global workspace facilitates the requisite open-endedness, which in turn allows the system to realize the potential of an open-ended space of affordances by exposing it to whole new vistas of possibility. The global workspace does this by being at once the locus of broadcast, from where influence and information are disseminated throughout the system, and the medium by which arbitrary couplings of processes (alerted by broadcast to their potential relevance to the ongoing situation) can be established.

One major thesis of the book, then, is that the connectivity and dynamics of the global neuronal workspace underwrite cognitive prowess. But an extra element is required if a creature endowed with a global workspace is not to be confined to the here and now, if it is to have an inner life that stretches back into its past and reaches forward into its possible futures. The required extra element, according to the present theory, is an internally closed sensorimotor loop that can simulate interaction with the environment. An internal simulation, in the intended sense, is not some sort of movie that is played in the head. Rather, it involves mutually coupled activation within sensory and motor brain regions in the absence of external stimulation and without the production of external motor output. A further major thesis of the book is that the inner life of a human being arises from the combination of a global neuronal workspace with such an internal sensorimotor loop.

The closing chapter of the book is devoted to this idea and its ramifications. Thanks to its internal sensorimotor loop, the brain of a cognitively well-endowed animal is able to rehearse possible courses of action off-line, and to assess their likely impact from an affective point of view without actually carrying them out, in effect causing previously hidden portions of the space of affordances to be exposed. Moreover, thanks to its global workspace, the brain of such an animal is able, not only to explore the space of affordances circumscribed by tried-and-tested coalitions of sensory and motor processes, but also to find novel ways of blending together those processes, expanding its repertoire of coalitions and thereby opening up new vistas of affordance. Far from being the exclusive province of celebrated artists or inventors, the stance of this chapter, following the work of Fauconnier and Turner, is that a multitude of tiny creative leaps of this sort enables every child to acquire, blend upon blend and layer upon layer, the rich stock of concepts it needs to function in society.

The concluding sections of the book deal with three topics: working memory, episodic memory, and inner speech. It's noteworthy that each of these facets of mental activity is associated with consciousness. Not only do they modulate behaviour, they also contribute to the inner life of a human being. In the present framework, this is no surprise because each is explained in terms

of the exercise of an internal sensorimotor loop whose operation is mediated by a global workspace. Three specific hypotheses are explored—first that working memory locates the subject in space, second that episodic memory situates the subject in time, and third that inner speech places the subject in a reflective relation to itself. Despite appearances, this is not a philosophical investigation into the nature of the subject. These are empirical claims, grounded in the neurodynamics of integration and internal sensorimotor activity. In the final paragraphs, the book returns to its point of departure with the observation that self-referential inner speech (thinking about thoughts) foreshadows philosophy.

Chapter 1

The post-reflective inner view

In this chapter we open hostilities with those traditionally philosophical ways of thinking about ourselves that lead to dualism—a metaphysical division between the public, external world and our private inner lives—and consequently to the sense that we are forbidden from truly investigating ourselves using the scientific method. Recruiting Wittgenstein as our ally in a confrontation with the most treasured categories of traditional philosophy—truth, meaning, experience, existence, identity, and so on—we progress towards a kind of silence on metaphysical matters. Along the way, a number of well-known arguments and thought experiments in the philosophy of consciousness are revisited.

1.1 The supposed dualism of inner and outer

In the *Meditations*, having cast into doubt the existence of sky and earth, of bodies and other minds, Descartes finally asks whether he might even be deceived about his own existence, and concludes that this, at least, is impossible since 'the proposition: *I am, I exist*, is necessarily true every time I express it or conceive of it in my mind'.[1] Pursuing the ramifications of this simple but incendiary argument—the *cogito*, as it is traditionally called—Descartes finds that he has, metaphysically speaking, severed the connection between his mind and his body and has lost touch with the world. In his earlier *Discourse on Method*, the whole troubling series of thoughts is compressed into a single short passage.

> Then, examining attentively what I was, and seeing that I could pretend that I had no body and that there was no world or place that I was in, but that I could not, for all that, pretend that I did not exist … I thereby concluded that I was a substance, of which the whole essence or nature consists in thinking, and which, in order to exist, needs no place and depends on no material thing …[2]

[1] Descartes (1641/1968), p. 103.

[2] Descartes (1637/1968), p. 54.

It has become customary, in contemporary cognitive science, to blame Descartes for almost everything. The celebrated passage above inaugurated 'Cartesian dualism', a by-word for a position now widely considered so silly that reduction to it is tantamount to refutation. The 'Cartesian theatre' is Dennett's label for a supposedly naive view of consciousness that he is concerned to debunk.[3] According to Damasio, 'Descartes' error' (one among so many) was to divorce human reason from the animal passions.[4] Wheeler attributes the failings of 'orthodox' cognitive science to its commitment to 'Cartesian psychology'.[5]

Broadly speaking, the stance we shall adopt here is sympathetic to the specific 'anti-Cartesian' views just listed. However, the radical challenge of the *cogito* is not at all easy to dismiss. The argument encapsulated in the quoted passage is of the utmost depth and difficulty, and the image of mind it impresses on us is impossible to ignore. So it's hardly surprising that we find powerful echoes of it in the contemporary philosophical debate about the nature of consciousness, specifically in arguments that seek to limit the scope of scientific enquiry by establishing an inviolable metaphysical barrier between subject and object, between inner and outer.

For example, a survey of contemporary arguments purporting to establish various types of fundamental inner/outer divide is at the heart of Chalmers' influential book *The Conscious Mind*, leading him to distinguish what he calls the *easy* problem of consciousness from the *hard* problem.

> [A lot of writing] addresses the 'easy' problem of consciousness: How does the brain process environmental stimulation? How does it integrate information? How do we produce reports on internal states? These are important questions, but to answer them is not to solve the hard problem: Why is all this processing accompanied by an experienced inner life?[6]

Though couched in contemporary terms, Chalmers' distinction bears the same hallmark as Descartes' reflection. In both cases, a wedge is driven between inner and outer. For Descartes, body and place (outer) are divided from thought (inner), whereas for Chalmers, the information processing taking place in the brain (outer) is divided from phenomenal experience (inner). Of course, there is a sense in which information processing occurring in the brain is 'inner' relative to the goings on in the 'outer' environment. But this is not the

[3] Dennett (1991).

[4] Damasio (1995).

[5] Wheeler (2005).

[6] Chalmers (1996), pp. xi–xii.

sense of 'inner' at stake here. The *inner life* of a human being comprises thoughts, feelings, perceptions, sensations, and so on.

The mark of the inner is that it is supposedly *private*. A person's inner life is his or her own, and—we are inclined to say—no one has closer acquaintance with it than the person in question. By contrast, the outer is *public*. It is everybody's, and we all have access to it, at least in principle. If I announce that I am feeling happy today, it's not for you or anyone else to argue with me, whereas if I claim the Moon is made of green cheese then you have the right to challenge my claim with an empirical investigation. My happiness is in here, whereas the Moon is out there. We might say that the former is a *subjective* matter, nobody's business but my own, whereas the *objective* nature of the latter can be studied by anyone with the appropriate means.

In everyday speech we deploy commonplace distinctions between inner and outer, between private and public, or between subjective and objective, without taking a metaphysical stand. But the wedge Descartes and his present-day counterparts attempt to drive between inner and outer is resoundingly metaphysical. It divides inner from outer so completely that we are left with a profound mystery. What is the relationship between inner and outer? How do they interact? How do events in the inner realm influence physical activity in the brain and affect overt behaviour? Conversely, how do events in the outer world penetrate the inner, and give rise to private, subjective experience? Assuming such interactions occur, as surely they must, how can we accommodate them within a scientific understanding of Nature and the physical laws that are presumed to govern it? However well a scientific theory explains the relationships among the merely outer phenomena of human behaviour, human psychology, and human neurophysiology, it seems the inner, private realm of subjective experience must remain forever beyond its grasp.

Hence we arrive at the sort of impasse characterized by Chalmers as the hard problem. It's perfectly natural to arrive at this position, and there is no straightforward route around it. Simply to deny metaphysical weight to the inner/outer divide without properly engaging with it is to bury one's head in the intellectual sand. Nor is it sufficient simply to launch surface attacks on the arguments that purport to establish the division, the *cogito* being among the most profoundly rooted. Rather, our confrontation with inner/outer dualism must take place at a much deeper level, at the level of the very foundations of philosophy itself. To put it bluntly, we must learn to do without the habit of metaphysical thinking.

What is meant here by 'metaphysical thinking'? Metaphysics, as we shall construe it, is the study of fundamental abstractions, such as mind, truth, reality, meaning, thought, and consciousness. Its ambition is to reveal the *essence* of these things, where a thing's essence is that without which it would not be

what it is, that which distinguishes it from everything else. A sure indicator of metaphysical thinking is the conviction that fundamental abstractions are, in some sense, a matter of objective necessity, so metaphysics can proceed by reasoned argument alone. Thought is computation. Truth is correspondence. Meaning is use. Consciousness is a physical process. These and all similar claims, wherein an innocent little copula is invested with overwhelming significance, betray metaphysical thinking. Moreover, to deny claims of this sort is equally a sign of metaphysical thinking. Likewise, materialism, functionalism, behaviourism, physicalism, dualism, idealism, Platonism, antirealism, verificationism—all these are metaphysical doctrines, as are their negations. All are symptoms of metaphysical thinking, and all must be discarded.

1.2 Introducing the private language remarks

The proposed assault is clearly audacious, so it's fortunate that we have a powerful ally, namely Wittgenstein. The relevant portion of Wittgenstein's oeuvre is not the early philosophy that found concise and careful expression in the *Tractatus Logico-Philosophicus*,[7] but rather the style of critical thinking characteristic of his later period and exemplified by the posthumously published *Philosophical Investigations*.[8] Our initial focus will be a much-discussed series of remarks at the heart of the *Investigations* concerning sensations, how we talk about them, and the extent to which it makes sense to speak of them as private.[9] Taken together, these remarks are usually referred to as 'the private language argument'. However, as we shall see, despite appearances, they don't really constitute an argument in the conventional philosophical sense. Rather, their purpose is to help break the habit of dualistic thinking at a level beneath that of traditional philosophical discourse. If this is successful, the intuitions that motivated the distinction between the easy problem and the hard problem of consciousness will lose their hold over us.

Bearing this agenda in mind, let's now take a look at the private language remarks. Here are the passages in which Wittgenstein introduces the notion of a private language. First, he invites us to consider some related ideas.

> A human being can encourage himself, give himself orders, obey, blame, and punish himself; he can ask himself a question and answer it. We could even imagine human beings who spoke only in monologue; who accompanied their activities by talking to themselves.[10]

[7] Wittgenstein (1921/1961).

[8] Wittgenstein (1958).

[9] Wittgenstein (1958), §243, §§256–258.

[10] Wittgenstein (1958), §243.

The idea of such a monologue is unproblematic. This is *not* the sort of private language under consideration. Indeed, much of the *Philosophical Investigations* is written as a sort of inner monologue, in which Wittgenstein—often using the voice of an imaginary interlocutor—asks questions of himself and then answers them. No, the sort of private language at issue here is something quite different.

> But could we also imagine a language in which a person could write down or give vocal expression to his inner experiences – his feelings, moods, and the rest – for his private use? – Well, can't we do so in our ordinary language? – But that is not what I mean. The individual words of this language are to refer to what can only be known to the person speaking; to his immediate private sensations. So another person cannot understand the language.[11]

Suppose a subject inspects a set of colour samples and claims to enjoy two different colour sensations. Even though they all look identical to the rest of humanity, the subject insists that, for her, some patches evoke one colour sensation whereas the others elicit a quite different experience. Suppose the subject elects to label her two sensations of colour 'S1' and 'S2'. Now imagine that a scientific investigation establishes a real difference in the spectra reflected by the various patches, a difference that correlates perfectly with the subject's claims. In this case, an objective distinction between the two labels has been found, and this means 'S1' and 'S2' are not part of a private language. The subject turns out to have been labelling something public all along. Suppose instead that no scientific test can distinguish the colours themselves, but that some neurological difference of state can be found, through fMRI or some future brain scanning technique, that correlates perfectly with the subject's use of the two labels. Once again, the potentially private language has ended up in the public sphere. If 'S1' and 'S2' were words in a genuinely private language then no scientific investigation would, in principle, be capable of supplying an objective test that correlated with the subject's labelling. This is the notion of privacy at issue.

Having circumscribed the idea of such a language, Wittgenstein muses on a hypothetical process by which its vocabulary might be developed. He asks us to imagine a diary in which he records all the days on which he has a certain sensation. To do this he invents a symbol, say 'S'. In his diary he marks each day on which he has the sensation in question with this symbol. Now, Wittgenstein asks, how would he know whether he was marking the diary correctly? How could he ever be sure that he was using the symbol 'S' for the same sensation on different occasions? If the sensation is truly 'private', what criterion for *correctness* could there possibly be? Marking the diary, Wittgenstein contends,

[11] Wittgenstein (1958), §243.

is nothing but an empty ceremony, like the left hand giving the right hand money: 'One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about right'.[12]

One way to arrive at a proper understanding of these remarks, and at an appreciation of their full force, is to begin with a provisional exegesis—couched in somewhat traditional philosophical terms—which can then be revised into a more radical interpretation. Provisionally then, we can characterize these remarks as an attack on the possibility of establishing identity criteria for private inner sensations. This, of course, puts enormous pressure on the very concept of a private inner sensation. Pressure is applied by casting doubt on the notion that the same word could reliably be ascribed to the same sensation twice. The very definition of a private inner sensation is intended to entail that there can be no external, objective arbiter of reliability here. If no coherent identity criteria can be found for private inner sensations, then their ontological status becomes very precarious, and arguments that are based on them are duly discredited. So what we apparently have is a conventional, though powerful, argument for a conventional, though damaging, metaphysical proposition. The proposition states that there is no such thing as a private sensation. The argument hinges on identity criteria, but appeals to language on the assumption that anything that exists can be named (a relic, perhaps, of the verification principle discerned by many philosophers in Wittgenstein's earlier work in the *Tractatus*).

On this provisional reading, there is plenty of scope for mounting a defence of the concept of private inner sensation. We might call into question, for example, the assumption that a public criterion of success is necessary for a word to latch onto its referent. We might even be so bold as to appeal to an external arbiter that transcends the public/private distinction in play—the mind of God, perhaps, or some higher metaphysical reality. We might challenge the role of language in the argument. But even if we dismiss all such counter-attacks, even if we accept the conventional conclusion of the private language remarks on this provisional reading, we end up not with a resolution of the philosophical problem, but instead with a profound dilemma. For the opposing intuitions of dualistically minded philosophers such as Descartes and his intellectual descendants retain all their original potency. The provisional reading of the private language remarks merely brings about a standoff. It is a useful stepping stone, but we must go deeper.

[12] Wittgenstein (1958), §258.

1.3 How philosophers talk

For a more complete understanding of the private language remarks, we need to situate them within the overall project of the *Philosophical Investigations*. Wittgenstein's purpose, in a nutshell, is to persuade the philosophically inclined to give up the habit of thinking metaphysically, and to show that nothing is sacrificed in the process. He does this by demonstrating over and again that a claim or a question couched in metaphysically loaded language can be met with a response expressed in non-metaphysical language. These demonstrations do not supply answers. Rather, each one is a provocation—an invitation to renounce one little piece of metaphysics or another, an invitation whose very mode of presentation makes it evident that nothing is really lost in the renunciation.

The private language remarks are one such demonstration, for the case of inner experience. That is to say, they offer a challenge to language that is metaphysically loaded with the inner/outer distinction, yet they themselves are couched in language that is metaphysics-free. But the difficulty in coming to terms with the *Investigations* is that individual passages make only incomplete sense before the habit of metaphysical thinking has actually been broken. So, although the private language remarks have some power on their own, and although they may be of some help in breaking the metaphysical habit, they actually arise from a deeper wellspring of *silence* on metaphysical topics. Ultimately it is that wellspring of silence that we need to tap if we are to be able to cope with the most intransigent manifestations of dualistic thinking, such as Descartes' *cogito*. But this wellspring of silence on metaphysical subjects lies beneath the edifice of traditional Western philosophy, and the only way to reach it is to undermine the foundations of that ancient and well-guarded citadel.

So where should we start to dig? Which is more fundamental—mind, language, or world; thought, meaning, or reality; consciousness, knowledge, or being? Of course, the question itself is poorly framed, and already betrays too much metaphysics. But one thing seems clear. There is no philosophy without language. Philosophy surely cannot begin before speech and writing. So an effective way to treat the whole complex of symptoms from which the philosophically afflicted suffer is perhaps to begin with language itself. Accordingly, the opening pages of the *Philosophical Investigations* deal with language, confronting the temptation to think metaphysically about meaning. This quickly leads Wittgenstein to a view of philosophy that is intimately bound up with the extraordinary way philosophers talk.

The first remark in the *Investigations* is a comment on a passage in Augustine's *Confessions* in which he speculates on the process by which he acquired language as a child.[13] Wittgenstein writes,

> [Augustine's] words give us a particular picture of the essence of human language. It is this: the individual words in language name objects – sentences are combinations of such names. – In this picture of language we find the roots of the following idea: Every word has a meaning. This meaning is correlated with the word. It is the object for which the word stands.[14]

Wittgenstein begins his challenge to this received 'picture of the essence of human language' by telling a kind of story, albeit a rather dull one. He asks us to imagine a shopkeeper and a customer, and describes a severely stylized exchange between them. The customer presents the shopkeeper with a piece of paper. The words 'FIVE RED APPLES' are written on it. The shopkeeper responds by opening a drawer marked 'APPLES' and looking up the word 'RED' in a book containing colour samples. Then, he counts out five fruits from the drawer—saying out loud the numerals as he goes until he reaches 'FIVE', and carefully matching the colour sample for 'RED' with each apple he takes out. Eventually the customer receives the goods, and presumably departs with an air of satisfaction.

> 'But how does [the shopkeeper] know where and how he is to look up the word "red" and what he is to do with the word "five"?'– Well I assume that he *acts* as I describe. Explanations come to an end somewhere. – But what is the meaning of the word 'five'? – No such thing was in question here, only how the word 'five' is used.[15]

In this short commentary on his little story, Wittgenstein presents an alternative to thinking metaphysically about language and meaning and to speaking of them in metaphysical terms. He is not proposing that 'meaning is use', as a naive reading might suggest. That too would be a metaphysical claim. Rather, he is exhorting us to *replace* questions about meaning with questions about use. As his treatment of the shopping scenario makes plain, any description

[13] In the passage Wittgenstein quotes (*Confessions*, I, 8) Augustine writes: 'So, by hearing words arranged in various phrases and constantly repeated, I gradually pieced together what they stood for, and when my tongue had mastered the pronunciations, I began to express my wishes by means of them' (Augustine, 398/1961, p.29). A comparable view was articulated by Aristotle as long ago as 350 B.C.E: 'Spoken words are the symbols of mental experience and written words are the symbols of spoken words. Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images' (Aristotle, 350 B.C/1941).

[14] Wittgenstein (1958), §1.

[15] Wittgenstein (1958), §1.

of language *in use* will implicate its worldly role in human affairs. The scene is one of interaction between two people. The interaction involves not only signs and symbols, but also physical objects—apples, drawers, pieces of paper, and so on. So an account of the episode that omitted the words, the people, or the apples, would be useless.

The treatment of language we find at the beginning of the *Investigations* is exemplary of Wittgenstein's overall method. Wittgenstein is addressing the philosophically inclined, who are apt to enquire into the 'nature' or 'essence' of meaning—almost as if it were some kind of stuff, something 'out there', so to speak, that is amenable to rational investigation. This is metaphysical thinking. But Wittgenstein advises us to set aside the vexing question, with its taint of metaphysics, of what a word like 'five' means. Instead he simply offers a description of the way the word 'five' is used, a description that emphasizes its place in the hustle and bustle of daily human activity. There is no presentation of propositions and arguments in the style of traditional philosophy, no attempt to define the concept of number or pin down the nature of mathematical reality. Yet (we can imagine Wittgenstein saying) what has been left out of the description?

In a sense that will be elaborated more fully at the end of the chapter, language itself flows from the same wellspring of silence on metaphysical topics that is hinted at here. So even metaphysical thinking originates in this silence, including the deeply entrenched dualism that hampers contemporary attempts to acquire a scientific understanding of the mind. Its means of expression are found in the harmless language of everyday human affairs, and it erupts when certain ordinary words and phrases are used in extraordinary ways, exceeding their remit and giving rise to disquietude. Wittgenstein's strategy for neutralizing this insidious effect is to remind us of how these words and phrases are commonly used.

> When philosophers use a word – 'knowledge', 'being', 'object', 'I', 'proposition', 'name' – and try to grasp the essence of the thing, one must always ask oneself: is the word ever actually used in this way in the language-game which is its original home? –
>
> What we do is to bring words back from their metaphysical to their everyday use.[16]

So the treatment of language we find at the beginning of the *Investigations* has a twofold effect. First, it sanctions a strategy for tackling metaphysical thinking, which is to bring language back to its original home. Second, it supplies an exemplar of this approach for the case of meaning itself. It is not an *argument* for a new style of philosophy. Indeed, to level such an argument—at

[16] Wittgenstein (1958), §116.

least to do so with serious as opposed to ironic or deconstructive intent—would be a concession to the old way of thinking. Ultimately it stands only as an *invitation*, like a Zen master's finger pointing at the Moon.[17]

1.4 Doing battle with the interlocutor

In case it seems we are straying too far from the advertised concerns of this book, let's be clear that this excursion into Wittgenstein's thought is absolutely necessary. Whether explicitly or implicitly, contemporary debate on consciousness remains heavily under the influence of dualism, and the failure to escape from its grip is attributable to a lack of appreciation of the critical resources made available by such writers as Wittgenstein.[18] But Wittgenstein's thinking cannot be assimilated piecemeal. It has to be taken in as a whole. To benefit from his radical approach to consciousness, sensation, experience, and subjectivity, it is necessary to understand the stance he takes towards language and meaning, because from this he derives his overall attitude towards the sort of metaphysical thinking that dualism exemplifies. Accordingly, we must persevere a little longer in our exploration of the *Philosophical Investigations*, in order to fully apprehend its relevance.

Two thousand five hundred years of philosophical tradition will not succumb to a handful of trite aphorisms about the way we use words. So the invitation to a new way of thinking that inaugurates the *Philosophical Investigations* is unlikely to be taken up by many professional philosophers, least of all by those in the Western analytic tradition. They are far more likely to fixate on the Zen master's finger than to look at the Moon. (Scholars impressed by the finger's vigour may even choose to devote their lives to its study.) But, fortunately for the philosophically inclined who are troubled (rather than excited) by their metaphysical urges, and who are open to fundamental change, Wittgenstein goes on to present a sustained engagement with traditional philosophy in which the numerous twists and turns of the philosophical mind are anticipated and given their due.

These turns of mind find expression in the voice of an imaginary *interlocutor*, a very important character in Wittgenstein's later writings. Sometimes the interlocutor sounds like an ordinary person veering unsuspectingly towards metaphysical thinking, sometimes it seems to represent a sophisticated philosopher in the conventional mould, and sometimes the interlocutor is an alter

[17] The parallels between Zen Buddhism and Wittgenstein's later philosophy are many and deep, and have been noted by several authors, including Fann (1969, Chapter 10), Canfield (1975, pp. 383–408), and Phillips (1993, Chapter 12).

[18] Eilan (2002) and Shanahan (2005) are rare exceptions.

ego of Wittgenstein himself. The *Investigations*, together with the rest of Wittgenstein's later oeuvre, is a collection of exchanges with the interlocutor on a multitude of philosophical topics—meaning, reference, logic, mathematics, representation, knowledge, certainty, thought, sensation, and so on. The result is not a body of doctrine but a compilation of case studies in philosophy, like a collection of Zen kōans, whose aim is not to instil belief but to guide the student away from the dark forest of metaphysics and back onto the path of clarity, to 'show the fly the way out of the bottle'.[19]

Here is an instructive example, in which the interlocutor's voice, because it articulates thoughts that are so close to his own, merges with Wittgenstein's itself. It is of particular interest to us because it relates to artificial intelligence and the possibility of machine consciousness, topics that are not centre-stage in our present discussion, but which will be revisited later.

> Could a machine think? – Could it be in pain? – Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine.
>
> But a machine surely cannot think! – Is that an empirical statement? No. We only say of a human being and what is like one that it thinks. We also say it of dolls and no doubt of spirits too. Look at the word 'to think' as a tool.[20]

Wittgenstein, here as ever, is working through his own thoughts on a particular matter. He asks whether a machine could think. Then he pauses, and moves to a seemingly more daring possibility—that a machine could feel pain. He notes that there is a sense in which the body is such a machine. And we might stop there. But this is not the end of the matter, because Wittgenstein (or his interlocutor) exclaims that a machine *surely* cannot think. This is not an empirical statement, but a consequence of the way we use the word 'think'. The punch line is a familiar reminder that even difficult words—words with philosophical import—must be considered in relation to their everyday use. That is to say, we must ask what they are for.

The exercise of interpreting these passages is helpful because it trains us not to mis-read the rest of Wittgenstein's later writings. In particular, it is important to see that he is *not* making a claim about the *a priori* limits of artificial intelligence research. The declamation 'A machine surely cannot think!' should not be lifted out of its context and taken as an item of doctrine. It is the interlocutor speaking, even though the sentence is not enclosed in quotation marks, and it is only a waypoint in the exchange. (The punch line—the only sentence

[19] Wittgensetin (1958), §309.

[20] Wittgensetin (1958), §§359–360. Compare §281: '… only of a living human being and what resembles (behaves like) a living human being can one say: it … is conscious or unconscious.'

that carries anything like a stamp of final authority—is several sentences away.) It is immediately followed by the gloss, 'we only say of a human being and what is like one that it thinks'. The point is to steer the philosophically inclined away from the metaphysical chasm that opens up when they ask whether a machine could think. Finally, because the metaphysical temptations will return, perhaps in a different guise and in relation to a different topic, the take-home message is delivered: words are best thought of as tools.

Attending sympathetically to the interlocutor—on topic after topic, in one exchange after another—is like watching a tennis player trying to defeat a brick wall. Whatever metaphysically loaded argument the interlocutor propels at Wittgenstein, a metaphysics-free response just bounces right back, sometimes in the form of a rhetorical question, often using a striking metaphor. Whatever clever angles the interlocutor tries, however powerful his shots and whatever spin they carry, the brick wall always has a reply, each time as if to say, 'given up yet?' The overall effect is gradual attrition. The interlocutor is slowly ground down. Whatever his initial stance, eventual defeat by his own hand is inevitable. In the end, even for the most tenaciously philosophical reader, it becomes possible to discern a kind of *post-reflective* condition, wherein metaphysical thinking no longer arises.

There is no final, authoritative voice in the *Investigations*, and no sharp division between Wittgenstein himself and the interlocutor. Although the last word is always given to the post-reflective voice, and it is always metaphysically neutral, we never encounter an author who is himself a fully enlightened, perfectly post-reflective thinker, a finished product. Rather we meet a philosopher who is perpetually deflecting his own metaphysical tendencies, who has to engage them afresh every time. Our challenge as readers is to learn the same skill, to learn how to do battle with our own internal interlocutor, the devil on our shoulder who whispers philosophy in our ear. Wittgenstein's talent as a teacher is to show us that the devil can be defeated, that metaphysical thinking can, in this way, be transcended.

Because of their bearing on the topic of consciousness, our especial interest here is the private language remarks. With a better appreciation of the overall scope of the *Investigations*, we can now revisit them. Recall that Wittgenstein asks us to consider a hypothetical procedure for filling out the vocabulary of the putative private language. He proposes to keep a diary, and to mark the diary with the symbol 'S' whenever he experiences a particular sensation. The 'S' is a label for his private, inner experience. It is not a symbol, like the words we all use every day, that could be used to convey something about the sensation to another person. Nobody else can understand this language. But then how could Wittgenstein be sure that he was marking his diary correctly? What does 'correct' even mean in such a case? In particular, what guarantee could

there be that the symbol 'S' was being used for the same sensation on different occasions? What criteria of sameness could there be?

Now, where have these reflections left us? Have we arrived at the conclusion that the sensation itself does not exist? Certainly not! Drawing conclusions is contrary to the spirit of Wittgenstein's later philosophy. The whole purpose of the *Investigations* is to break the habit of metaphysical thinking, not to generate more metaphysics. Rather, Wittgenstein says, the sensation itself

> is not a *something*, but not a *nothing* either! The conclusion was only that a nothing would serve just as well as a something about which nothing can be said.[21]

The private language remarks promote a radically new stance towards philosophy, an entirely different way of thinking and talking, a post-reflective way. Hence, when the interlocutor responds to the private language remarks with a metaphysically loaded challenge—'You're a behaviourist. You are saying the sensation itself is a nothing'—Wittgenstein archly replies 'not a something, not a nothing', showing by example how to counter a metaphysically loaded challenge with a metaphysics-free response. However—and this is very important—the response truly engages with the challenge, and shows respect for its depth. The private language remarks refer to one of the cornerstones of metaphysics—identity—but without ever resorting to metaphysical talk themselves. The post-reflective philosopher turns the tables on the philosophically inclined thinker, provoking him on a sensitive subject at the very heart of metaphysics, not by demanding identity criteria but by making him feel uncomfortable for being unable to supply them.

The interlocutor's discomfort is only temporary, however. He is the representative of metaphysical disquietude in all of us—especially the philosophically inclined—and he does not fold so easily. He can handle a little discomfort, and knows of many other angles from which to approach the metaphysical traps and holes to which he feels so powerfully drawn. Wittgenstein's task is to anticipate as many of these angles as possible, and to counter each one of them. Hence the texture of the *Investigations*, its numerous thought experiments and cameo exchanges, tackling the same topics over and again in a variety of ways. For example, consider the 'beetle in the box', one of Wittgenstein's most arresting metaphors.

> Suppose everyone had a box with something in it: we call it a 'beetle'. No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing.[22]

[21] Wittgenstein (1958), §304.

[22] Wittgenstein (1958), §293.

The beetle in the box is, of course, analogous to a private, inner experience. Wittgenstein is addressing the same reflective inclinations here as in the diary thought experiment, and he does so with the same critical intention, the same hope that the interlocutor will renounce metaphysics, and will re-emerge from its other side in a post-reflective condition.

> But suppose the word 'beetle' had a use in these people's language? – If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a *something*: for the box might be empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is. [23]

The contents of the box is not a something, but not a nothing either. Wittgenstein is pointing out the futility of pushing ordinary language into metaphysical territory. Words like 'pain', 'sensation', 'experience', and so on have an everyday use. Our compulsion to extend this use in ways that lead to puzzling questions is natural, but ultimately pointless, because the problems that are thrown up in this way have no real substance. We only end up with pretend uses for the words, and when we realize this, our compulsion is lessened.

1.5 **Philosophical zombies**

Let's take stock. Thanks to Wittgenstein, we are now acquainted with a kind of post-reflective stance towards philosophy. Our review of his later work has been brief, and we have only studied a handful of his most celebrated remarks. So it should not be forgotten that their aim is not to convey a body of doctrine (for which a small number of critical passages might provide a useful summary), but rather to effect a change of attitude. This change is brought about through the exercise of a certain skill, which the post-reflective thinker learns to emulate and practise on himself. This is the ability to offer up a non-metaphysical riposte to a metaphysical thought—not by using mere trickery or clever words, but acknowledging the full weight of the original thought while undermining it with a metaphysically neutral reply. The remarks we have chosen to study are representative, and are pertinent to our particular enquiry. But the philosophically inclined can only attain a truly post-reflective condition through repeated exposure to this method, by repeated practice of the skill.

Nevertheless, we are now in a position to respond to the philosophical literature on consciousness from a post-reflective vantage point. In due course we'll return to the Cartesian *cogito*. But let's start with the philosophical concept

[23] Wittgenstein (1958), §293.

of a 'zombie'.[24] A person's zombie twin, in the contemporary literature, is physically identical to that person, cell for cell and particle for particle, down to the very finest sub-atomic detail. But there is a crucial difference between the zombie twin and the real person. Unlike the real person, the zombie twin enjoys no conscious experience whatsoever. It is not *like anything to be* the zombie. For the zombie, the lights are out, so to speak. There is no one at home.

According to the so-called *zombie argument*—revitalized and popularized by Chalmers—since we can imagine such a zombie twin, since it is logically possible to have all those physical properties without consciousness, no merely physical description of a human being, of their brain, body, and environment, however complete and precise, could ever entail the presence of consciousness.[25] A limit is thereby set to the scope of any scientific theory of consciousness. Although a scientific theory might be able to explain many or all of the psychological and behavioural properties we associate with consciousness, it can never approach the holy grail of phenomenal experience. It can never explain why it is *like something* to be a human being whereas it is not like anything to be a brick. So a yawning 'explanatory gap' seems to exist between the physical world and phenomenal consciousness. Scientific investigation is apparently restricted to the outer manifestations of a private, inner realm, and it can never approach consciousness *itself*.

So the argument goes. There are many ways to attack it. We might deny the purported conceivability of a zombie, the coherence of the concept. We might accept the concept of a zombie, but deny the presupposed strong connection between conceivability and ontology. (Perhaps we can imagine not only things that do not exist but also things that *could not* exist.) We could embark on a lengthy investigation of the logical relations among imaginability, conceivability, possibility, and various notions of existence. But all of this would be to work at the level of metaphysics, which, strange as it sounds, is to remain on the mere surface of the matter. Instead, let's play with the idea of a zombie for a while—not with a view to establishing firm conclusions, but in order to uncover some of the curious questions it throws up.

One vital thing is supposed to be missing from the zombie (or perhaps we should say *for* the zombie). As Chalmers puts it, 'all is dark inside'. So let's say that what is lacking in the zombie is an 'inner light'. We don't need to say anything about the nature of the inner light. This is just our name for what is supposed to be absent from zombies. Now, if we can imagine X's zombie twin, a creature who is physically identical to X but who lacks an inner light, then can

[24] Kirk (1974, 2005); Chalmers (1996).

[25] Chalmers (1996), pp. 94–99.

we not equally imagine the inner light of the real X being switched off for a day or two and then being switched back on again? X's friends and loved ones notice no change in her behaviour during her temporary zombiehood, so have no inkling of the tragedy, if tragedy it is, that has befallen her. Now, when X's inner light is restored, will she know that anything unusual has happened to her?[26]

Well, let's suppose that she does. Will she tell us about it when we ask her? Recall that her behaviour after the episode of absence, including all the things she says, will be identical to that of her normal twin, whose inner light is permanently switched on. So her reply to the question will be the same as her twin's. But if she knows she has had an episode of phenomenal absence then her twin is presumably equally aware of *not* having had such an episode. So whichever reply we hear—whether the twins (in unison) report nothing untoward or whether they (in unison) report a gap in their phenomenal consciousness—one or other of them is mysteriously incapable of articulating what they know about their inner light.

On the other hand, if she has no awareness of what has transpired, then how can any of us be sure—even from the inside, even from the supposedly privileged, first-personal perspective—that this is not happening to us in ordinary life all the time? Perhaps I can indeed be certain that the inner light is on right now. But for all I know it might never have been on in the past, even a moment ago. This is not a statement of scepticism about memory. For I may be able to accurately recall and recount what has happened to me—to my body, that is to say—and even to provide a totally convincing description of my past thought processes. But, for all I know, everything I remember might have happened while my inner light was off. Now here is a disturbing thought: If that were true, why would it matter?

Here is a variant of Wittgenstein's diary thought experiment, using the inner light in place of private, inner sensation. Suppose I decide to try to discover whether and when I suffer from zombie episodes, episodes of phenomenal absence. So I keep a diary, and I write 'L' in the diary on those days when my inner light is switched on. Later, I tell myself, I will be able to look back through the diary and, seeing an 'L', be sure that I was phenomenally present on that day, that all was not dark inside. Well, how could I trust any previous occurrences of 'L' marked in my diary, even if (contra Wittgenstein) I were capable of unilaterally identifying when my private, inner light was on? After all, if I was phenomenally absent on that day—away with the zombies, so to

[26] See Güzeldere (1997), pp. 43–44.

speak—I would have written an 'L' in my diary even in my zombie state. So the 'L' can tell me nothing. It seems to have no use *even in a private language*.

Here is another puzzle. How do we know that an inner light is the kind of thing that is simply either on or off? In addition to zombies (and part-time zombies), can we not imagine *half-zombies*? A half-zombie is also physically and behaviourally identical to his normal twin (and indeed to his zombie twin). It is like *something* to be a half-zombie, but not much. The being-likeness of a half-zombie is, perhaps, analogous to that of a chicken, or a fish. The inner light is on, but dimmed somewhat. However, unlike a chicken, who behaves in a way that suggests a much-dimmed inner light, and who has a modest brain to match, the half-zombie behaves exactly like a normal human being and has a fully functional, human-rated brain. Surely such a thing can be imagined. But then how can we distinguish between a full inner light and a half inner light? That is to say, how could we distinguish these things even theoretically, in order to make further philosophical progress? There seems to be no place for us to lodge a rational hook into the concept of an inner light, no way to reel it close enough in to formulate an argument as to whether or not it admits of degree.

The inner light is, of course, just the same as the beetle in Wittgenstein's box, except now the box contains a 'firefly' (shedding the inner light of true phenomenal experience). Everyone says they have one, but no one knows what a 'firefly' is except by looking into their own box. So it's possible that all the boxes contain something quite different. Some fireflies might glow more brightly than others. Some boxes might even be empty. Some boxes might be empty on some days but not on others, or glow more brightly on some days than others. None of this makes the slightest difference to how people use the word 'firefly'. So the word 'firefly' cannot be a name for the contents of the box, because the firefly itself—the inner light—has no place in the language-game whatsoever. It divides through, whatever it is.

So have we arrived at the conclusion that there is no such thing as an inner light, that there being something it is like to be something is no different from there being nothing it is like to be something? Not at all! Our error—albeit a natural and forgivable one—was to indulge in metaphysical thinking in the first place, and to avow such a conclusion would only be to repeat the error. The conclusion of the zombie argument is not wrong. But it is not right either. Metaphysical thinking—dualistic thinking—starts with the very first attempt to describe a zombie, and the ensuing discussion is saturated with it from start to finish. A suitably critical foray into the absurd world of the philosophical zombie should nudge us towards a post-reflective view of the inner, a condition wherein metaphysical talk loses its power to impress, wherein we feel less attracted to questions that previously seemed deep because we know how to

think through them, exposing the emptiness of the language used to conjure them in our heads, but without resorting to further metaphysics. However, to attain such a state is no easy thing. It requires work.

1.6 The subjectivity of exotic life forms

Even for those among the philosophically inclined who feel burdened by their inclinations, who are drawn to the idea of reaching a post-reflective condition, it's not possible simply to withdraw from abstract, rational thought. For the supposedly slain monster of dualism has a way of persistently re-surfacing in different guises, like Grendel with an endless supply of mothers. To work through one line of dualistic thinking, undermining its foundations by rejecting its language, is not enough to acquire the means to work through every line of dualistic thinking in the same way. This is why the *Philosophical Investigations* and the rest of Wittgenstein's later writing is a compendium of case studies.

In our own compendium of case studies, our attention now turns to Nagel's well-known essay on subjectivity and to the question posed by its magnificent title 'What is it like to be a bat?'.[27] The point of the title is that a bat is a creature quite unlike a human being. Humans run, walk, and sit on sofas, whereas bats fly and dangle upside-down in holes. Mostly confined to the ground by gravity, humans navigate a world of more or less two dimensions, whereas bats have the freedom of all three. To negotiate their flattened world humans use their eyesight, whereas the visually impaired bat employs echolocation. Also, bats eat flies, which is uncommon in human beings. Someone who enjoys hang-gliding knows what it is like to fly, and a blind person who can 'hear' the walls of a room by tapping a white stick on the floor knows what it is like to use echolocation. Such people may be in a better position than the rest of us to understand what it is like to be a bat. But surely no human being, however prosthetically enhanced, could ever *really* know what it is like, could know what it is like for a bat to be a bat, so to speak, and no amount of scientific investigation can change this.

According to Nagel, this thought suggests an insurmountable obstacle to the provision of an objective, scientific theory of consciousness, because 'every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view'.[28] How could the subjective experience of a bat be described except from a bat's point of view? Much the same point might be made for humans alone, without appealing to bats or other strange forms of life. An objective,

[27] Nagel (1974).

[28] Nagel (1974), p. 437.

scientific theory of human consciousness would have to abandon the subjective point of view, something that is surely essential to conscious experience. But confined to humans, this 'subjective point of view' sounds little different from the elusive 'inner light' discussed in the previous section, a notion that in itself no longer has much of a grip on us. However, Nagel has found a new means of provoking us, of arousing our philosophical anxieties. Using exotic life-forms, he can evoke once more the dualistic picture of a metaphysical chasm between inner and outer.

Consider the octopus. The octopus displays many behavioural attributes associated with intelligence, and it seems reasonable to assume it is like something to be an octopus. Yet the octopus belongs to the class of cephalopods, who sit on a branch of the phylogenetic tree that split from our own before evolution settled on the basic pattern of neuroanatomy common to all vertebrates. The central nervous system of an octopus, which includes separate mini-brains for each of its eight tentacles, doesn't even have a thalamus. So we lack the potential handle on its subjective experience that a familiar underlying neuroanatomical blueprint might provide. And if we allow our imaginations to roam further into the realm of exotic life forms, we might wonder how we could ever know what it was like to be an extraterrestrial that evolved on a different planet, or a putatively conscious robot whose internal workings were engineered without reference to biological precedent.[29]

But let's stay with bats and see if we can unpack Nagel's reflections. At their core is a commonsense distinction between subject and object. To know that a bat has wings is not the same as actually to have wings, and to know that a bat navigates by echolocation is certainly no help in the dark. However diligently I study aerodynamics I will never sprout wings, and no amount of expertise in acoustics will enable my tongue to emit ultrasonic clicks. These are hardly profound observations. However it is surely *like something* to be a bat, in the way it is like something to be human. So, remarkably, for a bat, unlike an aircraft, it is like something to have wings. Now, I *know* what it's like to be a human being with two legs. But I could *never* know what it's like to be a bat with wings. And to know something is surely to *have knowledge* of some thing, which means to be *in possession* of certain *facts*. So it looks as if we have discovered a whole new realm of facts. Let's call them *phenomenological facts*. And doesn't the example of the bat show that the phenomenological facts about a sufficiently exotic form of life are, in principle, inaccessible to us?

But this is all smoke and mirrors. We started out with a perfectly harmless distinction between subject and object. Then we contaminated it with the

[29] See McFarland (2008).

already discredited distinction between inner and outer, lending it a spurious air of metaphysical mystery. And finally, to reinforce the effect, we introduced epistemology and insidiously shifted from one use of the word 'know' to another. All we have really done is take an elegant and useful locution—the phrase 'like something to be'—and serially abuse it until it yields metaphysics. The conjuring trick needs to be exposed and the harm it has done reversed, so that we can return to the path of post-reflective tranquillity. When we say: 'However much I know about bats, I will never know what it's like to be a bat', we have said no more than the following. It is like something to be human, and it is like something to be a bat. But humans and bats are very different, and I am a human not a bat. What is missing from this description?

But if nothing is missing from this metaphysically inert reformulation then what is the answer to Nagel's original question? Does the question of what it is like to be a bat have *no answer at all*? Surely it is like something to be a bat, and it seems fair to wonder *what* it is like. If confronted with an extraterrestrial creature, would we not be curious what it was like to be that creature? And if we built a robot in the laboratory that behaved like a human, would we not want to know if it were capable of suffering, whether we had ethical responsibilities towards it? Surely there are facts of the matter here, regardless of whether we can discover them. But the answer to Nagel's challenge, and the reply to all these questions, is that *nothing is hidden*.[30] That is to say, nothing is metaphysically hidden. Of course, you may dissemble and keep your thoughts from me. And we may be prevented by time and space from encountering extraterrestrials. But these senses of hidden are no more mysterious than a ball under a magician's cup. If further investigation were feasible, all would be revealed. However ignorant we are of octopuses, aliens, and robots, nothing about them is truly hidden from us, that is to say on the other side of a metaphysical veil.

1.7 Thought experiments with peculiar scientists

To sharpen the point, let's conduct another thought experiment. Imagine a perfect mechanical scientist—a machine whose internal workings bear no resemblance to our own, but which is capable of acquiring knowledge, acting on it, communicating it, and so on. The mechanical scientist can build perfect theories of the brain, of psychology, and of physics, and it can hold a normal conversation about these theories that impresses the world's leading experts. But there is nothing it is like to be that mechanical scientist. It is just a knowledge cruncher. As with the philosophical zombie, there is no one at home, and

[30] Wittgenstein (1958), §435.

all is dark inside. Whether a machine with this combination of attributes is empirically possible is not at issue here. As with all such thought experiments, what matters is the impact the imaginative excursion has on the way we think and talk about the relevant concepts. Now, what could the mechanical scientist possibly say about what it is like to be a human being, or a bat, or anything else? Indeed, how could it know that it was like anything to be anything? What inkling could it have that the universe contained consciousness at all?

Well, it does not follow from the fact that a certain property does not hold of a person, that the person in question cannot understand properties of that sort. A doctor can treat a patient for a disease from which she has never suffered herself, and discuss such diseases with similarly blessed colleagues. A car mechanic can replace the spark plugs in an engine even though he lacks an engine himself, and can teach an apprentice to do the same. Likewise, the mechanical scientist is no more handicapped than a human scientist when it comes to finding an empirical theory of human consciousness, of what it is like to be human. We might remark that it will never 'know' what it is like to be human in the way that we do. But what more does this amount to than the empty observation that it will never *be* human? The insidious word 'know' here misleads us into metaphysical thinking. In what useful or interesting sense could it follow that we *know* more than the mechanical scientist solely because of certain things we *are* that the mechanical scientist is not?

Is there perhaps a danger here of trivializing the sense of 'know' in knowing what it is like to be something? To know what it is like to be something is surely more than simply to be that thing. Someone who knows what it is like to eat pheasant, for example, can also *communicate* something of the experience to others. And a person who knows what it is like to make love to their spouse can also *imagine* what it might be like to make love to their neighbour. Much of this, surely, is barred to the mechanical scientist. Whereas the mechanical scientist may, through judiciously chosen sentences, be able to evoke in a human feelings it does not itself enjoy, the most skilled human poet is powerless to evoke feelings of any sort at all in the mechanical scientist. Very well, but there is nothing in this more nuanced treatment of the word 'know' capable of reinstating the original argument. What is barred to the mechanical scientist is so because of certain facts that do not hold for it, and this is no barrier to its understanding facts of that sort. Nowhere here do we find a new kind of fact.

The mechanical scientist is a conceptual relative of another peculiar imaginary scientist, namely Mary, who plays the leading role in the so-called *knowledge argument*, another attempt to establish a form of dualism.[31] Mary, who lives in

[31] Jackson (1982); Ludlow, *et al.* (2004).

a future when our knowledge of neuroscience is total, has a complete grasp of the physics, neurophysiology, and psychology of colour vision. However, Mary herself has never seen any colour other than black, white, or grey, as she has grown up and lived all her life in a specially constructed monochrome room. Now consider what happens when she leaves the room for the first time to see the world in its full colourful glory. None of her expertise in colour vision, so the argument goes, will diminish the shock and pleasure of seeing colours for the first time. When she sees red for the first time, she learns something. She discovers what it is like to see red. As, according to the thought experiment, she was in possession of all the objective, physical facts beforehand, this entails that what she has learned is not an objective physical fact. So it must be an altogether different kind of fact, something that is beyond the reach of objective scientific enquiry.

But the appropriate question to ask is not what Mary learns when she sees red for the first time, but what it is about her that changes. On one level, we might say, what happens is this. Mary exclaims. She smiles. In due course she becomes able to pick out the red objects among others. She becomes able to recall the colours of objects she has seen, and to report them to others. She might learn to paint. On another level, we might say, what changes is something like the following. Light of a particular wavelength hits her retina for the first time, and sets off a cascade of neural events. These put her brain into a state it has never been in before. As a result of this, and in combination with further ensuing events, her behavioural repertoire is duly enlarged. Whichever way you look at it, certain objective properties that had never previously been applicable to Mary become applicable, and this happens regardless of whatever previous understanding she may or may not have had of properties of that sort. Likewise, there are many properties that will never be applicable to a human being but this in no way limits the scope of a human scientific enquiry into properties of that kind.

The final hypothetical scientist we shall consider is perhaps the most peculiar of all. Imagine an extraterrestrial whose scientific understanding of the human being is as full and sophisticated as the mechanical scientist's. However, the extraterrestrial scientist does not speak any human language. Indeed, the extraterrestrial is physiologically and behaviourally so alien that all attempts to establish communication with it have failed. Of course, the extraterrestrial has an excellent *theory* of human language, explaining the strange noises and marks humans are prone to make and the role those noises and marks play in human society. But the theory is expressed in the scientific language of the extraterrestrial and deployed for its own inscrutable ends. The extraterrestrial scientist either cannot or will not make use of that theory to open a channel of

communication with the human race. Finally, it is not like anything to be the extraterrestrial scientist. Somehow, evolution on the extraterrestrial's planet managed to produce intelligence without consciousness.

In essence, then, the extraterrestrial scientist is very much like the mechanical scientist, except that its theories are supposedly incomprehensible to us. The one point of contact that we had with the mechanical scientist, namely a shared language, is absent. Because the mechanical scientist is inducted into our language and society, it can present its theory of human consciousness in human terms. So even though it is not like anything to be the mechanical scientist, it can answer questions about why it is like something to be human (and indeed why it is like nothing to be the mechanical scientist itself). The resources to do this come with the language, are part of the package, so to speak. But the extraterrestrial scientist is in a less favourable position. So whereas the mechanical scientist might be in a position to develop a scientific theory of consciousness, the extraterrestrial scientist surely is not. Something will always be missing from its objectively flawless theories, namely subjectivity, which it not only lacks, but cannot even talk about. Does this not then entail that the set of objectively describable facts about the universe does not encompass all the facts about the universe? The chasm of dualism gapes open once more.

Now, what exactly is supposed to prevent the extraterrestrial scientist from developing a full theory of human consciousness? Nothing is hidden, so how is the extraterrestrial scientist not in the same position as the mechanical scientist? Everything is available to it. So where in setting up the thought experiment did we go astray? At what point did we begin to indulge in metaphysical thinking? The science fiction looks impeccable. As with the mechanical scientist, it begs the question of the empirical possibility of intelligence without consciousness. But notwithstanding this, our imaginations are not overly stretched by the proposed creature. No, the mistake was to announce that the extraterrestrial theory omitted something that was included in the human theory, and then to take for granted the notion that the omitted something might be part of some fundamental reality. Yet the omitted something is nothing more than the dubious 'inner light' that is lacking in the philosophical zombie.

In conventional philosophical terms, we would characterize the extraterrestrial and human theories as having incompatible 'ontologies'. But what does that mean? All we have here are two different forms of life, two different languages, two different kinds of activity, and nowhere in this two-ness is a metaphysical division exposed. Sometimes scientific communities use words in highly distinctive ways ('molecule', 'gene', 'wave', and so on). If a scientist points out that these are the things that 'exist' according to her theory, then this is just the right way to talk given the practices of the scientific community,

practices that are especially rigorous and that demand a strong empirical sanction for using words in that sort of way. But for anyone—scientist, philosopher, or layperson—to go a step further and claim that the 'fundamental nature' of reality is revealed by a scientific theory is to make a dangerous and unnecessary metaphysical move. So talk of the superiority of one theory's ontology over another's that appeals to some altogether *hidden* order of reality—such as the realm of private, inner experience—is doubly misplaced.

1.8 The subject adrift in time

Having confronted a series of arguments like those in the preceding compilation, and having safely deflected them all without resorting to metaphysical thinking, the philosophically inclined individual may perhaps think she has attained a sufficient degree of post-reflective composure to wrestle with Descartes' *cogito*. But she would be wrong. For despite the passage of 350 years, the *cogito* remains the most potent weapon of dualism. Although we have repelled several of its most talked-about contemporary successors, the *cogito* itself will not submit so easily. The Cartesian thought is clearly a threat to the condition of post-reflective quietude we have been cultivating. It is a metaphysical argument with an overtly dualistic conclusion. But in contrast to the metaphysical stirrings engendered by zombies, exotic life forms, peculiar scientists, and so on, no suitable metaphysics-free response to the *cogito* is suggested by the private language remarks.

The power of the *cogito* derives from its parsimony. Each of the contemporary arguments we have examined trades on our acceptance of a problematic relationship between the physical world and the sensations, feelings, thoughts, and so on that occupy the conscious mind. That is to say, each of these arguments depends on the distinction between inner and outer. The private language remarks do their remedial work by using the inner/outer distinction against itself, causing the supposedly problematic relationship to vanish. But the *cogito* relies on no such distinction. It doesn't need the physical world. It doesn't need sensations or feelings. All it calls upon for its argument to work is the naked, thinking subject.

The *cogito* continues to bewitch us because it's hard not to embrace the conception of the subject that it conjures with, a conception whose essence is 'self-presence'. When a philosopher avows that the proposition 'I am thinking' is necessarily true each time she entertains it, the thought in question is essentially reflexive. It is a thought about itself, and it evokes a thinking subject that is present to itself. Moreover, this self-present subject is also self-sufficient, and this is evident from the meagre resources required to evoke it. Floating free of the material world, the self-present subject's existence presupposes nothing

but the self-present subject itself, its own presence to itself. It is this notion of the Cartesian subject—self-present and self-sufficient—that leads to a renewed irruption of the dualistic impulse. How could the material world accommodate such a thing? Surely no objective scientific theory of consciousness could ever hope to answer this question.[32]

When the method of doubt is spent, the hard, indestructible nucleus of this self-sufficient subject is all that is left. But in this barren wasteland where ordinary life is forgotten and the philosopher is left with nothing but the purity of her own self-given, self-present consciousness, an all-consuming ogre lurks—Time. Self-given subjectivity, if the concept is to be given any credence, must be amenable to preservation for later reflection. There can be no *mere instant* of self-presence. A moment of self-presence that left no trace of itself behind would be just a chimera, no better than a nothing.[33] So the Cartesian philosopher is obliged to address the following question. How is a fleeting moment in the 'flow of consciousness'[34] saved from annihilation the instant it comes into being? How is it preserved for later re-appropriation by the flow? As Husserl wrote,

> … all experiences flow away. Consciousness is a perpetual Heraclitean flux; what has just been given sinks into the abyss of the phenomenological past and then is gone forever. Nothing can return and be given in identity a second time.[35]

James expressed a similar thought.

> Let any one try, I will not say to arrest, but to notice or attend to, the *present* moment of time. One of the most baffling experiences occurs. Where is it, this present? It has melted in our grasp, fled ere we could touch it, gone in the instant of becoming.[36]

---

[32] For an overview of standard responses to the *cogito*, see Williams (1978), Chapter 3.

[33] A related point is made by Derrida (1967/1973, Chapters 4 & 5), commenting on Husserl (1911/1991). According to Zahavi (2005, p.70), Derrida's reflections have the 'disturbing implication that consciousness appears to itself, not as it is, but as it has just been' which suggests 'a blind spot in the core of subjectivity'. From the present standpoint, this thought is liberating not disturbing.

[34] The phrases 'flow of consciousness' and 'lived experience' are frequently used by scholars and translators of Husserl. James's notion of the 'stream of consciousness' (1890/1950, p. 239) is comparable.

[35] Husserl (1911/1991), p. 360. Husserl's way of saving the subject from the abyss of the phenomenological past is to elaborate a theory of the structure of inner time-consciousness, according to which our awareness in the present moment has three components – (immediate) primal impression, (backward looking) retention, and (forward looking) protention.

[36] James (1890/1950), vol. 1, p. 608. See also Andersen & Grush (2009), who exhume a number of relevant influences on James and Husserl.

Indeed the present is always gone. But some trace of it must linger if the conscious moment is to be given to the subject for later reflection. For the philosopher who adheres to the doctrine of self-sufficiency, who believes that her self-present, self-given consciousness does not depend for its being on world, society, or (public) language, all that can be preserved of a conscious experience after it has receded into the phenomenological past is what consciousness can clasp to itself, what it can offer to itself for later re-appropriation. Without recourse to any form of external recording, lacking any means of repository in the world, this is all that can endure *of* consciousness, and all that can endure *for* consciousness, in its pure, self-present, self-sufficiency.

Of course, for the idea of preservation to make sense, there has to be something to preserve in the first place. Let's call whatever it is that must be preserved for later reflection, or what is at least amenable to preservation for the very idea of self-presence to make sense, the *original conscious experience*. Now, for the philosophical champion of self-sufficiency the following question requires an answer. What criterion could be used to measure the original conscious experience against whatever trace of itself it leaves behind? What criterion could be used to assess the fidelity of the trace, and thereby to distinguish successful preservation from a mere sham? As the trace in question belongs to consciousness alone, this criterion cannot appeal to any outward behaviour that the original conscious experience might have given rise to at the time. Nor can it appeal to any aspect of whatever neurological activity that might have accompanied the original conscious experience. Bodies and brains are in the world and have thus been discarded by the method of doubt.

The only possible criterion for measuring the original conscious experience against the trace within which it is allegedly preserved for the self-sufficient conscious subject would be a strictly private one. (By now it should be clear that the strategy we are pursuing echoes Wittgenstein's private language remarks.) Only from the point of view of the subject would it be possible to assess the trace of a conscious experience against the original of which it is supposed to be a record. But from the subject's point of view, the original experience is always gone, and the trace is all that is left of it. So what could possibly distinguish a veridical trace from a false one? Indeed how, using nothing but private criteria, could any distinction be drawn between a later reflection wherein the original conscious experience has been faithfully restored and a later reflection wherein the original conscious experience is radically compromised yet *seems* like it has been faithfully restored?

Very well. Perhaps no such criteria are to be had, and the idea of a *trustworthy* trace is suspect. But surely for the *cogito* to establish the self-sufficiency of the conscious subject, any trace will do. It does not have to be trustworthy. It just has to be a thread that joins the past of the conscious subject—the immediate

past will suffice—to the present of that same subject. Yet what distinction could be drawn, calling only on the resources that are available within the subject's sphere of privacy, between a genuine trace and the mere impression of a trace? What sense can be made of the distinction between a trace that really does connect the past of a conscious subject to its present, and the instantaneous flashing into being of a phantom trace that connects nothing to nothing and belongs to a subject that is equally ephemeral?

These rhetorical questions are not an expression of scepticism about memory (much as the rhetorical question at the heart of the private language remarks is not an expression of scepticism about memory). The issue here is not how we could ever *know* that a private, self-given trace of the original experience was trustworthy or genuine. Rather, the issue is what it could even *mean* for such a trace to be trustworthy or genuine. So, is the conclusion that the subject is radically divided from itself by time? Have we shown that self-awareness is at best flawed and at worst just an illusion? Not at all. To affirm any such thing would be implicitly to approve the terms of a metaphysical debate on self-presence when we would prefer to wash our hands of it altogether. The conclusion, rather, is this. When the concept of self-sufficient subjectivity—a concept that we are tempted by thanks to the *cogito*—is submitted to critical examination, it is found wanting. And this is enough to quell the resurgent threat of dualism.

We're almost done with Descartes and the *cogito*. But not quite. Because the denouement we are moving towards is conveniently characterized by comparison and contrast with the project of the *Meditations*, which opens as follows.

> I had to undertake seriously once in my life to rid myself of all the opinions I had adopted up to then, and to begin again from the foundations, if I wished to establish something firm and constant in the sciences.[37]

A similar quest for foundations motivates the present chapter. But the foundations we are unearthing are not those Descartes believed he had found when doubt had run its course. Towards the end of his life, Wittgenstein wrote a series of remarks that were posthumously published as *On Certainty*. Many of these remarks engage directly with the sceptical elements of the *cogito* (although he never names Descartes explicitly), and it is here that we find the clearest allusions in the literature to the position we are approaching.

> If you tried to doubt everything you would not get as far as doubting anything. The game of doubting itself presupposes certainty.[38]

[37] Descartes (1641/1968), p. 95.

[38] Wittgenstein (1969), §115.

> My *life* consists in my being content to accept many things.[39]
> Doubt itself rests only on what is beyond doubt.[40]

Wittgenstein is gesturing at what is simply taken for granted in our lives, such as our shared world, our language, each other. These things (which we should not call 'things') form a necessary backdrop to the expression of the sceptical propositions on which the argument of the *cogito* rests. In other words, its conclusion denies the conditions necessary for its formulation. When we dig beneath the *cogito*, when we uproot it, only then do we expose a true foundation, the common source of language, reason, and science. All are seen to originate in a kind of silence, what we might call the *fundamental starting point*. This is the everyday silence on philosophical matters that comes before metaphysical anxiety, a silence that we can hope to re-inhabit by entering a post-reflective condition. Within the silence of the fundamental starting point, which precedes the *cogito* both conceptually and chronologically, subject and object are not divided. We might say that it manifests an ineffable harmony between mind, world, and language, a harmony that underwrites our being at home in the world. But to say this would be to violate the very silence in question.

1.9 The proper silence of first philosophy

We have come to this pass because we are interested in some very large questions. What is a human being? What am I? And we have alighted on two subsidiary questions. First, what is the nature of the conscious subject? Second, what are the limits of empirical enquiry? In the context of the *cogito* these two subsidiary questions are subtly but intimately entwined, because the Cartesian invocation of the naked subject threatens to prescribe a limit to the scientific study of consciousness. Science, in this conception, is condemned never to touch the secret inner plasm of consciousness itself. It can only fumble with its outer manifestations. The danger is that when we pull up the metaphysical roots of this way of thinking, removing the limitations it imposes, we will excavate the foundations of scientific enquiry itself. This is a very real danger because, before we can complete our project, before we are reduced to the silence of the fundamental starting point, we will be forced to use arguments that could be misconstrued as leading to a position that conflicts with both common sense and the ideals of rational investigation.

So it's important to see reduction to post-reflective silence not as a philosophical terminus, but rather as the discovery of a true first philosophy, not as

[39] Wittgenstein (1969), §344.

[40] Wittgenstein (1969), §519.

justifying a destructive or nihilistic standpoint from which reason and truth can excusably be degraded, but rather as a temporary resting place from which to reinstate our commonsense license to talk to each other about the world and our engagement with it, and to accord (more or less) the usual privilege to such concepts as truth, reason, and objective reality. To be reduced to post-reflective silence is, in a manner of speaking, to die to philosophy. But from the silence of this fundamental starting point we can be reborn into philosophy, and the project of gaining a greater understanding of our inner lives can be safely revisited.

But we're not quite there yet. Nothing less than the wholesale overthrow of metaphysics will fulfil our quest. We have confronted the concept of meaning and the concept of subjective experience. But no single set of foundational concepts exists, small in number and the same for every enquirer, whose fracture indisputably revokes metaphysics.[41] The dam, if it cracks at all, will crack differently for each thinker. Putting the concepts of meaning and consciousness under pressure greatly weakens the whole edifice. But for those brought up in the analytic tradition, a number of obvious additional targets are salient. These include the concept of truth, and indeed the very concept of a concept itself. There is a problem with analytic philosophy's concept of a 'concept' to the extent that it aspires to denote something self-identical, singular, and fixed, because our critique commends the view that the way we use a word is, by nature, irreducibly dependent on context, subject to polysemic variation across the membership of a community, and open to never-ending revision and alteration over time.[42] What could it mean to 'clarify' a concept on this construal? On what could the process of clarification possibly hope to converge? Against what standard could clarificatory progress be measured? When challenged to clarify the concept of a 'concept', analytic philosophy is dumbfounded.[43]

[41] Quine repudiates the quest for foundations. Borrowing a metaphor from Neurath, Quine (1960, pp. 3–4) famously likens both philosophy and science to 'a boat which, if we are to rebuild it, we must rebuild plank by plank while staying afloat in it. … Our boat stays afloat because at each alteration we keep the bulk of it intact as a going concern'. Quine (1969, pp. 126–127) claims that 'there is no external vantage point [away from the boat], no first philosophy'. But in the silence that is proper to a true first philosophy, there is neither boat nor sea.

[42] Consider Wittgenstein's treatment of the word 'game', for example (Wittgenstein, 1958, §§66–68): '[The use of the word] is not everywhere circumscribed by rules'. For Derrida too, the potential for 'grafting' words into new chains of signification entails that no amount of context can fully enclose a concept (Derrida, 1982, p. 317).

[43] For a relevant treatment of the aims and limitations of conceptual analysis, see Moore (2001). We shall revisit the concept of a concept in Chapter 2.

What about the concept of truth? Surely this is inviolable. Surely its sanctity must be preserved at all costs. Well, even the golden light of truth must take its place in the great embroidered cloth of our language games. But this is not an affront to common sense. We are properly inclined to say that mathematical theorems are true irrespective of whether anybody ever proves them, and that there are empirical facts about the universe that hold whether or not science ever reveals them to us. Of course, we don't want to deny such things. But if we agree that questions about meaning should be met with descriptions of the way words are used, then we're obliged to apply this principle to the language of philosophy. When asked what truth is, what more are we asking than what the word 'truth' means, and how could we do better than to reply with a description of the ways words like 'truth' are used?

Consider the definition of truth in Aristotle's *Metaphysics*.

> To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true.[44]

Our interest here is not scholarly, so we're not concerned with what Aristotle actually intended. But this well-known pronouncement is a useful foil. We might take Aristotle to have articulated a worthless platitude. But it's more interesting if we take him to be stating something troublemaking, specifically that what we say can be overlaid on what is the case—on Reality, if you will—and the match between the two duly assessed. This is troublemaking because it is liable to rouse the dormant beast of metaphysics. For if we accept this notion, we're obliged to address some nasty questions. What is Reality, and how is what we say to be measured against it? Yet if we deny the notion we seem to be denying Truth and Reality altogether. So Aristotle's statement is either empty, or it is intolerably problematic.

If we refused to engage with metaphysics in the first place we would obviously avoid these difficulties. But those of a philosophical inclination are unlikely to find satisfaction while the big questions are left hanging in the air. On the contrary, the only way to become convinced that nothing is lost when metaphysics is jettisoned is to engage with it thoroughly, but to do so without becoming immersed in it. The standard repertoire of philosophical questions about the character of mind, language, and reality is insistent. But our response to it must have a neutralizing effect. Most importantly, we should never attempt to 'reach beyond the language-game', so to speak. That is to say, we should make no appeal to anything outside the activities of groups of users of a common language.

[44] Aristotle (350 B.C./1924), 1011b25.

Yet how can we tolerate the apparent tension between, on the one hand, high-sounding talk of the objectivity of mathematics and, on the other, commonplace descriptions of the ways we use words like 'true', 'correct', 'valid', and so on? Surely we are in danger, with the latter, of naively equating truth with mere agreement. The dissenter will point out that no amount of agreement that, say, π equals 3.2 would make it so. But no attempt has been made to equate or identify truth with anything. No metaphysical claim has been advanced about what truth *is*. There is no conflict here among competing metaphysical positions, and no tension arises. It is part of the language game of mathematics (so to speak) to agree that mathematical truth is more than mere agreement and is sanctioned only by proof. Moreover, it is part of the language game of everyday truth that on certain topics (such as medicine, or indeed mathematics) we defer to experts, whereas in other matters we defer to those who are simply better placed to know something than ourselves, such as the witnesses to a crime. In short, only a hopelessly inadequate description of the way we use a word like 'true' would reduce it to mere agreement.

To clarify matters, let's consider the pragmatist's conception of truth, which on first examination might seem compatible with the present project. According to Peirce, for example,

> [t]he opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real.[45]

Of course, it would not be 'true' according to Peirce's carefully phrased definition that π equals 3.2 even if everyone agreed that it did, because he charges us to investigate, and nothing that would count as a proper investigation could legitimize such a proposition. So far, then, the pragmatist's conception seems allied to our own reflections. But Peirce's characterization invites a string of other critical questions. What is it about an opinion that makes it *fated* to be agreed upon? When exactly is *ultimate* agreement reached? To rise to the bait, to seek answers to these questions, as a contemporary philosopher with pragmatist leanings might, would be to make a mistake. Our goal here is neither to define truth nor to construct a theory. The only aim is to break the habit of metaphysical thinking.

Now, suppose someone were to claim that there is something deep and important in common between, say, truth in mathematics, truth in science, and truth in everyday life. The use of the same cluster of words in these different contexts is more than coincidence. There is an underlying concept here—Truth. A logician, taking up the baton, might point out that the tautologies of predicate calculus remain valid whatever the subject matter. So we are free to

[45] Peirce (1878).

combine mathematical, empirical, and even ethical and aesthetic propositions in any way we choose, and remain confident in our ability to reason correctly with the results. Suppose we were challenged to account for this seemingly remarkable fact. Well, what is there to account for? What more is there to say? The logician has said everything a logician can say, and we should resist the temptation to embellish her words with further metaphysical pronouncements. (As to whether logic, propositions, tautologies, and so on play a significant role in ordinary life, well that's another matter altogether.)

The offence of metaphysics is the attempt to exceed the language game, to get at what is hidden, what lies beneath, to anchor the conceptual apparatus of philosophy (experience, truth, reality, and so on) in a firmament outside the language game. But it can only do this by appealing to something that is, so to speak, *further in* than the language game, namely the authority of the conscious subject, or to something *further out* than the language game, such as the Platonic heaven. But we have already disabused ourselves of these notions. So, do mathematics, empirical science, and honest everyday speech ever articulate truths that obtain independently of our minds? It would be improper to answer yes, and perverse to answer no. The fault lies in the question, which presents an illusory dilemma. To grasp either horn of the dilemma would be to indulge in metaphysical thinking. Instead we should offer the sort of response furnished earlier, whose aim is not to supply a philosophically satisfactory account of Truth and Reality or any such thing, but rather to displace the need for such an account—to gesture, that is to say, at silence.[46]

Yet how can the privileged status of science survive such an assault? If truth is characterized merely in terms of what we say, then on what grounds can one form of enquiry claim authority over another? Well first, no characterization of truth has been put forward. Instead, the aim is to dispel the feeling that truth stands in need of such a characterization. Second, to speak of the grounds for privilege is inappropriate. Certain practices simply assume authority over others in the right circumstances, thanks to their manifest success (and the powers of persuasion this success confers on their advocates). It would be disingenuous to pretend not to feel the need for a metaphysical explanation of why this is so, yet we must resist the temptation to supply one all the same. It is inherent

[46] The position we have arrived at, as well as the strategy for getting there, bears comparison with the thinking of the 2nd century Indian philosopher Nāgārjuna (2nd Century/1995): 'To say "it is" is to grasp for permanence. To say "it is not" is to adopt the view of nihilism' (15:10). Nāgārjuna's writing influenced the development of Zen Buddhism, and has drawn contemporary comparison with both Wittgenstein and Derrida (Kasulis, 1981, especially Chapter 2; Loy, 1992; Garfield & Priest, 2003).

in the language game of truth to say that truth is more than just a language game, and there we must let the matter rest.

Finally, in philosophy, 'one gets to the point where one would like just to emit an inarticulate sound'.[47] We see that the very moment we open our mouths to speak or lift our pens to write, we have already gone wrong. In the end, only a kind of silence can serve as a proper starting point, as the first philosophy that Descartes was seeking. What is meant here by 'silence' is not literally the absence of speech, but rather a silence in the presence of certain questions, questions that lead to metaphysical reflection. Yet this is not the silence of an infant or an animal in the face of such questions. Nor is it the silence of an adult human who is simply not philosophically inclined (although it has something in common with that kind of silence). Rather, it is a silence that can come only *after* reflection, a silence to which a person is reduced by sustained confrontation with metaphysical questions.

At the same time, the silence we are reaching for here is *prior* to the language of philosophy. It is the original source from which the language of philosophy springs. Indeed, it is the source from which all language springs.[48] As such, it is an end that contains the seeds of a new beginning. What gives depth to the post-reflective condition is what comes after it. Although the silence of the pre-reflective child also carries the seeds of future thought, she faces a future of reflective anxiety. In post-reflective silence, everything remains available to be said. Yet philosophy has been given peace, and the path back to silence is always open.

In the proper silence of first philosophy everything important remains the same. The daily round of human affairs continues.[49] Everyday human language is untouched. As Wittgenstein affirms,

> Philosophy may in no way interfere with the actual use of language; it can in the end only describe it. For it cannot give it any foundation either. It leaves everything as it is.[50]

---

[47] Wittgenstein (1958), §261.

[48] Caputo (1983, p. 675), while discussing Rorty's (1979) reading of Heidegger, also refers to 'the silence from which all language springs'. His meaning appears close to what is intended here.

[49] The idea of a post-reflective silence that 'leaves everything as it is' is reminiscent of certain motifs in the literature of Zen Buddhism, where a Zen master often responds to a student who remains on the reflective level (asking for instruction, say) by directly exhibiting the post-reflective condition, wherein the absence of philosophical difficulties is manifest in the commonplace activities of ordinary life (Aitken, 1991).

[50] Wittgenstein (1958), §124. Of course, when Wittgenstein says philosophy here he means philosophy 'done right'—that is to say philosophy as it is done in the *Philosophical*

The post-reflective stance is not an end to thinking. After all, it concerns only philosophy. Other forms of intellectual chatter go on as before. Moreover, the enlightened, post-reflective philosopher retains an important role, because she can guide the metaphysically afflicted towards the silence to which she herself always remains oriented. This role is especially relevant to the scientific study of consciousness. Like any nascent field, the scientific study of consciousness has to innovate to establish its methods and theoretical vocabulary. But there is an understandable tendency for its investigators to plunder everyday psychological language, and then to stray unwittingly into metaphysical territory they are ill equipped to traverse.[51] The challenge for the enlightened, post-reflective philosopher is to recognize when the exotic deployment of everyday folk-psychological terms such as 'thought' and 'consciousness' is a preliminary form of conceptual innovation, or an innocent rhetorical device to make difficult ideas accessible to a wider audience, and when it constitutes a dangerous foray into metaphysics.

The philosopher's job is made especially tricky because the ambitious scientist is always in quest of the deepest possible theory. Deep theories use abstract concepts and express overarching principles, and sometimes it's difficult to separate scientific theorizing at the highest level from metaphysical thinking. Nevertheless, these things are quite distinct. The theoretical musings of a circumspect researcher deal always in abstractions that are amenable to grounding in scientific practice and can thereby be legitimized empirically as well as rationally. In short, the careful scientist can always be brought back to the fundamental starting point. It is in this spirit that we shall now proceed to sift for a set of principles and abstractions to underpin a scientific understanding of cognition and its relation to our inner lives.

Investigations, not necessarily philosophy as it was (and is) typically done. Indeed, conventional philosophy is the one thing Wittgenstein does *not* want to leave as it is. There is a point of contact here with Heidegger, which is plainest to see with the aid of the commentary by Dreyfus (1991): '*Being and Time* seeks to show that everyday human activity … can disclose the world' (p. 58). Heidegger and Wittgenstein also agree, according to Drefyus, that 'the lack of an ultimate ground … is not an abyss. Counting on the shared agreement in our practices, we can do anything we want to do: understand the world, understand each other, have language, have families, have science, etc.' (p. 156).

[51] See Bennett & Hacker (2003).

Chapter 2

Cognition and embodiment

This chapter offers a characterization of cognition that assigns crucial importance to the fact of an animal's embodiment. Embodiment helps to explain what cognition is for and how it exercises its influence, as well as underpinning a scientifically respectable account of concepts, the building blocks of thought. A central role for cognition, in this view, is to enable the exploration, either on-line (through interaction with the world) or off-line (through internal operations), of an animal's space of affordances. The intellectual prowess of humans and other cognitive high achievers is reflected in the resulting ability not only to deal flexibly with novel situations, but also to open up whole new regions of affordance.

2.1 On having no body

When a person sits quietly in an armchair, stares blankly out of the window, and just thinks—about a loved one perhaps, or tomorrow's chores, or poetry, or philosophy—the body seems to play little part in what goes on. The heart beats, oxygen is drawn into the lungs, food is digested, and so on. But while these metabolic operations might be *empirically* necessary for thought—a sufficient blood supply to the brain being required to fuel its electrochemical activity—they hardly seem to be *logically* necessary. That is to say, we can imagine thought going on in their absence—in a computer perhaps, or a brain in a vat, or even some otherworldly spirit. Moreover, thought requires no immediate sensory input and does not immediately give rise to behaviour. It might involve the recollection of past experiences (of kissing, washing up, reading, and such like), and it might bring about various intentions (to kiss someone, to wash up, to read something, and so on), but at the time of the unfolding of a thought, the body seems to be, conceptually speaking, superfluous.

Spurred on by this line of reasoning, a philosopher of mind (herself comfortably seated in an armchair) might be motivated to produce an account of thought as a disembodied process, essentially a matter of computation, the

manipulation of symbolic representations according to syntactic rules.[1] Similarly enthused, a computer scientist might attempt to build a system that emulates human thought processes, which, when installed on a desktop computer, would be capable of convincingly human-like conversation, as well as being unbeatable at chess. Indeed, agendas of these sorts were pursued by a great many philosophers of mind, artificial intelligence researchers, and other cognitive scientists in the 20th century. But towards the end of the century, this style of research came increasingly under attack from those who believed that embodiment and cognition go hand-in-hand.[2]

Before enlarging on this theme, an important distinction needs to be drawn between thought and cognition.[3] Our immediate concern is thought in humans, and human thought, as the term is used here, is necessarily conscious. We can report what we are thinking and recall what we have thought in the past, albeit fallibly, and when our thought processes lead first to resolution and subsequently to action, we say that the action in question is deliberate, that we have exercised our will. By contrast, many attempts to characterize cognition from the standpoint of empirical psychology make no mention of consciousness at all. Cognition might (conventionally) be described as, say, a combined process of gathering information from the senses, storing it, processing it, and using what has been gathered, stored, and processed to guide behaviour. According to such a characterization, a cognitive process may be conscious, or it may not. Likewise, it's commonplace in philosophy of mind to gloss over the conscious/unconscious distinction altogether—to speak, for example, of a mental state without declaring whether the mental state in question is conscious or not. But throughout this book, we shall strive to keep this distinction to the fore.

In the next few sections, we'll review three lines of argument that purport to establish an intimate link between cognition and embodiment. These arguments are derived from the following three questions. First, what is cognition for? Second, how does cognition exert an influence? Third, what are the building blocks of thought? The first two questions, being couched in terms of 'cognition' rather than 'thought', are correspondingly neutral on the matter of consciousness. The third question, which concerns thought, by implication

[1] There are clear echoes of the *cogito* here. But the context is empirical cognitive science not philosophy. Cognitive scientists and AI researchers can (rightly or wrongly) downplay the importance of the body without adopting a metaphysical stance.

[2] Brooks (1991); Clark (1997).

[3] Contemporary scientific usage has strayed from the Latin root of the word 'cognition', which is *cognoscere*, to know (rather than *cogitare*, to think).

also concerns consciousness. The pro-embodiment lobby—and we shall count ourselves among their number—takes comfort from the fact that there are convincing answers to each of these questions that assign a central role to the body and its worldly interactions. What is cognition for? Crudely speaking, it improves an organism's ability to preserve, sustain, and reproduce itself. How does cognition exert an influence? In broad terms, it is incorporated into an organism's sensorimotor loop and thereby perturbs its behaviour. What are the building blocks of thought? The building blocks of thought are concepts, and all concepts are ultimately founded on the set of sensorimotor skills we exercise in our ordinary commerce with the physical and social environments.

2.2 **The biological roots of cognition**

Let's now examine the proposition that the purpose of cognition, as it is found in Nature, is to help an organism to sustain and preserve itself and to perpetuate its genes. Of course, there is a strict and important sense in which cognition, biologically realized, has *no* purpose. It is the product of an evolutionary process that has neither goal nor direction, and as such cognition takes its place alongside such marvels as the flowers of an orchid, the song of the whale, and the tail of the peacock, as well as such horrors as bowel cancer, deadly nightshade, and the jaws of the great white shark. So when we speak of the purpose of cognition, of what it is *for*, we are speaking elliptically of its role in determining the evolutionary fitness of an animal, and the teleological overtones of the phrase are to be ignored.

The notion of evolutionary fitness only makes sense relative to a particular *ecological niche*. It is no discredit to a pathogenic bacterium that its cognitive capacities are lacking, because within its microbial niche advantage is conferred by other attributes, such as the ability to migrate easily from one host to another and to multiply rapidly while under attack from the host's immune system. Moreover, fitness within a given niche is a complex business. Although straightforward attributes such as speed, size, or strength are often contributory, the matter is complicated by phenomena such as symbiosis and sexual selection, which permit peculiar forms of specialization within an ecosystem. Consider the way the long, curving beak of a single species of hummingbird can co-evolve with a uniquely shaped nectar-bearing flower, or how the complex song of the male nightingale has evolved concurrently with the discriminating powers of the female nightingale's auditory system.

Notwithstanding its relativity to ecological niche, the fitness of an organism (or a population), is a function of its ability to preserve and sustain itself and to perpetuate its genes. So when we say that human beings occupy an ecological

niche that favours cognitive prowess, we are claiming that cognition subserves these things. But in what way does cognition subserve these things? Our provisional assumption will be that cognition *helps an animal decide what to do when the possibilities afforded it by the environment are combinatorially structured*. It helps by *exploring* the space of affordances. This can be done either 'on-line'—through play, with the aid of training, and so on—or 'off-line'—that is to say by means of purely internal operations. The thoroughness with which an animal can explore a combinatorial space and reveal its hidden affordances is a measure of its cognitive prowess. Much of the rest of this chapter is devoted to making this formulation clear. What are the possibilities afforded an animal by its environment? What does it mean for this set of possibilities to be combinatorially structured, and how does exploring it help the animal?

The concept of an *affordance* was introduced by Gibson, for whom '[t]he affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or for ill'.[4] Crucially, what the environment affords an animal depends on the animal's capabilities. A drainpipe affords a rat a way up to the roof of a house, but all it affords a dog is a convenient place to leave a scent marker. For a human its affordances are abundant. The drainpipe might be detached and used as a makeshift didgeridoo, sawn in half and installed as a gutter, or made into a traditional game for the village fete. Moreover, what the environment affords is dependent not just on the physical capacities of an animal, but also on its psychological propensities, which in humans are a function of upbringing, education, and cultural background. For a Palaeolithic toolmaker, the affordances of a pile of flints are many and complex, whereas for a 21st century computer programmer, they are negligible.

Even for a given individual, there is an important distinction between what the environment *transparently* affords and its merely *potential* affordances. For example, consider the following task, which has been used to study tool-use in a variety of animals, including macaque monkeys.[5] A pellet of food is placed within sight of the monkey, but out of reach. However, within reach of the monkey is a small rake. With several days of training, the macaque can be taught to use the rake as a tool to pull the pellet close enough to be grasped. Now, there is a sense in which, even prior to training, the environment (including both pellet and rake) affords the monkey the means to satisfy its hunger, as

[4] Gibson (1979), p.127. Convincing neuroscientific support for Gibson's conception was obtained by Grèzes and Decety (2001), who showed that motor regions are activated by the perception of 'meaningful' objects (ones with familiar uses) but significantly less so by the perception of non-objects (meaningless blobs).

[5] Maravita & Iriki (2004); Povinelli (2000); Santos *et al.* (2006); Taylor, *et al.* (2009).

it is physically capable of obtaining the pellet (using the rake) and psychologically predisposed to do so. But this affordance is not apparent to the monkey. It is merely a potential affordance. After training, this potential affordance becomes transparent.

Of course, to count even as a potential affordance, a possibility has to be discoverable. Consider a burglar standing in front of a safe with a combination lock, and suppose he is ignorant of the combination. In principle, the environment affords the burglar the possibility of turning the dial to the right sequence of digits and opening the safe. But it is not possible, even in principle, for the burglar to work out the right sequence just by thinking hard enough. By contrast, the rake's potential as a tool for retrieving food pellets, though hidden in the combinatorial space of possibilities, is a discoverable affordance for an animal capable of grasping a rake. Cognition exposes such potential affordances – affordances that are discoverable but concealed in the combinatorial space of possible actions and outcomes—and makes them apparent. By making apparent potential affordances that would otherwise have remained hidden, cognition reveals to an animal opportunities and threats that it would otherwise have missed, and thereby increases its evolutionary fitness.

Now, what is meant by a *combinatorially structured* set of affordances? Mathematics and computer science often deal with combinatorial sets—sets of structures composed of parts that can be arranged in many different ways according to a systematic set of rules. For example, a simple set of grammatically correct sentences in English can be defined in the following way. A *sentence* is a noun phrase followed by a verb phrase. A *noun phrase* is an adjective followed by a proper name. A *verb phrase* is a verb followed by a noun phrase. The *adjectives* are 'pretty', 'funny', and 'silly'. The *proper names* are 'John', 'Mary', and 'Paul'. Finally, the *verbs* are 'loves' and 'hugs'. Examples of grammatical sentences according to this definition are 'Pretty Mary loves funny John' and 'Silly Paul hugs funny Mary'. With negligible effort it can be seen that the definition admits 162 possible sentences in total. This is a small set, but only eight components were required to define it. It's obvious that even with such a simple grammar, the number of possible sentences increases dramatically with the number of nouns, verbs, and adjectives. Rapid growth in the cardinality of a set when plotted against the size of its description is characteristic of combinatorial structure. Moreover, if recursive definitions are allowed, an infinite combinatorial set of sentences is straightforward to define.

Combinatorially structured sets are ubiquitous in mathematics and computer science. They include, for example, the set of sentences of first-order predicate calculus, and the set of syntactically correct computer programmes in the language C. According to Chomsky, the *productivity* of human language—that

is to say our theoretical capacity to generate and understand an infinite number of sentences—is a consequence of its underlying combinatorial structure.[6] In a similar vein, Fodor and Pylyshyn draw attention to the *systematicity* of human mental states. If a person is capable of entertaining the thought (or holding the belief) that, say, John loves Mary, then they should also be capable of entertaining the thought (or holding the belief) that Mary loves John. Likewise, the ability to draw inferences from the former proposition should go hand-in-hand with the ability to draw inferences from the latter proposition.[7] Fodor and Pylyshyn's general point is that a hallmark of human cognition is the ability to handle combinatorial structure, and the point is compelling whether or not we endorse the computational theory of mind to which they adhere.

The sense in which certain *mathematical* objects are combinatorially structured is precise and formal. The notion that such mathematical objects might serve as theoretical approximations to certain real-world phenomena, such as the utterances of a natural language, is unobjectionable. But the claim that there is combinatorial structure in what the environment affords an animal stands in need of clarification. (For now, we'll confine our clarificatory efforts to affordances in the physical sphere, as opposed to the social or cultural spheres.) In particular, combinatorial structure in mathematics is a feature of discrete domains—that is to say domains that comprise countable sets of distinct objects. Yet the real world of space, time, and matter is continuous, not discrete. So the question arises of how the continuous world presented to an animal's senses is to be conceptualized in discrete terms.

The world as it appears to an animal, what von Uexküll calls its *Umwelt*, is a product of its particular needs, concerns, and capabilities.[8] These vary from species to species, from individual to individual, and are subject to alteration throughout an animal's lifetime as it adapts and learns. An animal's Umwelt is reflected in what we might (cautiously) call the 'categorical scheme' superimposed on the physical world by its perceptual apparatus. Under each animal's categorical scheme, certain discrete objects and certain spatial relations among those objects stand out against the backdrop of the rest of the world. Where an urban human sees only a homogenous mass of foliage, a goat beside a hedge sees a mosaic of edible and unpalatable leaves. Similarly, under each animal's categorical scheme, certain discrete events and temporal relations among those events stand out from the ongoing flux. While a dog pricks up his ears at every rustle and crackle in a nearby flowerbed, barely noticing the babble of human

[6] Chomsky (1957).

[7] Fodor & Pylyshyn (1988).

[8] Von Uexküll (1957).

voices, the little girl on the swing beside him hears only the sound of her name followed by the words 'ice cream'.

Under a sufficiently rich categorical scheme, combinatorial structure emerges. If an animal's categorical scheme allows it to discriminate scenes in which object A is to the left of object B then, in general, it should allow the animal to discriminate scenes in which object B is to the left of object A. If its categorical scheme allows it to recognize that event C was followed by event D, then, in general, it should allow the animal to recognize that D was followed by C. The combinatorial structure the world discloses to suitably endowed animals is mirrored in the combinatorial structure of the set of options for action available to those animals. If an animal can place A to the left of B, then, in general, it can place B to the left of A. If it can perform action C before action D, then, in general, it can perform D before C. (We say 'in general' here because, of course, there are physical constraints and laws that forbid certain relations among objects and events while insisting upon others.)

So, for some animals, a combinatorial tree of possibilities perpetually branches out from the present situation into the future, according to what the environment potentially affords (Fig. 2.1). Because of its size and complexity, only a few branches of this tree can be anticipated by evolution, and the repertoire of behaviours an animal is born with reflects this limitation. Thanks to learning and adaptation, the animal moulds and expands this repertoire while it is alive. But not even a lifetime of experience can equip it with a tailored

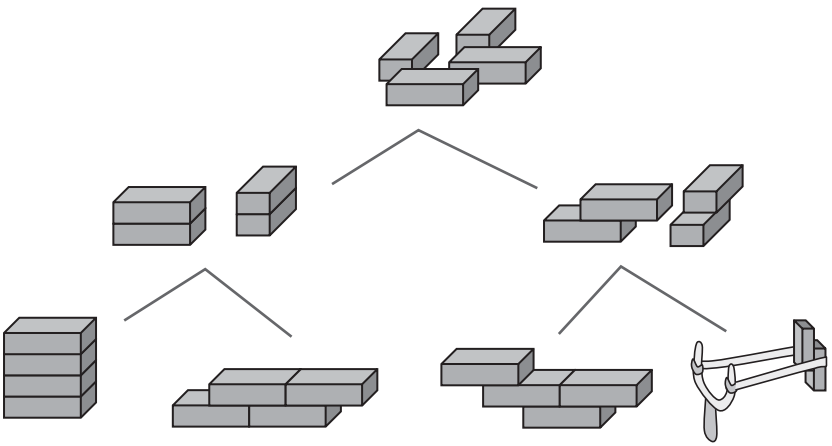

Fig. 2.1 A combinatorially structured space of possible affordances. The bricks can be assembled into structures of endless complexity by applying a small number of rules. But as the possibility of using an assembly of bricks as a missile illustrates, the space of possible affordances for a human being is open-ended.

response for every contingency. The tree of possibilities will always contain hidden surprises. So in order to take the fullest advantage of what the environment affords, it is profitable for the animal to explore this combinatorially structured space on the fly, making forays into it on a situation-by-situation basis, hopefully to bring out potential affordances that would otherwise have remained concealed. This is what cognition does, and this is how it benefits an animal and promotes the perpetuation of its genes. Or to place the emphasis differently, this is what makes the cognitively endowed animal a viable phenotype.

2.3 **The sensorimotor loop**

With the role that cognition plays in boosting the evolutionary fitness of an animal firmly established, we can probe more deeply the second in our series of three questions. How does cognition exert an influence? The proposal under the spotlight is that the only way for cognition to exert an influence and benefit an animal (or the population to which it belongs) is by perturbing and contributing to the dynamics of that animal's sensorimotor loop. To bring out the implications of this proposition we will assess the extent to which cognition is still comprehensible when sensorimotor considerations are downplayed. To this end, it's instructive to look not only at natural cognition but also at attempts by artificial intelligence (AI) researchers to replicate human cognitive skills with computers and robots.

In the 1960s, 1970s, and 1980s many AI systems were developed whose core was a process operating along the following lines. Having received as input a symbolic description of some problem, the process would embark on a lengthy computation without further input from the user or the world, finally presenting the symbolic description of a proposed solution as its output.[9] Such a process can straightforwardly be incorporated into an interactive system, such as a chess-playing programme where turns have to be taken with an opponent. This is done simply by embedding it into a loop that repeatedly obtains input (the opponent's move), carries out the required computation (searches for a good move of its own), then presents the corresponding output (the move it has found), maintaining between iterations whatever record of internal state it requires (such as the board position). Robots too can be programmed this way. Such a robot repeatedly senses its environment, plans its next action, then executes that action, while maintaining as accurate a model of the world as possible. The core process here—the part with a cognitive flavour—is the planner.[10]

[9] This research programme was inaugurated by McCarthy (1959a), whose fictional 'advice-taker' programme may be considered the forerunner of all such systems.

[10] The prototype for all such robots was Shakey (Nilsson, 1984).

All such work proceeds under the assumption that cognition can be treated as a black box that is given an input of finite length, performs a computation that terminates in finite time, and then produces an output of finite length. By way of contrast consider homeostasis, the closed-loop control system that maintains the body's internal milieu. If the water level in the blood is too high, for example, this is detected in the brain by the hypothalamus, which instructs the pituitary gland to decrease production of a certain hormone. This in turn causes the kidneys to absorb more water from the bloodstream producing more urine, which is why pubs have busy toilets. Conversely, if the water level is too low, the hypothalamus tells the pituitary gland to secrete more of the same hormone, which inhibits absorption of water by the kidneys. Under normal circumstances and in the absence of pathology, this maintains the water level in the blood within just the right range for effective metabolism.

When these homeostatic processes terminate, it is not because they have computed some final output. On the contrary, it is because the parameters of the body's internal milieu have strayed outside tolerable bounds or because the body's integrity has been catastrophically violated, that is to say when the organism dies. A normally operating homeostatic process is non-terminating. If we had to specify its correct operation, we would not be able to appeal to the functional relationship between the finite input and the finite output of a finite computation, even if this computation were embedded in a loop. Rather, we would employ the language of control theory to describe the system's ability to bring the required parameters back to within certain bounds in a given time following a sufficiently bounded perturbation.

This does not entail that such a system is not amenable to description in computational terms. Although early mathematical treatments of computation, drawing on the ideas of Turing and Church, were confined to terminating processes, modern computer science has a wealth of theoretical tools for describing interaction, concurrency, and non-terminating processes.[11] This is not the point. Neither does it entail that conventional computation cannot be used to implement a closed-loop control system. Indeed controllers built out of microprocessors are ubiquitous in modern technology. This is not the point either. The point, rather, is that we cannot grasp the character of such a process unless we see it as part of a sensorimotor loop dynamically coupled to the external environment, and we cannot make sense of its behaviour unless we examine it over an extended period of time.

Homeostasis is best conceived as a perpetual process. But the point also applies to more goal-directed forms of closed-loop control, processes that do

[11] The seminal work is that of Milner (1980) and Hoare (1985).

terminate, usually with certain conditions fulfilled. As an example, consider the robot model of cricket phonotaxis—the female cricket's ability to locate a male cricket by walking in the direction of its mating call—developed by Webb and her colleagues.[12] Unless the male happens to be directly ahead of the female, its call will arrive at the female's left and right ears at slightly different times. In Webb's model, the signals from each ear propagate via neurons in the cricket's auditory system to their counterparts on one or other side of its motor system. But one signal, left or right, will arrive at its destination slightly sooner than the other, and the neurons are so arranged that the winner of this race inhibits the effect of its rival. So the legs on one side of the cricket's body will be activated more vigorously than those on the other, causing the cricket to orient itself towards its potential lover.

The role of feedback here is to facilitate repeated corrections to the system's trajectory to compensate for sensor noise, motor play, and unexpected environmental perturbations. Robustness in the presence of these uncertainties is essential and unachievable without constant sensory input. Again, such a process can only be understood in the context of a sensorimotor loop operating over an extended period within an environment. Isolated from the environment and from the feedback the environment supplies, the process is incomprehensible. Likewise, if the behaviour of the process is observed over too short a period of time, its essential self-correcting character is indiscernible.

The contention here is that these considerations also apply to cognition. Cognition is best viewed not, as early AI researchers thought, in terms of a problem-solving module that, mathematically speaking, computes a finite output sentence from a finite input sentence and then terminates. Rather, it should be understood as part of a feedback control process, extended in time and dynamically coupled with the environment, a process that continuously adjusts its output according to its input in order to maintain an animal's well-being and fulfil its needs and desires. But many different feedback-based control loops contribute to an animal's behaviour, and not all of these deserve to be seen in a cognitive light. Indeed, our contention was that cognition comes into its own for an animal confronted with a combinatorially structured set of affordances. The advantage it confers is to reveal opportunities and threats that, to the cognitively less well-endowed, remain hidden in the space of combinatorial possibilities. The question is how this is done in the context of a feedback control loop.

[12] Reeve & Webb (2002). For details of the likely mechanisms at work in real crickets see Hedwig (2006).

2.4 **Behaviour selection**

To answer this question, we'll make use of a general framework for characterizing behaviour that originated in the field of ethology and was subsequently elaborated by roboticists.[13] (Roboticists influenced by ethology typically come from a school of thinking opposed to the way cognition is conceived by classical AI.) According to this framework, the overall behaviour of an animal (or robot) is the outcome of a perpetual process of selection from a repertoire of more simple behaviours (Fig. 2.2). Each behaviour in this repertoire is a response to some or other cue in the animal's internal or external environment. Each behaviour is to a large extent pre-programmed and fixed, and each involves the execution of a number of component actions. To illustrate the idea, let's think about chickens (free-range, of course).

A chicken's natural repertoire of behaviours includes foraging, feeding, laying, fleeing, dustbathing, roosting, and so on. Consider feeding. This behaviour is induced by the combination of an internal deficit (hunger) and an external cue (the appearance of a farmer at a gate, say, which is associated with the arrival of food). The behaviour itself can be divided into two phases—approach and consumption. The approach phase involves its own miniature control loop, wherein the chicken uses visual feedback to guide it towards the farmer. This is terminated when the food is attained, and is followed by the consumption phase. The consumption phase involves repetitive pecking at

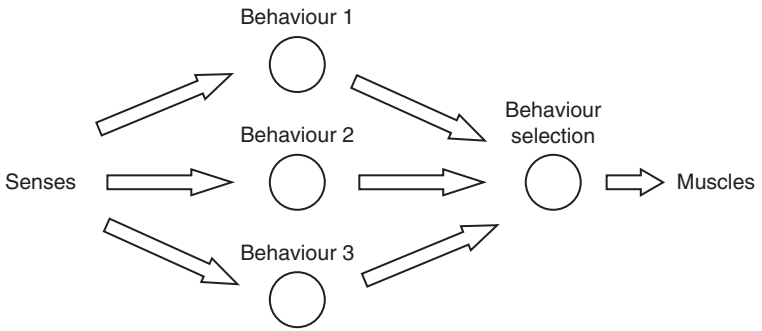

Fig. 2.2 A behaviour-based architecture. Separate processes, each one spanning sensation and action, are responsible for distinct behaviours. Each behaviour is triggered by a combination of external cues and internal deficits. When more than one behaviour is triggered, the competition is resolved by a selection mechanism.

[13] McFarland & Bösser (1993).

grains, and is terminated by satiation or when the food runs out. Each of the behaviours in the chicken's repertoire might be similarly specified.

A day in the life of a chicken can be characterized as a patchwork of distinct episodes of behaviour, each one taken from this repertoire. In general, more than one of the behaviours in its repertoire is applicable at any given moment, so the chicken is faced with the continual problem of choosing between them. This is known as the *behaviour selection problem* (or *action selection problem*), and a variety of neural mechanisms for tackling it have been hypothesized.[14] Certain principles that any proposed mechanism should respect are obvious. In general, the competition between behaviours will be settled according to some function of the strengths of external cues and the levels of internal deficit. Nevertheless, it must be possible for some behaviours to take priority over others when the occasion demands. If a potential predator appears, even a hungry chicken must abandon its food and flee. Furthermore, effective switching between behaviours requires hysteresis. If a chicken is both thirsty and hungry, it is not a good policy to drink a little, run to the feeding trough, eat a little, run to the water dispenser, drink a little, and so on. Rather, there should be a degree of commitment to a behaviour once it has been initiated.

According to this conception, the overall behaviour of an animal, on a timescale of hours or days, is seen to be regulated by a large feedback control loop in much the same way that its behaviour on a timescale of seconds is governed by various smaller, tighter feedback control loops. The small control loops, such as those involved in locomotion towards a target or reaching for an object, make rapid adjustments to the current motor output in reaction to immediate sensory input, on a millisecond-by-millisecond basis. By contrast, the large feedback loop makes infrequent adjustments, and several tens of seconds might elapse between one and the next. Rather than modifying immediate motor output, these adjustments take the form of switchings between behaviours in response to relatively major events in the animal's life involving sensory cues and metabolic thresholds. The overall effect is to maintain the integrity and well-being of the animal and to fulfil its needs in the face of constant change within both its external environment and its internal milieu.

Almost every behaviour in an animal's repertoire can plausibly be cast as a response to the perception *of* something that affords something to the animal. However, it's not always appropriate to describe a behaviour as a response to

[14] Redgrave, *et al.* (1999); Cisek (2007); Houk, *et al.* (2007); Humphries, *et al.* (2007). Cisek (2007) appeals to the concept of affordance in a way that is compatible with the present account.

the perception *that* the environment affords it something. The latter wording implies a level of sophistication not implied by the former. A patch of dusty ground affords a chicken the possibility of cleansing its plumage, and this, in evolutionary terms, is why the perception of a patch of dusty ground can trigger dustbathing.[15] But there is no need here for the chicken's brain to discriminate between means (dustbathing) and end (improvement in plumage condition), because the unfortunate chicken has only one means to that particular end in its behavioural repertoire. So there is no useful sense in which the chicken can be said to perceive *that* the patch of dusty ground affords an opportunity for plumage cleansing. It simply dustbathes when it sees dust.

In this respect dustbathing contrasts with feeding. A full grain-trough affords a chicken relief from hunger, and this is why, thanks to what the chicken has learned, the sound of pouring grain triggers a dash for the trough. But the trough is not the only way the environment makes food available to the chicken. There are multiple means to the same end, and to exploit this, the chicken's behaviour has to be sensitive to the difference between means and end, between approaching the grain-trough and the relief of hunger. (A minimal sign of such sensitivity in an animal is the capacity to extend its behavioural repertoire with new ways to achieve a desirable outcome, or flexibly to adapt its old ways.) In this sense, the chicken can be said to perceive *that* the trough affords food, whereas it does not perceive *that* dust affords cleaner plumage. It's only this stricter sense of affordance that's relevant here.

Now, recall our provisional characterization of cognition, according to which it helps an animal decide what to do when faced with a combinatorially structured set of affordances. We are at last in a position to see how this role can be accommodated within the animal's sensorimotor loop. Simply put, cognition perpetually updates the repertoire of behaviours an animal can choose from with potential affordances that were previously concealed in the space of possibilities. So as not to interfere with the animal's ability to respond in a timely fashion to ongoing events in a fast-changing world, it must operate in parallel with other, more primitive mechanisms. If, thanks to whatever internal dynamics it is founded on, cognition discovers a behaviour (an action or a series of actions) with a salient expected outcome (desirable or undesirable), and does so in time for it to be taken into consideration by the behaviour selection process, then so much the better. If it does not, no harm is done, and the animal behaves as it would have done anyway.

[15] Olsson & Keeling (2005).

2.5 **The human edge**

We have attempted to pin down what cognition is for in evolutionary terms, and to account for its influence in behavioural terms. Our treatment has ranged over a spectrum of cognitive abilities, including those of non-human animals. To address the third of our questions about cognition—the question of the building blocks of thought—we'll have to focus more narrowly on the human case. But first we shall pause to consider what, if anything, is special about the human animal, by examining a series of examples of behaviour displaying increasing levels of cognitive prowess, culminating with a showcase of human ingenuity.

We'll begin with avian nest-building, which seems cognitively demanding on first acquaintance, but turns out to be only modestly so. Let's consider a particularly gifted species, namely the village weaver (*Ploceus cucullatus*).[16] Its nest-building procedure involves several component behaviours, including the attachment of initial material to a branch, the selection and carriage of plant matter for inclusion in the walls, and the weaving of individual threads into the fabric of the nest. Having established an initial ring of woven grass suspended from a branch, the male village weaver builds up the walls of its nest chamber by repeatedly threading strands of grass through the structure until a near-spherical shape is achieved. The weaving of even a single strip of grass into the nest is a complex operation. The bird must press the end into the nest wall with its beak, release it, then grasp the end again from a slightly different location and pull it back out, as if it were using a needle and thread, then repeat this process, moving in a circle inside the partially built nest, until the whole strand is woven in. The exquisite structure that results might easily be taken for the product of intentionally thought-out design, planned and executed, it would seem, with considerable help from cognition. However, as Hansell points out, such a nest can be woven 'using a fairly limited repertoire of stereotyped movements'.[17]

Now let's consider an Olduwan flint-knapper from the Lower Palaeolithic (early stone age) period (beginning approximately 2.5 million years ago).[18] An Olduwan blade is manufactured by repeatedly striking a hammer stone held in one hand against a flint core held in the other. The process only takes a few minutes. Every successful strike causes a flake of material to be sheared from the core, and by this means the knapper gradually sculpts the flint into a desired overall shape, then trims and sharpens its edges. Each separate strike requires the knapper to select a suitable site on the flint surface, which involves a visual

[16] Hansell (2000), Chapter 4.

[17] Hansell (2000), p. 84. See also Hansell (2007).

[18] Wynn (2002); Stout & Chaminade (2007).

examination while turning the core. This is followed by the application of the hammer stone to that site with such a force and at such an angle that a flake of material is removed and the flint core is brought closer to a useful shape. The final tool might easily be taken for the product of deliberate and careful design. But there is little evidence that this is the case.[19] Again, a limited repertoire of stereotypical actions is sufficient for the task. Despite their flashy performances, neither the nest-builder nor the early stone-age flint-knapper is an obvious cognitive high-flier.

Moving the dial 2 million years forwards to the Upper Palaeolithic (late stone age), we encounter an altogether more sophisticated knapper. An Upper Palaeolithic handaxe has a comparatively regular, symmetrical shape and a carefully finished edge. Its production requires the removal of a large number of flakes of different sizes, and the manufacturing process has several distinct stages. A compelling case can be made that to sculpt these more refined tools requires a degree of spatial cognition, goal-directed action, and planning.[20] Furthermore, the late stone-age tool-maker sits at the cusp of a dramatic expansion of human culture and technology. We would like to understand the cognitive foundations for this. And arguably, what sets this modern human tool-maker apart from the avian nest-builder is not discernible in any established behaviour, even one that incorporates goal-directed actions and planning. Rather, it is the capacity to innovate, to invent new forms of behaviour.[21] To illustrate this, let's advance the dial another 50,000 years or so and, instead of an Upper Palaeolithic flint-knapper, consider a 21st century city-dweller preparing a meal in her kitchen, a figure we can more easily relate to.

Suppose our cook needs to add tomatoes to a saucepan. She has a tin of tomatoes, but the tin-opener seems to have gone missing. Perhaps it has been incorporated into one of the children's games. She conducts a search of the playroom, but fails to find it there. Well, one of the neighbours is sure to have a tin-opener she can borrow. She rings the doorbells of two neighbours, but no one is home. So she resolves to open the tin by other means. (The dish would be a failure without tomatoes.) She retrieves her toolbox, which contains a common assortment of tools. First she pierces the top of the tin using an awl and a hammer. But the resulting hole is tiny. So she enlarges it with a screwdriver, again using the hammer to apply enough force to penetrate the tin. Now the fissure is large enough to take the head of a small pair of pliers. Using the pliers she prises up the lid, widening the aperture until she seems to

[19] Stout & Chaminade (2007).

[20] Wynn (2002).

[21] Mithen (1996); Coolidge & Wynn (2009).

have gained access to the tin's contents. However, the opening is still quite small, and when she up-ends the tin over the saucepan the tomatoes remain lodged inside. So she finds a tea-spoon and, holding it the wrong way round, pokes it into the tin. Using the handle of the tea-spoon to mash the tomatoes she eventually breaks them into pieces small enough to fall through the opening she has made and into the saucepan.

This is a most impressive performance, in which formidable cognitive powers are on display. There is clearly no explanation for it in terms of a small repertoire of stereotyped actions. Rather, the cook has carried out several narrowly focused explorations of the combinatorial structure of possibilities for action afforded by her environment—partly through experiment and partly through internal operations whose character we need not specify for now. If the lost tin-opener were a familiar situation, and if our subject had previously resorted to household tools to open tins, then it would be a less remarkable exhibition. In that case, exploration would not be required because the relevant unorthodox affordances of the household toolbox would not be hidden in the space of combinatorial possibilities. They would be transparent. But we are assuming this is not the case.

To explore the space of affordances would be infeasible without a degree of mastery of the causal relations that structure that space. Thanks to this mastery, innovative possibilities are properly entertained, yet resources are not wasted on the plain ridiculous. The awl is worth considering as a means of access to the tin because it can make a hole in a solid object, the sort of aperture through which a non-rigid body of matter might pass. But the possibility of, say, phoning the police to ask where the tin-opener might be, or of using a sponge to pierce the tin, is not even worth considering. In traditional epistemological terms, the cook's mastery of the causal structure of her environment's affordances might perhaps be said to include both procedural and declarative knowledge, both the knowledge *how* to use an awl to make a hole and the knowledge *that* an awl can be used to make a hole. But the umbrella term 'mastery' is better here because it allows us to sidestep epistemological discussions.

Human-level mastery of the causal relations that structure the space of affordances seems to come with fewer 'blindspots' than we find in other animals. Recall the task of retrieving food with a rake. Chimpanzees easily learn to use a rake to obtain food that is out of reach. But using an extension of this paradigm, Povinelli provided evidence of the limitations of a chimpanzee's understanding of the physics of everyday objects.[22] In Povinelli's modified

[22] Povinelli (2000); Penn & Povinelli (2007). However, it has been shown that chimpanzees and gorillas can solve equivalent problems when the experimental set-up is modified

experiment, the chimpanzee was presented with a table divided into two parallel sections, each containing a rake and a food item (a peanut). However, one of the sections included a shallow but visible trench between the peanut and the chimpanzee that was guaranteed to intercept and trap the nut before it could come within reach. In this and a number of variations on the same experiment, the chimpanzees demonstrated no overall preference for one rake over the other, suggesting a certain inability to anticipate the inevitable and unwanted consequence of drawing a peanut towards a trench.

How might an animal possess true mastery of the causal relations that structure the space of its potential affordances, a level of mastery that is not unduly marred by such blindspots? According to the hypothesis we shall pursue, this requires a generic facility for integrating distinct domains of expertise. On a coarse scale, these domains span the physical, psychological, and social aspects of the animal's Umwelt. On a finer scale, we find sub-domains within these broad regions of expertise. Within the domain of everyday physics, for example, an animal may or may not have expertise in the effects of gravity, the qualities of different materials, or the kinematic properties of various shapes. These sub-domains are ultimately grounded in sensorimotor skills, which can be thought of as micro-domains of proficiency, such as putting-things-in-holes, or poking-things-with-sticks. The sort of generic facility in question ensures integration across all of these, and would allow an animal to explore a region of the tree of affordances that included *both* pulling-with-a-stick actions *and* falling-in-a-hole events.

So, getting back to the Upper Palaeolithic tool-maker, established skills such as flint-knapping are perhaps less cognitively significant than the capacity to draw on and integrate multiple distinct regions of understanding, blending their elements together. Arguably this quality, which Mithen calls *cognitive fluidity*, was crucial in enabling our early ancestors to adapt to change, and to progress.[23] When stone tools ceased to be viable or competitive, they were able to develop new technologies and to adopt different ways of life. Likewise, this sort of integrative capacity enabled our cook to cope with the novelty of a lost tin-opener, and to bring her combined expertise in the psychological, social, and physical domains (the children, the neighbours, and the toolbox) to bear on the problem. Chapter 4 will propose a substrate for a generic integrative facility of this kind, namely a *global workspace architecture*, the heart of which is a broadcast mechanism that allows localized processes to exert global

(Martin-Ordas, *et al.*, 2008). Rooks and crows are not only able to solve such problems, but can also transfer their acquired expertise to different tasks (Seed, *et al.*, 2006; Taylor, *et al.*, 2009).

[23] Mithen (1996).

influence on the brain. Not only is the integrative provision of this architecture hypothesized to underpin the conscious condition, it is also proposed as the means by which *conceptual blends* are effected, enabling concepts of increasing abstraction to be layered one upon another.

2.6 **Founding concepts**

Thought, which constitutes a large portion of a person's inner life, is here conceived of as serially unfolding conscious cognition.[24] What then are the building blocks of thought? Concepts are the building blocks of thought. But what are concepts, and how are they acquired? Our special concern for the time being is with abstract concepts, such as those of infinity or electrical charge, things that are outside the purview of non-human animals. Although the treatment on offer here allows for the ascription of certain kinds of concept to non-human animals,[25] the issue at hand is that the quintessentially human facility with abstraction suggests that cognition might not, after all, be confined to the earthly subject matter of biological imperatives, that the conceptual reach of humans may extend to loftier realms.

But in the post-reflective condition, we are no longer tempted by philosophical doctrines that conjure up metaphysical divisions, and we are disinclined to locate concepts anywhere other than right here. Metaphorically speaking, their home is neither 'further in' than the everyday world we all share nor 'further out'—neither inside our heads nor in some non-physical reality. So it's not necessary to adopt a metaphysical stance to tackle the question of how it is that we come to do the things we do and think the things we think, and to comprehend the role concepts play in all this. These are questions about our shared world—a world of which possessors of concepts form a part and in which they wholly dwell. Such questions can be tackled by a worldly investigation, an empirical investigation.

Moreover, the attitude towards meaning that goes with our post-reflective stance encourages us not to ask the metaphysically loaded question of what a concept *is*, but rather to investigate how words like 'concept' are used. Indeed, when we ordinarily say that someone 'possesses a concept' (no doubt using a more natural turn of phrase) we do so for a purpose, as part of our daily lives, perhaps to help a third party to understand that person, and so to interact with her better or more reliably to anticipate her next move. Sometimes, perhaps,

[24] There are other facets to a person's inner life, of course. Not every image, memory, or fantasy is usefully construed in terms of the exploration of a combinatorially structured space of possible affordances.

[25] Such ascriptions are controversial, as we shall see.

what we say is best thought of as elucidating a person's inner life. Sometimes what we say is best thought of as elucidating her outward behaviour. Either way, the word 'concept' is just a useful tool. The concept of a concept, we might usefully say, is just a useful tool.

Consolidating everyday usage, a fruitful way to prosecute an empirical investigation of how we do the things we do and think the things we think is to assume that to acquire a concept is to master a systematic set of skills, and that to acquire an abstract concept is to layer a systematic set of thinking and talking skills on top of a foundation of more basic sensorimotor skills. To think, then, is to exercise these skills, not to sojourn in a mythical plane of pure reason. So even when we say—with perfect propriety—that a thought (or a belief) is *about* the prime numbers, for example, or *about* Ohm's law, there is a sense in which that thought (or belief) remains securely *founded* in the everyday world.

We shall further explore this idea shortly. But first, we must escape the charge of trivially equating truth with consensus.[26] Specifically, we want to sanction the difference between the proper grasp of a concept and a faulty one, between the correct application of a concept and its misapplication. After all, if we allow rampant relativism to overwhelm us there will be no distinguishing these things and, as a consequence, no distinguishing truth from falsehood. Epistemological mayhem will ensue (and perhaps civilized society will collapse). Very well. But nothing has been said that would undermine these important notions. It goes without saying that we sometimes accuse a person of not having understood a concept right. If someone has misunderstood a concept—that of a prime number, say—then he is liable to be put right when he claims that 15 is prime, and is owed an explanation of why he is wrong. Likewise, if a student fails to grasp what is meant by, say, capacitance, then of course we will correct her when she plots a distorted graph of current against time. To slip up in any of these ways is not yet to have mastered the relevant skill.

But who is to say it is not the maverick who has got the concept right, and the majority who are in the wrong? Well, insofar as they share a common foundation for making judgements, the maverick should be no less able to persuade the majority than the other way around. It is the nature of a language game to have methods and conventions for settling disputes—methods and conventions that sometimes involve consulting the world, as in a scientific experiment, and sometimes involve following carefully formalized procedures, as in mathematics—and the majority is no less subject to their rule than the individual. Sometimes in difficult cases, when the dust settles, new ways of talking

[26] See Wittgenstein (1958), §§241–242, as well as the discussion in Chapter 1.

and new ways of resolving disputes will have found their way into a language game. But these shifts in and adjustments to our common foundations are also part of the cut and thrust of ordinary life.

But surely our concepts must measure up against reality. Does the post-reflective stance entail a denial that, say, mathematical ideas have any reality independent of the human mind? No, not that either. Questions of the reality, or otherwise, of the concepts of mathematics are the residue of an abandoned mode of thought. We are interested, as scientists, in the processes and mechanisms by which abstract concepts are acquired and exercised. In consideration of these mechanisms, although it serves us to continue to speak, in the natural idiom, of a certain thought as being *about* the integers or *about* the laws of physics, it serves us equally to speak of those concepts as being *founded* on our sensorimotor interaction with the environment. If pressed on the question of what such thoughts are *really* about, metaphysically speaking, we decline politely (unless we have the talent to offer therapy, in the style of Wittgenstein, and our inquisitor is minded to receive it). Yet in keeping counsel, the truly post-reflective thinker has neither ducked her intellectual responsibilities nor gone into denial over her innermost philosophical urges. In the proper silence of first philosophy, nothing is left unsaid. There are no secrets whose revelation is left pending.

2.7 **Counting and infinity**

To get a flavour of the sort of empirical account we should hope for, let's consider how a system of mathematical concepts might be layered onto a foundation of sensorimotor skills. The sketch here will be very brief. But it draws on the work of Lakoff and Núñez, who offer a thorough, book-length treatment along similar lines.[27] First, how might a child acquire the concept of a positive integer, which is surely at the lowest level of mathematical abstraction? Here's a speculative description of a possible developmental path. It begins with counting. The child is taught to say the numbers from one to ten out loud. This might be done by teaching her to recite a nursery rhyme in which the numbers occur in order, such as 'One, two, three, four, five, once I caught a fish alive. Six, seven, eight, nine, ten, then I let it go again'. From here it's a small step to being able to recite the numbers on their own.

[27] Lakoff & Núñez (2000). See also Dehaene (1997). The present endorsement of Lakoff & Núñez (2000) does not extend to their philosophy of mathematics, however. By denying that 'mathematics exists outside us' (Chapter 15), they accept the terms of a metaphysical debate that we are here concerned to avoid altogether.

Building on this ability, the child is next taught to point to each member of a group of objects in turn while reciting the series of numbers she has learned, matching the rhythm of the finger movements to the rhythm of the chant. There are some subtleties here, but a parent or teacher will be on hand for guidance. In particular, the child must learn not to point to any object more than once, and not to omit any object in the group. To begin with, the objects can be arranged in a line within her field of view, to make it easier to conform to these constraints. With words of encouragement or gentle admonishment as appropriate, she soon learns the proper procedure. With this skill in place, the child is in a position to answer questions about quantity. How many biscuits are on the plate? She points and counts and offers the final word in the recited sequence as her answer. A useful skill has been acquired. It helps to ensure fairness in the distribution of chocolate among siblings.

By itself, though, this trick does not qualify as a *systematic set* of skills. Like the ape who has learned to use a rake to obtain food, but is fooled by a simple trap, the child may not yet be adept at variations of the basic counting scenario—counting by taking objects one at a time from a container, for example—or be able to generalize her skill to other kinds of problems. Moreover, at this stage she lacks a whole cluster of related skills, such as the ability to add and subtract numbers, without which any ascription to her of the concept of an integer would have to be qualified. Thankfully, ordinary discourse is more nuanced than the examples typically found in the philosophical literature. If asked by a concerned parent whether a child has 'acquired the concept of an integer', a primary school teacher—after a moment's pause at the odd phrasing of the question—might report that the child can count, can compare numbers, and can add and subtract, but has yet to understand that adding X to Y always yields the same answer as adding Y to X.

With the aid of more props and the scaffolding talents of a good teacher, the child's repertoire of arithmetic skills is soon enlarged and consolidated to the point where she is able to perform mental arithmetic, and to apply her understanding to novel classroom problems. This requires the ability to map from patterns of concrete sensorimotor activity to patterns of abstract thought, which Lakoff and Núñez relate to the human capacity for metaphor making. A thorough account of this or a similar mechanism would be pertinent. But such details do not concern us right now. Our only concern at present is to show that the human facility with abstract concepts does not lend credibility to the idea that the mind can 'float free' of the body. To press the point, let's extend our attention to a more abstract concept, namely infinity. How might the concept of infinity be founded on a sensorimotor substrate?

The concept of infinity does seem to be strictly the provenance of humans. Some non-human animals are able to determine the number of objects in a

group and can be trained to offer answers to questions about quantity.[28] But no animal can demonstrate its understanding by pointing to or picking up an infinity. Infinity is something that can only be thought about or talked about. So to have mastered the concept of infinity is to possess a thinking and talking skill rather than a manual or physical skill. It entails the ability to discuss such issues as how many integers there are or how many fractions lie between zero and one. How might such a thinking and talking skill be acquired? Consider the following fragment of dialogue, in which an imaginary pupil is introduced to the concept of infinity by his teacher.

TEACHER: What's the biggest number you can think of?

PUPIL: A million. No! A billion.

TEACHER: Can you think of a number bigger than a billion?

PUPIL: A billion times a billion!

TEACHER: But if you add one to that you'll get an even bigger number.

PUPIL: Yes …

TEACHER: In fact, however big a number you think of, you can always add one to it. Isn't that right? And that would give you an even bigger number.

PUPIL: You could keep on adding one forever, and never stop.

TEACHER: Yes! You can add an *infinite* number of ones.

It is not the aim here to offer a plausible piece of developmental psychology, but rather to highlight that there's no need to step outside the ordinary world of everyday physical activity to answer questions about the source and character of abstract thought.[29] In this instance, the teacher presents the concept of infinity through the idea of a process that has no end. It's always possible to add one to an integer, however large that integer is. So this process of incrementation can continue forever. The teacher's instruction builds on her pupil's ability to count, a sensorimotor skill that is already in place. The idea of counting forever is new to him, but this too has a sensorimotor foundation. For sure, many sensorimotor operations associated with counting are bounded by terminating conditions. The process of removing marbles one by one from a bag will come to an end when the bag is empty. But other operations, such as tapping a finger on a desk, need never come to an end.

Of course, the pupil cannot in fact tap on the desk forever. To grasp such an idea requires the capability, not to carry out an endless sequence of actions, but to *imagine* the perpetual exercise of the relevant sensorimotor skill. A process that, in the imagination, never ends is just one that, in the imagination, can always be continued, which is the case if it consists of actions whose execution

[28] Dehaene (1997), Chapter 1.

[29] Monaghan (2001).

brings about their own preconditions. After a finger has been lifted and brought back down onto the desk, it is in just the right position to be lifted again. Now, each finger-tap on a desk is much like the last. But in counting a set of objects, every touch of the finger is accompanied by the pronouncement of a unique name, the next number in the series. So to imagine the business of counting going on forever requires the ability to imagine a new, nameable thing, without having to actually name it. Plenty of metaphorical props, building on the ability inwardly to rehearse sequences of actions in the real world, are available to the teacher to nudge a child towards such an understanding. For example, he might be told to imagine a series of fence-posts that goes on forever, and to envisage painting them one by one. Because each fence-post is distinct from all those that precede it, he will always arrive at a fence-post that is unpainted.

A good deal more will be said about the workings of the imagination in Chapter 6. For now it suffices to give notice that the imagination will be understood in terms of a mechanism for internally rehearsing trajectories through sensorimotor space, that is to say for simulating interaction with the environment. Because talking is no less a form of sensorimotor activity than walking, whistling a tune, or cooking a meal, the imagination so conceived is also the locus of inner speech, and indeed of thinking generally. The interplay, governed by learned associations, between various different sorts of internally simulated sensorimotor activity—counting, painting fence-posts, talking about numbers—is what allows the rehearsal of a series of actions, and the resulting sense that this series is always open to continuation, to issue in verbal pronouncements about infinity. The teacher's instruction puts in place a new matrix of associations between certain patterns of perpetually iterable action and certain patterns of talking, and the upshot is the augmentation of the pupil's repertoire of thinking and talking skills with the ability to think and talk about infinite processes.

Before abandoning the infinite, some parting remarks are in order on the phenomenology of grappling with abstraction. The whole thrust of the foregoing remarks is against the tendency to speak as if, in conscious thought, we can become temporary travellers in some ethereal plane of mathematics or logic. Our aim has been to dissolve the image of disembodied consciousness at play in the realm of pure ideas. But suppose that someone in the grip of such a conception were to challenge the view that abstract concepts have a sensorimotor foundation by pointing out that it doesn't *feel* that way.[30] When we think a mathematical thought, sensorimotor interaction doesn't feature in the phenomenology.

[30] Penrose trades heavily on such intuitions in *The Emperor's New Mind* (1999).

Well, introspective verbal report is a primary source of data in the scientific study of consciousness, and it is not to be dismissed. It is most useful when it correlates with outward events, such as the presentation of a visual stimulus or a pharmacological intervention. But the notion that the conscious subject has a privileged viewpoint on the origin and character of the concepts with which it thinks is vulnerable to the critical reflections on privacy and authority that led to the post-reflective condition. There should be no need to revisit the relevant critical material here. Suffice to say, such a notion appeals to the metaphysically fissile inner/outer distinction that was disabled by the private language remarks, and it resurrects the suspect idea of a solitary inner world wherein meaning is underwritten.

2.8 The space of possible minds

To recap, we have examined what cognition is for, how it exerts an influence, and what the building blocks of thought are, and in each case we unearthed a significant link between cognition and embodiment. A question we might now be tempted to ask is whether any of these links amounts to a *necessary* condition for cognition, or whether the connection with embodiment is merely empirical and contingent. But there is no need to address the issue of necessity. We simply have to note the connections, and we shall allow them to inform our research programme. Cognition, that is to say, will be understood from the standpoint of embodiment. No formal opinion is required on the possibility or otherwise of disembodied spirits or artificially intelligent programmes resident in cyberspace.[31] But such fantasies will not intrude on our enquiries.

On the other hand, in choosing to understand cognition from the standpoint of embodiment we have not prejudiced our reflections against the possibility of cognitively endowed robots or intelligent extrabiological life forms. On the contrary, our remit should include as much as we can comprehend of what Sloman evocatively calls the *space of possible minds*.[32] In Chapter 1 it was emphasized that deep science seeks overarching principles, while retaining empirical legitimacy and remaining grounded in scientific practice. Our investigation has to proceed from cognition and consciousness as it is found in humans and (perhaps) certain other animals, the only exemplars we have. But our ambition must be to abstract away from those exemplars to a set of governing principles—principles of organization, dynamics, or architecture – whose application is more general, much as Darwin was able to articulate a theoretical foundation that not only underpins our understanding of all life

[31] In fact, the conception of embodiment in play here does not rule out the possibility of a virtual body inhabiting some virtual reality.

[32] Sloman (1984).

on Earth, but to which extraterrestrial life (should it exist) must surely also conform, as does the operation of a genetic algorithm in a computer.[33]

The principles of organization, dynamics, and architecture we shall alight on in the coming pages entail no particular commitment to a biological substrate. But for some authors, such as Thompson, 'life and mind share a set of basic organizational properties, and the organizational properties distinctive of mind are an enriched version of those fundamental to life'.[34] According to this way of thinking, an organism perpetually constitutes its own identity through metabolic exchange of matter and energy with the environment so as to maintain the boundary between self and non-self. At the same time this process of *autopoiesis* both brings forth a domain of concern, wherein features of the environment acquire significance according to their relevance to the organism's well-being and perpetuation, and opens up a spatial and temporal horizon for the organism. A domain of concern and a spatiotemporal horizon are seen as prerequisites for lived experience.

The set of organizational principles derived here, by contrast, is founded on an empirical account of the distinction between conscious and unconscious aspects of behaviour, both of which conditions arise in the everyday life of a human being. There is no reason to suppose that the dynamical signature of the conscious condition could not be realized in a system having the right connective topology and implemented on a digital computer. If suitably embodied (as a humanoid robot, say), there is no reason to suppose that such a system could not interact with us and our environment in such a way that our attitude towards it would be the attitude we take towards our peers and equals.[35] Such an artefact would surely have to be very life-like indeed. It would, no doubt, have its own domain of concern and display self-oriented purpose. But in the realm of possibility, metabolism is not a prerequisite for these features. In short, we should like to situate consciousness as we know it within the larger picture of consciousness *as it could be*,[36] and the possibility of *artificial consciousness*—of man-made artefacts with an inner life—is implicit here.[37]

[33] Holland (1975).

[34] Thompson (2007), p.128. Thompson's inspirations and allies in this matter include Jonas (1966) and Maturana & Varela (1980).

[35] See Wittgenstein (1958), p.178: 'My attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul'.

[36] This phrase echoes the ambitions of the field of Artificial Life, whose founders sought to locate 'life-as-we-know-it within the larger picture of life-as-it-could-be' (Langton, 1989, p.1).

[37] Haikonen (2003); Holland (2003); Aleksander (2005); Gamez (2008). See also Shanahan (2006) and the discussion of that work by Bringsjord (2007). If the prospect of artificial consciousness starts to look realistic it may be necessary to invoke Metzinger's proscription: 'we should ban all attempts to create (or even risk the creation of) [artificial

If artificial consciousness resides in a remote, and possibly non-existent, region of the space of possible minds, then animal consciousness—that is to say consciousness in non-human animals—should be less controversial. Indeed, the presumption of animal consciousness is enshrined in UK law. The 2006 Animal Welfare Act makes it an offence for a person to cause unnecessary suffering to a domestic animal. Although the act does not define what is meant by 'suffering', the very idea of an animal's suffering only makes sense if it is like something to be that animal, if the animal can experience pain, hunger, thirst, and so on. Yet the idea of animal consciousness is notoriously resistant to a scientific treatment.[38] The difficulty, of course, is that animals cannot tell us what they are thinking. The cautious researcher might accept that animals have *feelings*, in some sense, even that animals can *communicate* those feelings (by whimpering or wagging their tails, say), yet profess scepticism over whether a non-human animal can properly be said to *think*, and then compound that scepticism with doubt about whether there is any scientific way to settle the question if the animal's behaviour is all there is to go on.

The philosophical debate here hinges on the role of language. A philosopher might grant that animals are capable of suffering—that they experience pain, hunger, and so on—yet deny them thought, properly speaking, on the grounds that, lacking language, they lack concepts and the means to involve concepts in a rational process of deciding how to act or what to believe.[39] One form of riposte is that, although we cannot interrogate an animal over the reasons for its choices, nor elicit a verbal report of what an animal perceives or how it feels, it may still be appropriate to speak of its having wordless reasons for its actions or of thinking without language.[40] But at this stage, there is no need for us to take sides in this debate. The recommended approach here is to take ourselves as a yardstick. First we need to establish a scientific framework for understanding the conscious/unconscious distinction in humans. This framework should incorporate both behavioural *and* neurological indices. Then we can extrapolate from the human case and apply the principles we have uncovered to non-human animals.[41]

consciousness] from serious academic research' on the grounds that otherwise 'we might dramatically increase the amount of suffering, misery, and confusion on the planet' (Metzinger, 2003, pp. 620–622).

38 Griffin (2001).

39 Davidson (1982) is representative of this stance. See also McDowell (1994): a 'mere animal does not weigh reasons and decide what to do [and] the milieu it lives in can be no more than a succession of problems and opportunities, constituted as such by [immediate] biological imperatives' (p.115).

40 Bermúdez (2003); Hurley (2006).

41 This is also the approach advocated by Edelman & Seth (2009).

Chapter 3

Probing the internal

This chapter looks at the challenge of operationalizing the conscious/unconscious distinction. The strategy is to distinguish consciously mediated behaviour from automatic behaviour, taking introspective report as a gold standard while granting its limitations. Particular attention is given to certain distinctive phenomena, such as the Sperling effect, that make it difficult to devise practical experimental paradigms for contrasting conscious and unconscious conditions. A thought experiment involving an imaginary psychologist with extraordinary powers is used to characterize an idealized (and unattainable) contrastive data set. Finally, a number of cognitive and behavioural correlates of reportability are hypothesized, touching on flexibility, inner rehearsal, and memory.

3.1 The conscious/unconscious distinction

Although the word 'consciousness' makes regular informal appearances throughout this book, it would not be proper to characterize our aim as 'explaining consciousness'. Rather than trying to explain an amorphous something or other that no one can define clearly in the first place, our initial explanatory target is a *distinction*, the conscious/unconscious distinction. Some of what goes on around us we are conscious of. But some of what goes on influences our behaviour unconsciously. Similarly, some of what we do we do consciously, but some of what we do we do unconsciously. Our task is to understand the nature of these contrasts, and to begin to account for them scientifically. Let's begin with an everyday example.

Suppose someone sets out to brush her teeth. Usually there's a tube of toothpaste on the shelf. But even when it has run out, there's often a new tube in the medicine cabinet. Clearly there is a significant difference between the implicit and unconscious belief that there is toothpaste when it is there on the shelf as usual, and the conscious thought that there is toothpaste (in the medicine cabinet) although none is visible on the shelf. In the former case, the toothbrushing subject habitually and unthinkingly retrieves the tube from its resting place. But in the latter case, she must respond to an unexpected situation arising in the middle of a routine operation. The absence of toothpaste in its usual

place intrudes on what it's like to be that subject. The thought that there is toothpaste after all, because there is a new tube in the medicine cabinet, might result in an immediate verbal report (to a yawning partner), or be committed to memory and offered up later (as pillow talk, perhaps). The thought might elicit an emotional response (irritation at having to go to the medicine cabinet), and give rise to deliberate action (a short journey across the bathroom).

But if the toothpaste is in its usual place, none of this occurs. The toothbrusher locates the tube visually among the clutter of the shelf, reaches out, and grasps it, automatically taking account of its orientation and adjusting for its shape (slightly more squeezed and folded than yesterday). It's surely appropriate to say that she believes the toothpaste is on the shelf, that the tube is not empty, and so on. How else could she have completed her toilette? Yet after her ritual ablutions she is unable to recall anything about the action of picking it up. Not only is she unable to say where the tube was in relation to the various other objects on the shelf, report how full the tube is, or recall what brand of toothpaste she used, she confesses that she cannot remember picking up the tube at all, although she doesn't deny having done so. In short, the act of retrieving the tube of toothpaste made a negligible contribution to what it was like for her during the pertinent 5 minutes of her life.

This little cameo has introduced several themes that we shall return to in the ensuing discussions. Our waking life is a patchwork of activities that are in part automatic and in part consciously mediated. Activities such as making toast, driving to work, playing table tennis, or logging on to a computer are predominantly habitual, whereas others such as learning to play a board game, composing a letter, or navigating by map, require considerable conscious intervention. Some activities, such as composing a letter, will always demand consciousness. Others, such as driving or playing a video game, demand attention and concentration from a novice, but become increasingly automatic with practice. But as the teeth-cleaning example illustrates, even a habitual task has recourse to consciousness when something goes wrong, when the conditions for its normal execution are violated. When we respond to a familiar situation with habitual or automatic behaviour, the details are hard to recall afterwards. But novel circumstances, and tasks that command our full attention, are more memorable and their performance is more easily described later.

It's time to put some cards on the table, and to make plain an important thesis on which the argument of this book rests. The contrast between automatic and consciously mediated behaviour, as just drawn, is a foundation stone of our approach to the conscious/unconscious distinction, and the rationale for this is the following working hypothesis. Consciously mediated behaviour may be slower and more effortful than automatic behaviour, but it is cognitively efficacious under a variety of special circumstances—in situations

that are novel, for example. Special circumstances such as these requisition a global communications infrastructure in the brain that underlies the inner life of a human being. This communications infrastructure, which we shall call the *global workspace* after Baars,[1] receives information from, and disseminates information to, numerous parallel processes operating on multiple levels, and thereby integrates their otherwise segregated activity.

According to the proposal to be pursued, the integrative facility supplied by a global workspace gives rise to the conscious condition in general, a richly empowering condition in which a whole battery of cognitive faculties is concurrently engaged—learning, working memory, episodic memory, language, and so on. Perhaps it's not clear that a contrast inspired by such mundane exemplars as cleaning teeth will do justice to the rich pageantry of human inner life. What of internal speech, of dreams, or meditation? What of a chess master contemplating his next move, or a silent mourner at a graveside? But the claim is that even the most subtle aspects of our inner lives are realized by internal mechanisms built on top of firmware that has evolved to marshal the numerous distributed processing resources of the brain and to organize their combined activity into an effective response to the ongoing situation. This is the fundamental role of the conscious mode, the way in which it subserves cognition and the chief reason it has survival value.

Some effort is now required to make this standpoint more precise, and to see how we might validate and enlarge on it. First, we need to clarify and operationalize the concept of automaticity and the associated contrast with consciously mediated behaviour. Second, we must conceive experimental methods that will distinguish the kinds of behaviour associated with these two conditions. Third, we need to develop the means to understand the internal workings that underpin the behavioural distinctions duly drawn, and to do so under the auspices of a theory broad enough to cover more of the space of possible minds than has been actualized so far in the natural history of this planet. This will motivate us to take an architectural, systems-level view, and to embrace the language of dynamical systems.

3.2 **Being on autopilot**

Let's characterize a subject's behaviour as *automatic* to the extent that she is not conscious of the sensorimotor activity that constitutes it.[2] This is largely

[1] Baars (1988; 1997; 2002).

[2] Automaticity is characterized and indexed in a number of ways in the mainstream psychology literature. For a review see Moors & De Houwer (2006), and for an overview of the topic see Schneider (2009). Unconventionally, the present definition builds in the contrast with consciously mediated behaviour.

a terminological manoeuvre, as attaining a clear view of the conscious/unconscious distinction is itself work in progress. But it's a useful manoeuvre. For example, it's clear that automaticity so characterized has varying degrees. A long behavioural episode comprises many distinct sensory and motor events, and a person may be conscious of some (such as seeing toothpaste in the cupboard) but not others (such as picking up a tooth brush). Moreover, a single sensory or motor event may comprise a number of sensory or motor subevents, some of which might be conscious whereas others are unconscious. When she picks up the new tube of toothpaste, our subject may be conscious of its brand but not of its orientation on the shelf, even though the latter has guided her reach and determined the grip she has used.

Although this characterization of automaticity places emphasis on sensory events that influence behaviour notwithstanding that the subject is not conscious of them, the sensory events in question are not subliminal. Subliminal effects occur when a stimulus is presented for too brief a period to be reportable—perhaps for a few tens of milliseconds—in spite of which it still has a measurable influence on behaviour. Subliminal effects have been exploited in a variety of experimental paradigms, such as visual masking with priming, whose aim is to shed light on the conscious/unconscious distinction.[3] However, although subliminal effects may sometimes arise in Nature, they are surely a rare occurrence, and it's not clear how much insight is to be gained into the conscious/unconscious distinction by using them to exemplify the unconscious condition. As Bargh and Morsella put it, 'assessing the unconscious in terms of processing subliminal stimuli is analogous to evaluating the intelligence of a fish based on its behavior out of water'.[4]

Thus far we can concur with Bargh. None of the pertinent sensory events in the teeth-cleaning scenario, whether the subject is conscious of them or not, involves a presentation brief enough to elicit a subliminal effect. However, we will not follow Bargh in his shift of emphasis away from the study of sensory events of which a subject is not conscious, and onto the study of the unconscious *influences* on behaviour of stimuli that the subject is perfectly conscious of.[5] Such influences have been shown to arise, for example, when a social stereotype is activated that biases a subject's subsequent behaviour in ways that he fails to report when asked to explain the reasons for his choices.[6] This is, of course, an important research area. But it's not the right starting point for the

[3] Kim & Blake (2005); Breitmeyer & Öğmen (2006).

[4] Bargh & Morsella (2008), p.74.

[5] Bargh & Chartrand (1999); Bargh & Morsella (2008).

[6] Bargh & Morsella (2008).

present project. We shall be concerned to supply an account of the conscious/unconscious distinction that could, in principle, be applied to prelinguistic infants and non-human animals. These are creatures that (presumably) lack any conscious awareness of the reasons for their behaviour despite (presumably) being conscious of many of the sensory and motor events that contribute to it. So our initial focus must be consciousness (or the lack of consciousness) of the sensory and motor events themselves.

Very well. But how is this notion of automaticity to be operationalized? Are there methods for empirically establishing the extent to which a given behavioural episode is automatic or consciously mediated? If our subjects are adult humans, then the obvious answer is that all we need to do is ask them. Indeed, verbal report is taken to be the most reliable indicator of the conscious condition, a kind of gold standard, by many researchers engaged in the scientific study of consciousness. So all we need to do, it would seem, is detain our subjects for interview as soon as possible after the behavioural episode of interest, and to question them in detail about what they experienced and how they acted during the relevant period. However, this approach is beset with difficulties, and to get a sufficient grip on the relevant issues requires a standalone discussion.

3.3 **Introspective report**

For the sake of variety, let's bring out a new example—weeding a vegetable patch. A gardener has to discriminate weeds from planted seedlings, digging up only the former and discarding them onto a compost pile. Suppose we find him absorbed in his task. We watch as he selects an unwanted specimen for removal, sparing several lettuce seedlings nearby. He casts the weed aside, and at that point we interrupt him with some questions. (He is forbidden to look down at the vegetable patch to find the answers.) Roughly how tall was the plant he just dug up? What shape were its leaves? How many lettuce seedlings were there nearby? Did he dig the roots out from the left or the right? Perhaps his answers lack conviction. Some are accurate, but some are just guesses, and some are plain wrong. He wasn't really paying attention to what he was doing just then, he tells us. In fact, he was thinking about his tax return. Now, what do the gardener's honest answers tell us about his inner life? What, in general, do a person's verbal reports tell us about what it is like to be that person?

There is no doubt that there are cases in which introspective report is an excellent indicator of conscious experience, both its presence and its absence. Consider the credence given to verbal pain reports. If a patient says he feels nothing in a certain part of his body—thanks to the administration of a local anaesthetic, for example—the doctor does not dispute with him. It's perfectly

ethical then for the surgeon to apply the scalpel. Equally persuasive are cases of verifiable introspective report. For example, consider a hearing test in which tones of various frequencies are played at random times to a subject wearing headphones, who is instructed to raise her hand whenever she hears one. (Introspective report does not have to be verbal, as this example demonstrates.) Because the test administrator knows when a tone is presented in the headphones, there is independent verification of the introspective report represented by a raised hand. A raised hand that coincides with the presentation of a tone indicates that the subject is conscious of the tone, that the tone is contributing to what it is like to be that subject at that moment. Of course, this paradigm would fail if the subject were exposed to some other cue—a view of the administrator's hand on the switch, for example—that allowed her to work out when a tone was being played. But under laboratory conditions the means to cheat can be excluded.

The utility of other types of introspective report for establishing what a subject is and is not conscious of is more controversial, even assuming laboratory conditions. There are two distinct issues at hand. First, we need to examine the usefulness of introspective report for establishing the *presence* of conscious experience, in trickier cases than pain or a hearing test. Second, we need to consider what introspective report—or the lack of introspective report—can tell us about the *absence* of conscious experience, again in the tricky cases. We also need to consider both *retrospective* and *concurrent* introspective reports. Let's begin with retrospective report. The obvious difficulty with a delayed introspective report is that its veracity can be compromised by the medium of memory. Even a short delay between a sensory or motor event and the corresponding report opens up the possibility of corruption.

The gardener is a case in point. He states not only that he does not recall what shape the leaves were on the plant he just dug up, but also that he did not notice at the time. Yet his behaviour showed sensitivity to leaf shape, because he pulled up only weeds, leaving the lettuce seedlings unscathed. Maybe he was fleetingly conscious of the leaf shape. Maybe it contributed to what it was like to be the gardener for half a second or so, but the experience of seeing the leaves left no trace in his memory. Or perhaps a memory of this conscious event was indeed laid down, but faded so rapidly that by the time he came to ponder the matter it had disappeared altogether. Or perhaps the very act of verbally recalling the previous few seconds eroded the memory of those few seconds, so that certain details were lost. How can we distinguish these possibilities from each other, given that each one results in the same incomplete and potentially inaccurate retrospective report from the gardener?

This potential for confusion is no mere theoretical fancy. A celebrated experiment demonstrating just such an ambiguity under laboratory conditions was

reported by Sperling in the early 1960s, and its interpretation is still controversial today.[7] In Sperling's two-part experiment, subjects were briefly shown a grid comprising three rows of four letters each.[8] In the first condition, after the grid was removed the subjects were asked to recall as much of it as possible. In the second condition, immediately after the removal of the grid subjects were presented with a tone whose pitch could be either high, middle, or low, and their task was to recall the four letters in the row corresponding to the tone—high for the first row, middle for the second, and low for the third. In the first condition, subjects were only able to recall on average 4.3 out of the 12 letters. But in the second condition, they were able to recall on average 3.0 of the 4 letters in the requested row, despite not knowing in advance which row this would be. This led Sperling to postulate a short-term 'iconic' memory that retains the image of the grid for long enough to enable any given row to be read off, but which degrades too easily to allow the whole grid to be recalled.

The question that exercises students of consciousness today is this.[9] Is the subject of the Sperling experiment at any point conscious of all 12 letters in the grid? If he is, then how come he is able to report only a fraction of them? But if he is not, how come he is able to report the majority of any given row? According to Block, the Sperling experiment is evidence that 'phenomenology overflows access', which is to say that a person can have an experience—such as that of seeing the whole grid—to which her cognitive faculties, including those that underlie verbal report, are denied (full) access.[10] Regardless of whether we agree with this, and whether or not we accept Block's distinction between phenomenological and access consciousness, the Sperling experiment shows that the possibility of a conscious experience that is too fragile for full report is more than just a sceptic's plaything. It is one of several perfectly reasonable explanations of the Sperling effect.

According to the most obvious interpretation of the Sperling experiment, it is the very act of providing an introspective report that interferes with memory in a way that compromises the ability to report further. Perhaps the subject consciously sees the whole grid of letters with perfect clarity, but the memory of that experience is so delicate that to evoke it is partially to destroy it.

---

[7] Sperling (1960). A more recent experiment that raises similar issues is reported by Landman, *et al.* (2003), and discussed by Lamme (2003) and Block (2007). For a first-personal description of such effects outside the laboratory setting, see Blackmore (2009).

[8] In fact Sperling's paper describes a larger series of experiments, only two of which are related here.

[9] Block (1995). See also Bayne & Chalmers (2003), Shanahan (2005), and Block (2007).

[10] Block (1995).

To adapt a simile from James, it is rather like catching a snowflake in order to discover its structure, only to have it melt in the palm before the inspection is complete.[11] But a more radical interpretation is suggested by Dennett's critique of the relationship between consciousness and report.[12] Perhaps it is the very act of soliciting a report after the presentation of the grid that makes conscious what would otherwise have remained unconscious, namely the contents of a specific row.

At first, this possibility seems to make no sense. How could the act of soliciting a report *after* the event affect whether or not that event was conscious *at the time* of the event? But Dennett argues that 'what we are conscious of during a particular time period is not defined independently of the probes we use to precipitate a narrative about that period' and 'since these narratives are under continual revision, there is no single narrative that counts as the canonical version'.[13] In other words, there is no fact of the matter about what a subject is conscious of at any given time. Different answers can be obtained depending on when and how a report was solicited. On this account, there is nothing definitive to be said about what the Sperling subject was conscious of at the time of the initial presentation of letters.

The most potent example Dennett uses to bolster this position is not the Sperling experiment, but the so-called colour phi effect demonstrated in the mid-1970s by Kolers and von Grünau.[14] The basis of their experiment is the well-known phenomenon—thanks to which animation is convincing—of apparent motion when two or more images depicting an object a short distance apart are presented in rapid succession. In the Kolers and von Grünau experiment, the images in question are of coloured dots. When both dots are the same colour—green, say—subjects report seeing a moving green dot. But when the dots are different colours—green at time T_1 then red at time T_2, say—subjects typically report a moving dot that changes colour abruptly half way through its motion. An interesting question then arises. What was the subject conscious of just before T_2? According to the subject's report, the dot seemed red at that time, having already apparently changed colour. But the subject could not possibly have experienced a red dot before T_2 because at that time the second (red) dot had yet to be presented. Indeed, it might have turned out green. Dennett's counter-intuitive answer to the question is that there is

[11] James (1890/1950), vol. 1, p. 244.

[12] Dennett (1991).

[13] Dennett (1991), p. 136.

[14] Kolers & von Grünau (1976). A related illusion is the so-called cutaneous rabbit (Geldard & Sherrick (1972)).

no answer. There is no such thing as what the subject was conscious of just before T_2.[15]

With events on longer timescales, we might try to circumvent these difficulties by banishing retrospective report altogether, and insisting that our experimental paradigms use *concurrent* introspective report, perhaps using the method of 'protocol analysis' advocated by Ericsson.[16] As Ericsson argues, eliminating the delay between thought and report mitigates unwanted interference effects, and is helpful for obtaining reliable first-person descriptions of the thinking processes used in tasks such as mental arithmetic. And in a clinical setting, concurrent introspective report of symptoms such as pain or fatigue is useful for patient monitoring.[17] But these methods are inapplicable to the problem of operationalizing the distinction between automatic and consciously mediated behaviour. This is because the delivery of a running commentary during an episode of behaviour will force sensory and motor events to be brought to consciousness that would otherwise have remained unconscious. To vary a simile used by O'Regan and Nöe,[18] trying to use concurrent report to detect an unconscious sensory or motor event is like trying to find out whether a refrigerator light is off by opening the door. Or, to use another of James's figures, it is like 'trying to turn up the gas quickly enough to see how the darkness looks'.[19]

Despite its drawbacks then, it seems we must make the best of retrospective report. One way to mitigate its drawbacks is to solicit a response at just the right time. This should be long enough after the event both to eliminate the Sperling effect and to render the colour phi phenomenon unambiguous (the second dot has to appear before a report is offered), but not so long that significant corruption in memory is possible (Fig. 3.1).[20] To be more precise, we might posit two kinds of memory trace—a *sensory trace* and an *episodic trace*—which fade at different rates after a stimulus has disappeared.[21] To leave an episodic trace, and thereby permit report, is a hallmark of the conscious condition. Although the sensory trace of an unattended stimulus persists it can still participate in the conscious condition if attention is drawn to the stimulus

[15] Dennett (1991), pp.120–126.

[16] Ericsson (2006). See also Ericsson & Simon (1993).

[17] Stone & Shiffman (2002).

[18] O'Regan & Nöe (2001), p. 947. Also see Thomas (1999), p.219.

[19] James (1890/1950), vol. 1, p. 244.

[20] See Carman's (2007) discussion of the colour phi experiment and Dennett's interpretation of it (especially pp. 104–105).

[21] Sperling's 'iconic memory' is one type of sensory trace. The present concept is neutral with respect to sensory modality.

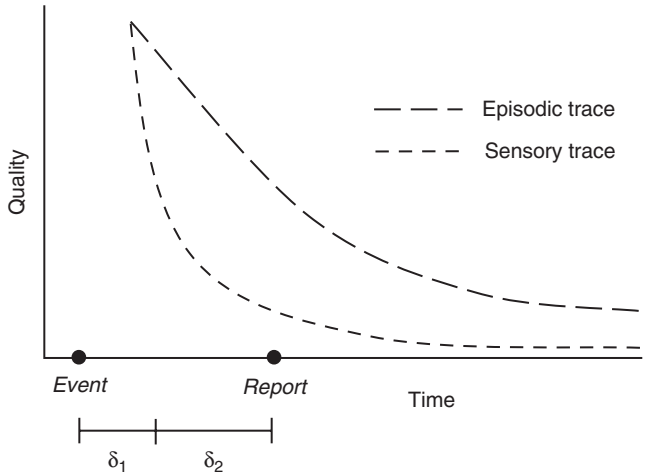

Fig. 3.1 The reporting sweetspot. The ideal time to solicit an introspective report for an event is long enough after the event for any sensory trace to have faded, but not so long that whatever episodic trace there may have been has also faded or become corrupted.

after its presentation, precipitating the start of a retrospective episodic trace. But after a sufficient delay ($\delta_1 + \delta_2$, where δ_1 is the minimum time required for any participation in the conscious condition), the sensory trace fades out, and only the episodic trace remains. This is the ideal time to solicit an introspective report, assuming no intervening events (such as masking stimuli).[22]

This proposal entails no particular stand on the question of whether or not there *really is* such a 'thing' as what a subject is conscious of at any given point in time, independently of our ability to measure it. To take such a stand, whether a denial or an affirmation, would be to stray into metaphysics. From a subject's point of view, it is only necessary to remark on how things seem, and from the standpoint of science the most that can be expected is an explanation of the best data it is possible to acquire, where what is best is what is deemed so by a self-critical community of scientists. What we are then left with is near-immediate retrospective report that, in its summary of the preceding moments of a subject's inner life, tends to run together periods shorter than a few tens of milliseconds.

[22] Appropriate values for δ_1 and δ_2 depend on the stimulus in question. Small values would suffice in the case of the Sperling experiment. But for the example of retrospectively counting chimes that go unnoticed until the third or fourth strike (Dennett, 1991, p.137), a longer period might be necessary.

3.4 Catching ourselves unawares

Its limitations notwithstanding, from a subjective, first-personal standpoint, verbal report authoritatively annunciates conscious experience. We take very seriously what we say (or could say) to each other, and indeed what we say (or could say) silently to ourselves, about our inner lives—and quite rightly so. If I say I can or cannot see, hear, or feel something, if I say I am happy or sad, excited or confused, then I do not expect to be greeted with full-on scepticism. To be sure, my confidence is sometimes misplaced, especially so in the sphere of affect. Consider the red-faced father who shouts at his children, 'I am NOT angry!', or the recently retired businessman who confesses, 'I never realized how unhappy I was in my job'.[23] Moreover, as the Sperling and colour phi experiments show, whatever I say will have limited temporal resolution. Nevertheless, from my point of view, what I have to say about my own inner life carries weight. This is the first-personal human perspective. It is from this perspective that things matter, that there are such things as suffering and joy, and that suffering and joy bear significance.

In short, introspective report can be thought of as a 'window' on the inner life—a window made of impure glass, we might say, but a window nevertheless—and in general we take at face value people's pronouncements about how things are for them. On this basis, we can investigate the *presence* of conscious experience by soliciting carefully timed verbal reports. However, convincing evidence of the *absence* of conscious experience, of what is *not* part of a subject's inner life, is harder to come by. On the one hand, the *inability* to produce a retrospective verbal report of a sensory or motor event does not warrant the conclusion that the event in question has *not* contributed to the subject's inner life, because it is hard to tell the difference *retrospectively* between something that never existed and something whose existence was too fragile to persist.[24] In the terminology of Merikle and Reingold, retrospective verbal report may be an *exclusive* indicator of certain types of conscious awareness, but it does

[23] Lambie & Marcel (2002).

[24] Such ambiguities have led to controversy over the existence and detectability of unconscious perception. Merikle, *et al.* (2001) overview experimental paradigms for detecting perception without conscious awareness, Snodgrass, *et al.* (2004) propose a further paradigm, and Eredyi (2004) emphasizes the need to take account of timing issues in such experiments. Yet Reingold (2004) questions the validity of these experimental methods, reiterating the criticisms of Reingold & Merikle (1988). Holender and Duscherer (2004) go further and question the very existence of unconscious perception. Dehaene, *et al.* (2006) clarify the issue from a neuroscience perspective. But the context for all this discussion is subliminal perception in the laboratory. None of these authors considers automaticity in an ecological setting, which is the present emphasis.

not seem to be an *exhaustive* one.[25] On the other hand, attempts to generate *concurrent* verbal reports of unconscious (yet supraliminal) sensory and motor events are inherently self-defeating. Yet without some means of detecting these things we will be hard pressed to operationalize the conscious/unconscious distinction by drawing a contrast between automaticity and consciously mediated behaviour.

Of course, absence of evidence is not evidence of absence—unless we have looked very hard in all the right places. If a verbal report is solicited from a subject soon after the events she is expected to report on, and if nothing occurs in the mean time to compromise her memory, then her inability retrospectively to describe some facet of the ongoing situation that manifestly influenced her behaviour can surely be taken as some evidence of automaticity in the sensorimotor arc that mediated that influence. Such inferences are widely used to gauge 'situational awareness' during occupations such as driving.[26] But science proceeds best through the accumulation of evidence from multiple sources, and thankfully we can supplement failure of retrospective report with other indices of unconscious sensorimotor activity.

One promising approach is to exploit a widely accepted feature of consciousness, namely its *limited capacity*.[27] Crudely put, it seems that there cannot be more than a very few distinct objects of consciousness at any one time. Exercising this hypothesis requires caution.—Perhaps it fails if the objects in question are presented in different modalities. Perhaps we can rapidly interleave consciousness of one object with consciousness of another object.—But if we accept the limited capacity hypothesis in some form, and if it can be shown thereby that at some time during the performance of a task the subject's conscious capacity is fully taken up with task-unrelated thought, then it can plausibly be argued that the subject has no conscious awareness of ongoing task-related sensorimotor activity at that time.

One method for achieving this is by means of what we shall term *cognitive masking*. The method here is to occupy the subject with a distractor task, for example by engaging her in conversation or giving her exercises in mental arithmetic. At certain times, and in some respects, we should expect a reduction in the level of performance of the primary task under these conditions. But at other times and in other respects, the subject's performance might not degrade, and this would be an indicator of automaticity.[28] For example, in

[25] Reingold & Merikle (1988).

[26] Strayer & Drews (2007); Ma & Kaber (2005).

[27] Baars (1988), Chapter 1.

[28] This is an example of a dual-task paradigm. Dual-task studies have been used to investigate the related idea of a *central bottleneck* in human cognitive processing (Pashler, 1984;

normal traffic conditions, a driver may be perfectly adept at maintaining her lane and keeping her distance from the car in front while conversing on a mobile phone. On the other hand, a driver conversing on a mobile phone might fail to notice something unexpected, such as a drunken pedestrian straying into the road, and be unable to react quickly enough to avoid an accident. (We shall encounter a different but related use of cognitive masking shortly.)

Another potential method for identifying automaticity by exploiting the capacity limitations of consciousness is to use *thought sampling* to identify periods of mind wandering. In this paradigm, the subject goes about her daily business equipped with a bleeper. The bleeper sounds at random intervals during the day, and as soon as it does the subject must stop what she is doing and record an immediate report of what she is thinking about at that moment.[29] Clearly this procedure is open to abuse and vulnerable to distortion of various sorts. But with suitably motivated and trained subjects, thought sampling is believed to be capable of detecting episodes during which the mind wanders and the subject becomes absorbed by thoughts that are irrelevant to her ongoing activity.[30] Insofar as this ongoing activity is oriented towards completing some task, and to the extent that there is no significant degradation in the performance of that task, it is reasonable to assume that these episodes of mind wandering are coincident with episodes of automaticity.

Of course, in each of these cases of alleged automaticity we would also expect the subject to be unable to offer a retrospective report about the pertinent aspects of her activity, and in this way we gain two vectors that converge on the same phenomenological hole. Moreover, we should aim to gather consistent results across a significant population. The early 20th-century demise of introspectionism as a respectable scientific methodology was partly due to the apparent irreconcilability of divergent first-personal accounts of certain subjective phenomena, such as 'imageless thought'.[31] By contrast, the reason that subjective phenomena such as the Müller–Lyer illusion remain acceptable as valid data in psychology is that they are consistently reported by large populations. Likewise, a robust conscious/unconscious distinction made on the basis of verbal report will only be acceptable if an experimenter can replicate the

Ruthruff, *et al.*, 2001; Lien, *et al.*, 2006). The relevant literature rarely discusses the conscious/unconscious distinction, but an obvious hypothesis is that the limited capacity of consciousness and the central bottleneck are manifestations of the same underlying architecture.

29 Stone & Shiffman (2002); Hurlburt & Akhter (2006).

30 Smallwood & Schooler (2007).

31 Costall (2006).

production/non-production of similar reports in comparable circumstances for significantly numerous sets of subjects.

3.5 **The omnipotent psychologist**

We are now in a position to pay another visit to our gardener. But we shall make him the victim of a thought experiment. Let us imagine what the findings might be of an experimental psychologist who is blessed with a certain kind of omnipotence. Thanks to some unspecified miracle—an especially large research grant perhaps—she has the capability to turn back time and replay the same episode of behaviour as often as she likes, making a different intervention on each occasion. Here is our gardener bending towards a weed. The species is unfamiliar, but it is certainly no lettuce. He plunges his trowel into the soil, prises up the unwanted plant, and discards it. Very soon after the rejected plant has disappeared from the gardener's field of view, the psychologist interrupts his work and subjects him to a short interview. What leaf-shape did that plant have? How many petals were on its flower?

With the answers duly noted, the psychologist presses the rewind button. The weed leaps from the compost pile and flies back into the ground. The gardener's trowel retreats from the soil, and the gardener himself is restored to an upright posture. Now the psychologist replays the scene up to the moment she interrupted him last time. But in this trial, when she arrests the gardener's progress, she poses a different set of questions. Where on the trowel handle was the tip of his thumb when its tip broke the soil's surface? Did the blade meet any resistance, from stones or other obstructions, or did it slide straight in? Where on the weed did he grasp it—by the leaves, or by the stalk? Again the psychologist makes due note.

Unlike a real psychologist, the omnipotent psychologist can repeat this procedure as often as she likes, varying the time of her interventions and the set of interview questions at will, compiling a catalogue of evidence as she goes along for which sensory and motor events have impacted on the gardener's consciousness and which (apparently) have not. For the gardener, the intervention is unexpected every time, the questions unanticipated. Moreover, each intervention can be made independent of every other. On the first trial, the psychologist might query the gardener's recall of the leaf-shape. Then, for the second trial, she can rewind and replay and query his recall of the petal count at exactly the same instant, knowing that the report elicited by the second request will be untainted by the corrupting influence of the first.

As well as soliciting different reports at multiple times for the same behavioural episode, the omnipotent psychologist can vary her experimental paradigm. For example, she might use the thought sampling method to look

out for periods of mind wandering. Suppose the gardener is not quizzed directly about the weed he has just pulled, but is instead asked simply to describe his thoughts, and suppose he tells the psychologist that he was thinking about his tax return. Specifically, he was wondering whether the cost of a recently purchased piece of equipment was tax deductible. Now, in a real experiment, whatever conscious awareness the gardener *might* have had of the appearance of the last weed he pulled, this is liable to have decayed by the time he has finished recounting his financial ruminations. So little could be concluded from his inability to provide a subsequent report about leaf-shape or petal count. But our omnipotent psychologist is in a better position. She can simply replay the scenario, and this time ask the gardener directly about the weed in order to confirm the evidence of automaticity already gained by thought sampling.

Using her god-like powers, our imaginary psychologist is in a position to build up a detailed picture of what goes through the gardener's mind as he tends his vegetable patch and, most importantly for our present agenda, which route is taken as it goes through—the conscious route or the unconscious. To visualize what's going on, let's suppose she builds an interactive chart on her computer, with time laid out from left to right. The chart is initially populated with information gathered by observing the gardener's behaviour without intervention. Every motor event, and every sensory stimulus that influences the gardener's behaviour, appears on the chart as a blue horizontal bar whose length corresponds to the duration of the event in question. Often, distinct sensory and motor events are concurrent—sighting a weed, hearing a bird, grasping the trowel, and so on, can all occur together—and these are separated vertically on the chart, resulting in a stripy appearance (Fig. 3.2).

Now the experiment begins in earnest. The psychologist makes a series of interventions, and further coloured bars are overlaid on the chart representing what the subject reports. Whenever the subject reports an experience that matches what observably took place, a faint purple stripe is overlaid on the chart, frequently over the top of an existing blue stripe representing the same event. Reports of sensory or motor events that do not correspond to what is observable appear as red stripes. Having sensibly automated the rewind–replay–interview procedure, the psychologist can recline in her chair and watch as the data floods in and the chart fills with colour. As more verbal reports come in, many of them pertaining to the same stimulus but obtained at different times, there is an increasing build-up of ever-deepening purple stripes. The less of a blue hue a stripe has and the more purple it appears, the more confident the psychologist can be that it signifies an event of which her subject is conscious. However, much of the chart remains blue, and as more data come in, the absence of evidence of consciousness indicated by these

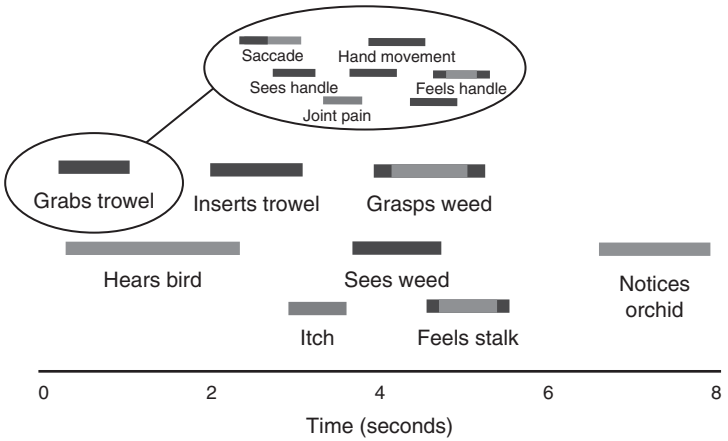

Fig. 3.2 An idealized data set showing conscious and unconscious sensory and motor events in an ecological (i.e. non-laboratory) setting. Verifiable conscious events are shown in magenta, unverifiable conscious events are shown in red, and verifiable unconscious events are shown in blue. Zooming in on an event shows up finer detail – a predominantly unconscious event might have many sub-events, some of which are partly conscious. The conscious condition is established (according to the thought experiment) by repeated interventions and elicitations of verbal report for the same replayed episode. (See Plate 1).

stubbornly blue stripes starts to look increasingly like evidence of the absence of consciousness.

At the end of the experiment, the chart is a patchwork of blue, purple, and red. Using its interactive facilities, the psychologist can now select any event on the chart and zoom in on it. Suppose she chooses a moment several minutes into experimental time, when the gardener puts down his trowel in order to look at his watch. She zooms in on the put-down-trowel event, which appears blue (unconscious) on the large-scale chart. When opened up, this event, which lasts approximately 1.5 seconds, turns out to comprise several distinct arm, hand, head, and eye movements mediated by several distinct visual, haptic, and audio stimuli (the position of the ground, the edge of the trowel, the feel of the trowel handle, and so on). This appears as another patchwork of colours. It is mostly blue (unconscious), but the psychologist spies an isolated little streak of purple, a conscious sensory event. On further inspection, the stimulus turns out to be a strange clang audible when the trowel touches the ground. There is a short bar of pure red parallel to the streak of purple—a brief moment of illusion. The clang sounds to the gardener like metal on metal. He responds to this with a tiny flick of his eyes. He fleetingly sees the true cause of

the noise, just a stone—another small streak of purple on the chart—and his gaze moves on.

The imaginary chart now spread out before our fictional psychologist represents a fundamental form of contrastive data set, wherein conditions of reportable and unreportable sensorimotor activities are laid out (Fig. 3.3). Of course, it is an idealization, a piece of science fantasy. In reality, no single episode in a subject's life can be scrutinized so closely, especially not in a natural, 'ecological' setting, and psychologists must rely on successive trials with different subjects to build up the best picture possible. (Obviously, to compile a truly thorough compendium of data, the fictional psychologist too must study many subjects, and many scenarios for each subject.) But the thought experiment helps to make precise the proposed distinction between automaticity and consciously mediated behaviour, the possibility of replay mitigating many of the methodological problems of introspective report.

Now let's suppose the omnipotent psychologist embarks on a new phase of her research programme. The purpose of this second phase is to determine how the conscious condition influences the gardener's behaviour, over and

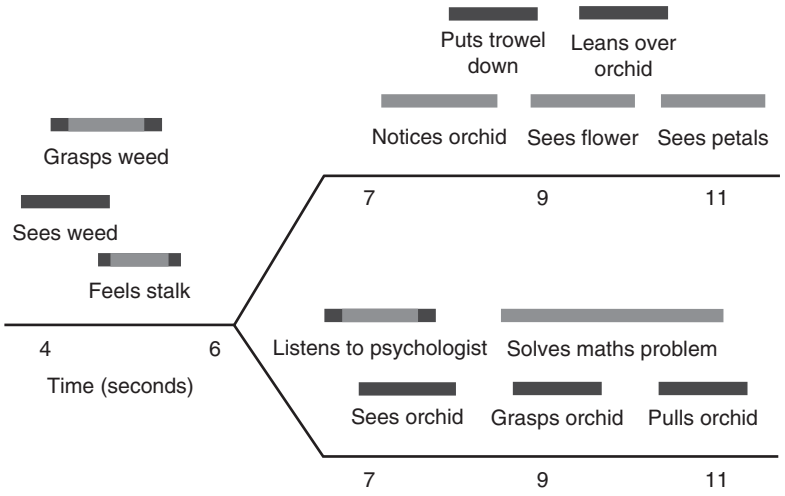

Fig. 3.3 Part of an idealized contrastive data set. According to the thought experiment, the psychologist is able to replay the same episode many times with different interventions, in order to compare closely matched variants with conscious and unconscious versions of the same sensory or motor event. In this example, the psychologist uses cognitive masking (a distracting mental arithmetic task) to render a sensory event unconscious that would otherwise have been conscious. The behavioural consequences of the distinction are then open to study. (See Plate 2).

above enabling him to make verbal reports. To establish this, she uses a cognitive masking paradigm, once again exploiting her unusual powers. For each stimulus of which the gardener is conscious according to the chart, the psychologist replays the relevant episode in the gardener's life yet again. But this time she devises an intervention that causes her subject to be occupied with an additional, cognitively demanding task at the time of presentation of the stimulus in question. This task could involve mental arithmetic, or everyday conversation, for example. By adjusting the timing and difficulty of the task appropriately, there should be many cases in which the psychologist is able to achieve the effect that the gardener is no longer able to report the stimulus—as determined by further replays, using the method already described—even though it is clear that his behaviour has been influenced by it.

For example, suppose a rare orchid has by chance seeded in the vegetable patch. The gardener spots the unusual specimen when his weeding brings him close to it, and he responds by stopping, looking more closely, then lovingly digging it up and setting it aside for replanting elsewhere. On the psychologist's chart, this episode is streaked with purple. In other words, when verbal reports about this episode were solicited at various times on successive replays, the gardener had a good deal to say, much of which closely matched what was objectively measurable. He could report the condition of the orchid's flowers, its approximate height, its position with respect to the plants around it, and so on. (He could not have had prior knowledge of any of these facts, so the psychologist is confident of the relationship between the reports and the stimulus.) Now, in the second phase of experimentation, the psychologist gives the gardener a piece of simple mental arithmetic to do just before he arrives at the orchid. Sure enough (let us suppose), if timed just right, this additional task is sufficient to prevent the production of most of the usual verbal report, and this is taken to indicate a lack of conscious awareness of the unusual plant.

Now the question is this. What difference does the cognitively masked condition make to the gardener's behaviour? One possible outcome is that the gardener will fail to notice anything unusual, and treat the orchid as if it were a weed. If this occurs, it's clear that the gardener has still, in some sense, seen the orchid. Otherwise he would not have inserted his trowel into the earth at its base, nor could he have picked it up in his fingers. But the fact that it's an orchid has seemingly failed to penetrate his consciousness, and his response to the stimulus of the flower is automatic. As a result he has missed an opportunity, and the psychologist has acquired one small example of the contrast between consciously mediated and automatic behaviour in otherwise closely matched circumstances.

It's worth noting that cognitive masking can be used in different ways. Earlier we saw that it can be used to support the hypothesized absence of

conscious awareness. If performance on a task in a cognitively masked condition is *comparable* to that in the unmasked condition then, under the assumption that a cognitively demanding task occupies the limited capacity of consciousness, it can be concluded that behaviour in the unmasked condition is not consciously mediated. Now we are seeing it used in a different way, namely to contrast consciously mediated and automatic behaviour in otherwise matched conditions. If performance on a task is *poorer* in the cognitively masked condition than in the normal, unmasked condition then, under the assumption that there is no separate limited capacity central bottleneck that could be dissociated from the limited capacity of consciousness, it can be surmised that conscious mediation is required for effective performance on the task in question.

Different scenarios will be more or less susceptible to manipulations of the sort just envisaged. If the gardener pricks himself on a thorn, then no amount of absorption in mental arithmetic is likely to eliminate the disagreeable sensation. Likewise, if the gardener breaks his trowel, then the intrusion of this event into his consciousness is inevitable, because he simply cannot continue without addressing the crisis. Moreover, even in cases where a reportable stimulus can be rendered unreportable, there might be no behavioural difference between the two conditions. The gardener might be consciously aware of the unexpected sound of his trowel hitting metal, but carry on weeding without bothering to look for buried treasure. However, the most interesting cases are those in which the extinction of a conscious stimulus through cognitive masking goes hand-in-hand with a behavioural change, as in the orchid example. The question such cases raise is this. What, if anything, do they all have in common? An answer to this question would clarify the sense in which the conscious condition is cognitively efficacious, and show us how this is manifest in behaviour that better subserves a subject's goals and needs.

3.6 **Novelty and flexibility**

An idea often mooted in the literature is that the conscious condition facilitates *flexibility* of behaviour, whereas automatic behaviour is rigid and stereotypic and therefore less able to handle the unfamiliar.[32] So perhaps the cognitively masked condition, when it affects behaviour at all, will give rise to less flexible behaviour. But what is meant by flexibility here? What makes one motor response stereotypic and inflexible whereas another is not? To answer

[32] The link between consciousness and flexibility is frequently alluded to in Baars (1988), for example, and Searle (1992, Chapter 4) discusses the connection in the context of automaticity in epileptic patients during seizure. The topic of flexible cognition is discussed at length in Carruthers (2006), although the issue of consciousness is not to the fore in his treatment.

this question, let's revisit the scale we encountered in Chapter 2, according to which the capacity to innovate in the face of novelty is a sign of cognitive prowess. At the low end of the scale we have the humble chicken who, though perhaps more adaptable than an industrial robot messily executing its usual motions even though the plastic parts it was designed to assemble have been substituted with cream cakes, nevertheless lacks the ability to innovate in the face of opportunity or adversity. To thrive in the ecological niche into which it has evolved, the chicken has no need of a sensorimotor loop that makes especially fine discriminations. A limited repertoire of simple behaviours is sufficient for it, and switching effectively between them requires only the recognition of simple cues. The world does not disclose to the chicken a combinatorially structured space of affordances, and the chicken lacks the means to explore such a space.

Further along the scale we find cognitively well-endowed non-human animals, such as the corvid family, which includes crows and rooks. The cleverness of these birds is apparent from their performance on the *trap-tube test*, a benchmark experimental paradigm devised by animal cognition researchers (Fig. 3.4).[33] The trap tube in question is a transparent cylinder at the centre of which an item of food is placed, visible to the animal but inaccessible without the aid of a tool. A stick or plunger is provided, and using this tool the animal can pull the food item towards one or other end of the tube. But there's an additional difficulty. At one end of the tube, a trap is interposed between the food and the exit. If the animal chooses to pull from the wrong end the food will fall into the trap and be lost. (A similar set-up involving rakes and a table with a trench (a trap table) was described in Chapter 2.) Although there is

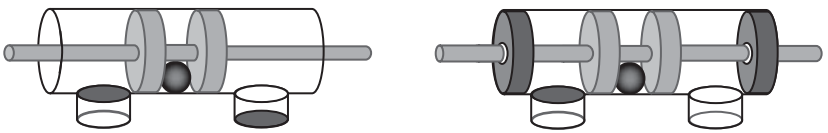

Fig. 3.4 The trap-tube test and a variation. In both versions of the test, a food item is lodged in the middle of a plunger inside a transparent tube. To pass the basic test (left), an animal must pull the plunger to the left so that the food passes over the false trap and pops out. If the animal pulls to the right, the food falls into the trap and is lost. In the variation (right), the animal must pull the plunger to the right so that the food item falls out of the hole. No associatively learned rule is sufficient to allow an animal to pass the variation on first presentation.

[33] Seed, *et al.*, (2006); Taylor, *et al.* (2009).

some variance in individual performance, many crows and rooks learn to pull out the food, avoiding the end with the trap.[34]

However, this can be achieved through trial-and-error and associative learning. What truly impresses is the crow's ability to apply what it has learned to previously unseen variations of the test, variations that would confound an associative rule based on a simple visual cue. For example, as Taylor, *et al.* have shown, in a population of crows, certain individuals are capable of transferring what they have learned on the trap tube to a trap table, which has no visual features in common with the trap tube yet shares the relevant spatial and causal properties.[35] Rooks, crows, and other corvids have evolved a variety of behaviours that is more complex than the chicken's—including tool-use and food-caching, for example—and a larger, more richly structured set of affordances is available to them as a result.[36] Moreover, they seem to have the cognitive capacity to explore this space of possibilities through a combination of experiment and insight, and so to expose whole regions of affordance that were previously hidden from them.

In general, a situation (or a particular stimulus in the context of a situation) invites flexibility when an animal's innate, habitual, or previously learned responses prevent it from best prosecuting its interests in that situation. If an animal repeatedly offers the same unfavourable response in such circumstances then its behaviour can be classed as inflexible. For example, a chicken that reacts to the sight of food through the perspex of the trap tube by repeatedly pecking the exterior of the apparatus will not gain any reward. On the other hand, if such a situation inaugurates a period of exploration of the space of potential action—an exploration whereby hidden affordances are revealed, either on-line through exploration and play, or off-line by means of some internal process—then the animal may exhibit a response that better promotes its well-being. Hence the talented crow, after a period of trial and error, learns to pull the food out of the tube, avoiding the trap. Whereas the chicken's behaviour is inflexible, the crow manages to adapt to the novelty of the circumstances.

So the crow's behaviour is adaptive. We might think of it as more flexible than the chicken's. But real flexibility, as we shall use the term, involves more

[34] Seed, *et al.* (2006) and Tebbich, *et al.* (2007) tested rooks, and Taylor, *et al.* (2009) tested crows.

[35] Taylor, *et al.* (2009).

[36] In the experiment reported by Seed, *et al.* (2006), an individual rook (Guillem) passed the variation of the trap tube shown on the right of Fig. 4.4 on first presentation. This is especially remarkable, as rooks are not known to use tools in the wild. However, captive rooks have been shown to be proficient tool-users (Bird & Emery, 2009).

than just the re-tuning of an existing behaviour, more than just a new marriage of cue to response.[37] Real flexibility is innovative, and transcends an animal's existing behavioural repertoire. It arises, for example, when the animal combines the elements of its existing repertoire in new ways. A hint of real flexibility, in this sense, is detectable in the rook or crow that succeeds in transferring its success at the standard trap tube to a novel variation. The bird has done more than merely adapt an existing behaviour using trial and error, more than just marry a cue with a response. Its behavioural repertoire has properly expanded, perhaps drawing on both its tool-use expertise and its food-caching expertise in order to foresee the consequences of using the plunger (which is like a twig) to push the food item into a hole (which is like a cache site). Human beings enjoy unparalleled flexibility, an ability to blend the elements of their behavioural repertoire in ways that open up whole new regions of affordance, and then to blend with the newly blended repertoire itself.

So the human is the point of reference at the high end of the cognitive scale. Recall the resourceful cook in the kitchen, who had to open a can of tomatoes using a variety of household tools instead of a tin opener. The word 'resourceful' is apt, because the cook's achievement was to marshal the numerous resources of past experience and apply them to a challenge she had never faced before, somehow selecting from a wide repertoire of skills and expertise a novel yet appropriate combination of actions to achieve her aim. Such innovation is the epitome of flexibility. The space of affordances that opened out before the cook was combinatorially structured. The behavioural possibilities could have been re-ordered and recombined in an indefinite number of ways, and the human has the capacity to explore this space both on-line, by interacting with the environment, and off-line, by rehearsing the outcomes of actions without performing them. Moreover, the space of human affordances is *open-ended*. The set of behaviours that can be combined and recombined is not prescribed in advance. It is forever being made and remade thanks to the productivity of human invention, and transmitted through the medium of human culture.

If the field of psychology were able to compile the sort of idealized, ecologically situated contrastive data set envisioned in the thought experiment described earlier, it seems likely that it would show severely degraded performance

[37] The term 'flexibility' is often used in a more liberal sense that encompasses mere adaptivity. For example, the Wisconsin card-sorting task, which is often characterized as a test of flexibility in cognition, assesses the ability to adapt to changing rules (Berg, 1948). Similarly, the classic experimental work of Schneider and Shiffrin (1977) is often characterized as showing that automaticity compromises flexibility, but again the paradigm used only assesses adaptivity in the present sense. (Of course, automaticity is even more likely to compromise flexibility in the more demanding sense being used here.)

in the cognitively masked condition compared to the consciously mediated condition in tasks that demand flexibility in the sense defined. Existing experimental evidence backs this up. It has been shown, for example, that drivers engaged in telephone conversations have poorer situational awareness of traffic conditions than those who are not distracted by any such task.[38] They are more likely to fail to respond to unexpected events pertinent to their journey, such as road signs relevant to their route, and are less likely to be able to report such events when questioned immediately after driving. Although the link to the conscious/unconscious distinction in these and similar experiments is open to challenge, the evidence they supply is broadly supportive of the hypothesis that the conscious condition facilitates flexibility.

3.7 **Accounting for the conscious condition**

If *every* instance of consciously mediated behaviour could be characterized in terms of flexibility in the presence of novelty, we would have a simple answer to our earlier question of what distinguishes the conscious condition from the automatic condition in behavioural terms. However, this is surely not the case. To see this, we need only consider examples of human behaviour that is mediated by thinking. Suppose a pupil is asked to divide 129 by 30 without the aid of pencil and paper. The pupil might approach the task in a number of different ways, depending on how he was taught and his preferred method. But he will certainly have to carry out several mental operations in sequence before offering an answer, such as first dividing 120 by 30, committing the intermediate result to memory, then working on the sub-problem of dividing 9 by 30. Similarly, consider a chess player contemplating her next move. Suppose she considers it advantageous to advance her rook, but she must check out the consequences of doing so. The move might entice her opponent's bishop out to threaten her knight. But the knight is protected by the rook. Aha! The chess-player quickly realizes her mistake. In the imagined scenario the rook will no longer be there to protect the knight. So she thinks again.

In both cases, there is reportability. If our psychologist conducts an interview with either of these subjects, they will be able to describe their reasoning processes.[39] The pupil is likely to be able to break down her thoughts into two or three specific episodes, and the chess-player is likely to be able to describe

[38] Strayer & Drews (2007); Ma & Kaber (2005).

[39] Alternatively, they could be asked to think out loud while performing their tasks (Ericsson & Simon, 1993; Ericsson, *et al.*, 2006). Of course, the omnipotent psychologist is in a position to try out both retrospective and concurrent reports on the same occasion, and to compare the results.

the moves she has rehearsed and rejected, specifying the order in which she mentally tried them out. We need not concern ourselves here with the question of the accuracy of such introspective reports (lacking, as they do, any external means of verification). No one would dispute with the bare claim, made by both the pupil and the chess-player, that they had thought through the problems facing them. In both cases, we have consciously mediated behaviour. In both cases, it is hard to imagine how the same results could have been achieved without conscious involvement. Yet in neither case is flexibility a hallmark of the behaviour in question. Rather, we are inclined to characterize both in terms of internal processes. The pupil's success is the outcome of a (familiar) mental procedure, whereas the chess-player's move is the result of inwardly rehearsing different (familiar) possibilities and weighing them against each other.

So we now appear to have two independent candidate indices of the conscious condition, in addition to the gold standard of reportability—enhanced flexibility in the face of novelty, and the ability inwardly to execute a sequence of problem-solving steps. Further possible indices of consciousness are not hard to imagine. Suppose a person is cleaning his house. He pushes the vacuum cleaner back and forth, periodically stopping to remove a stray object from the floor. But his mind is elsewhere. He's thinking about where to go on holiday. So when he picks up his glasses case and puts it on top of the bookcase, he does so distractedly. Later, when he needs his glasses, he has no recollection of where he put them. By contrast, if he is being especially mindful while cleaning—perhaps our psychologist has rewound time and asked him to name each object as he picks it off the floor—then he is more likely to notice the glasses case, and more likely to know where it is when he needs his glasses several hours afterwards. It's clear that the visual stimulus of the glasses case influences behaviour in both cases. Moreover (let's suppose), our psychologist finds non-reportability in the distracted case and reportability in the mindful case, so she is onto another contrastive data set. But the behavioural correlate of the conscious condition here has nothing to do with either flexibility or mental problem solving. It seems to have to do with the ability to lay down memories of a certain type.

In short, there is no single cognitive or behavioural correlate of the conscious as opposed to the unconscious condition. In these examples, besides enabling introspective report, it sometimes enhances flexibility in unfamiliar circumstances, sometimes accompanies a sequence of internal problem-solving steps, and sometimes facilitates memory. The chapters to come will support the contention that, properly to account for the conscious condition, it is a mistake to confine theoretical consideration to the cognitive or behavioural levels. Without neglecting these, we should also be looking at the

brain mechanisms that underlie consciously mediated behaviour. The claim we shall promote here is that the hallmark of the conscious condition is that it integrates the activity of brain processes that would otherwise be insulated from each other's influence. But as a prelude to an account of this sort, a few paragraphs are in order on the very concept of a brain process, as brain processes will feature prominently in the architectural explanation to come.

The formation of a star, the life cycle of a frog, and the construction of a house could each be described as a process. However, a *mental* process is something special. As Wittgenstein pointed out, inner processes 'stand in need of outward criteria'.[40] In everyday life we talk about each other's inner lives as freely as we talk of bricks, or cars, or pets—'Kerry is happy' is no more peculiar a sentence than 'Liam is holding a frog'—and we get by perfectly well without raising questions of outward criteria. But when neuroscientists or psychologists talk of the mind's internal workings—whether they speak of mental processes, cognitive processes, or brain processes—they are in each instance speaking of things whose existence and characteristics are inferred in complex ways from experimental data. Scientific conventions for talking about the inner life are very different from those of everyday speech, and philosophers of mind worry about the exact relations between them, applying such labels as reductionism, eliminativism, and so on. But there is no need for these labels here. What matters to us is the actual practices of working scientists, how they resolve their differences, and how they achieve consensus.

For example, consider the Stroop test, in which a subject is presented with a series of colour words (RED, GREEN, YELLOW, and so on), in coloured fonts. The task is to name the font colour (not to read the word). When the font colour conflicts with the word itself (the word GREEN is presented in a red font, say), a subject's response time typically goes down and their error rate goes up.[41] This classic experimental paradigm permits a psychologist to argue coherently that insofar as there are distinct processes in the brain for handling colour and for dealing with the semantics of written words, their operations must interfere with each other. But fellow psychologists are at liberty to question this interpretation, devising extensions to and variations of the test to support their own views. Likewise, the much debated mental rotation experiments of Shepard and Metzler—in which the length of time subjects took to verify that one figure was a rotational variant of another was shown to be proportional to the length of the rotation—supported the tentative conclusion

[40] Wittgenstein (1958), §580.

[41] Stroop (1935); Macleod (1991).

that gradual processes of inward mental rotation were taking place, a hypothesis that is open to refinement and refutation by further experiment.[42]

In this way, thanks to rather than in spite of its controversies and conflicts, psychology has the potential to converge on an agreed inventory of inner processes and their characteristics, an agenda that is furthered by neuroscience, especially with the advent of *in vivo* recording methods such as EEG and scanning technology such as fMRI. As an example, consider the elucidation in the 1980s and 1990s of two parallel streams of visual processing, the dorsal and ventral.[43] Behavioural studies on patients with visual impairments suggested a dissociation between two kinds of deficit. Visual agnosia patients are compromised in their ability to verbally identify objects or to report differences in their shapes and proportions. Despite this, some visual agnosia patients retain the ability to grasp and manipulate an object in ways that depend on the object's shape. This suggests that perhaps distinct visual systems (processing streams) are responsible for recognizing form and for guiding hand movement, and that one system can be damaged while the other remains intact.

Though compelling, this dual-systems hypothesis would be easy to challenge if supported by behavioural studies alone. However, it is buttressed by a variety of findings in neuroscience. For example, Ungerleider and Mishkin compared monkeys with inferotemporal (IT) cortical lesions to monkeys with posterior parietal (PP) lesions.[44] Echoing the behavioural studies with human patients, the monkeys with IT damage were impaired at tasks involving object recognition whereas those with PP lesions were impaired at reaching and grasping tasks requiring sensitivity to form in visual cues. With the aid of results such as these, the dual-systems hypothesis was reinforced, and augmented with neuroanatomical detail. Two distinct visual processing pathways had apparently been identified—a ventral stream carrying information pertaining to the shape and type of an attended object, and a dorsal stream bearing information relating to the way to grasp and handle an object. Using fMRI, these results were subsequently confirmed in visually impaired human patients.[45] Nevertheless, the basic hypothesis of two visual streams has benefited from subsequent debate, and has undergone development and refinement as a result.[46]

In short, contemporary behavioural and brain sciences are methodologically entitled to make claims about inner processes. They do so with a legitimacy and authority that we gratefully inherit when we attempt to pin down the

[42] Shepard & Metzler (1971).

[43] Ungerleider & Mishkin (1982); Goodale & Milner (1992); Milner & Goodale (2006).

[44] Ungerleider & Mishkin (1982).

[45] Milner & Goodale (2006), Chapter 8.

[46] Milner & Goodale (2008).

contrast between conscious and unconscious mental activity. However, the ambition of the present book is not to supply piecemeal interpretations of a growing compendium of empirical data, but (rashly perhaps) to offer an overarching theory. A central tenet of this theory is that cognition and consciousness are intimately connected. Moreover, as argued in the previous chapter, cognition is essentially related to embodiment. Attempting to understand cognition without comprehending its situation within a sensorimotor loop and its immersion in the environment is like trying to clap with only one hand. So our investigation will proceed at the architectural level. Mindful of the availability of relevant and useful empirical findings that are interpreted in terms of isolated mental processes, our descriptions will be of whole systems, systems that are embodied and embedded in the environment. The aim is to understand how these systems are organized, how their component processes interact, and how their orchestrated activity issues in behaviour that is pertinent to an organism's well-being and life goals, and within this context to study the conscious/unconscious distinction.[47]

We should be wary of construing processes as neatly bounded computational units, and take care not to characterize cognitive architecture in terms of an ordered flow of information through the system from input to output, passing through various intermediate stages of processing on the way. Although this way of thinking has its uses, we should be cautious of the pitfalls of this kind of boxology. The connectivity of the biological brain is densely recurrent, and it makes little sense to think of information as flowing through the brain in one direction, from input to output. Rather, as Lamme and Roelfsema demonstrated for the case of the ventral visual stream, waves of activation can move back and forth between multiple regions until the activity in those regions settles into a temporary state of mutual equilibrium.[48] These so-called *metastable* states are easily disrupted by new stimuli or, more generally, by activity in the numerous other regions to which they are connected. A potentially productive way to conceive of the performance of the whole system is in terms of a fluctuating series of these metastable states, a series whose order and membership is tightly constrained by the environment to which the system is coupled via the body whose actions it controls.[49] When, in the material to come, we present an architectural blueprint, this is a better picture to have in mind for the system's dynamics than that of a large-scale, well-engineered piece of computer software.

[47] It should be obvious from this description that the 'phrenological' style of theorizing criticized by Uttal (2001) is also repudiated here.

[48] Lamme & Roelfsema (2000).

[49] Kelso (1995); Bressler & Kelso (2001); Werner (2007).

This page intentionally left blank

Chapter 4

Broadcast and the network

This chapter introduces the global workspace architecture, and pins down certain concepts essential to its presentation, notably those of a process, computation, influence, and information. The chapter then moves from a high-level, abstract account of the architecture to a more detailed characterization closer to the biological brain. With the global neuronal workspace conceptualized as a communications infrastructure, the scene is set for its description in network terms. A specific brand of hierarchically modular small-world network is hypothesized to facilitate the required combination of broadcast and competition, and empirical evidence that human brain structural connectivity conforms to this topology is reviewed.

4.1 The elements of global workspace theory

Where do we stand? Having engaged with, and worked through, our metaphysical inclinations, we emerged with a license to investigate consciousness empirically. Then, with the fact of our embodiment brought to the fore, we developed a characterization of cognition in terms of the exploration and opening out of an animal's space of possible affordances. Returning to phenomenology, we took a critical look at experimental methods for studying the conscious/unconscious distinction. The conclusion was that no overarching set of principles governing the distinction is likely to emerge at the purely behavioural level, and that the most promising basis for a successful account would combine an architectural framework with a dynamical systems perspective, while drawing on the insights of both psychology and neuroscience. The aim of this chapter is to advance a proposal along such lines, one that simultaneously gels with the view of embodied cognition arrived at earlier.

The chief inspiration for the proposal is the *global workspace theory* of Baars.[1] The heart of this theory is a specific information processing architecture, the major components of which are a set of parallel specialist processes and a global workspace (Fig. 4.1). Computation in the architecture proceeds through

[1] Baars (1988; 1997; 2002).

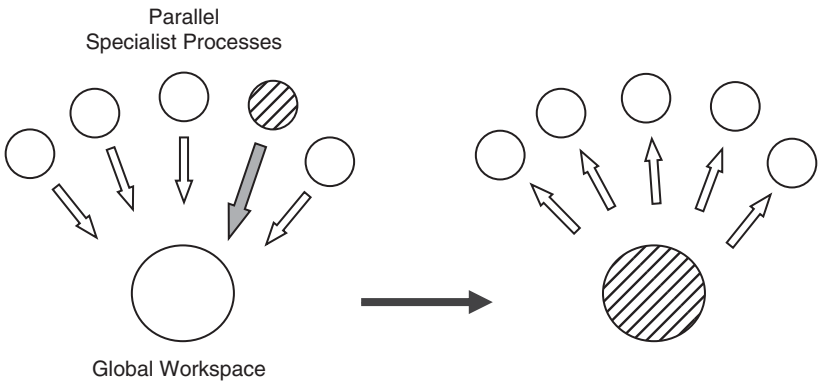

Fig. 4.1 The global workspace architecture. The overall dynamics alternates periods of competition (left) and broadcast (right). Parallel specialist processes (or coalitions of processes) compete to influence the global workspace, whose state is broadcast back to the whole cohort of parallel specialists.

a series of episodes of *broadcast*, during which information in the global workspace is disseminated to the whole cohort of parallel specialists, punctuated by bursts of *competition*, during which processes—or co-operating consortia of processes—attempt to influence the global workspace to the exclusion of their rivals. The central tenets of global workspace theory are that 1) the brains of humans and certain other animals conform to this architectural blueprint, and 2) the conscious/unconscious distinction mirrors the division between processing that is mediated by the global workspace and processing that is localized within the specialists A major attraction of the theory is that it supports the intuition that the conscious condition promotes flexibility in the face of novelty, because the blend of broadcast and competition at its core fosters integration among otherwise segregated brain processes.

One element of Baars's original presentation that plays a less prominent part here is the idea of a *context*, 'a system that shapes conscious experience without itself being conscious at that time', which encompasses 'currently unconscious *expectations* … and currently unconscious *intentions* that shape voluntary actions'.[2] Here we shall assume that the influence of context is subsumed within the general melee of parallel unconscious processing. When we unexpectedly encounter a friend on the street, when we walk through the front door of our home, when we arrive at an unfamiliar railway station—upon the occurrence of each such event a particular set of processes becomes active, priming

[2] Baars (1988), p. 138.

a collection of expectations and habits that goes together with the context in question. This set of processes remains active, modulating the activity of other processes, until the context changes.

Global workspace theory owes its pedigree to work in Artificial Intelligence.[3] Baars was inspired, in particular, by so-called *blackboard architectures*,[4] where the global workspace is analogous to the blackboard. Blackboard architectures, in turn, resemble the *pandemonium architecture* of Selfridge, in which a crowd of competing demons, each with its own specialist task, all shout to gain the attention of a decision-making super-demon, who is only influenced by the loudest of them.[5] In the printed commentary to Selfridge's 1959 paper, John McCarthy, progenitor of the field of Artificial Intelligence, presciently anticipated the main tenets of global workspace theory, commenting on

> the advantages of the pandemonium model as an actual model of conscious behaviour. In observing a brain, one should make a distinction between that aspect of the behaviour which is available consciously, and those behaviours, no doubt equally important, but which proceed unconsciously. If one conceives of the brain as a pandemonium – a collection of demons – perhaps what is going on within the demons can be regarded as the unconscious part of thought, and what the demons are publicly shouting for each other to hear, as the conscious part of thought.[6]

The central ideas of global workspace theory can be conveyed in a single paragraph. But the concepts deployed in the summary earlier merit closer scrutiny. What is a 'process' in the context of the theory, and when is a process a 'specialist'? What sort of entity is the putative 'global workspace'? What kind of 'information' is traded among the processes and passes through the global workspace? In what sense does 'computation' take place in the architecture? Each of these concepts in turn will now be further elucidated, beginning with that of a process.

---

[3] In Franklin & Graesser (1999) we find an AI system based on the global workspace architecture, and the chain of influences comes full circle.

[4] Nii (1986).

[5] Selfridge (1959).

[6] McCarthy (1959b). In September 2006, I reminded McCarthy of his comment, presenting him with a photocopy of the relevant page from the Teddington proceedings and pointing out its affinity to global workspace theory. He had no recollection of his remarks (almost half a century earlier), but quipped that he stood by them, because he had only said 'perhaps' that is what goes on. Nevertheless, he folded my photocopied sheet and put it in his pocket.

4.2 **Parallel specialist brain processes**

A warning was issued at the close of the previous chapter. A process in the brain should not be thought of as a neatly bounded computational unit with clearly defined inputs and outputs. How, then, are the processes of the global workspace architecture characterized? How are they individuated? Well, let's explore a similar but unrelated question. Consider the people moving around a department store on a busy shopping day. How should we define a *group* within this crowd? How are groups individuated, that we may better understand the organization of the crowd? From a bird's-eye view, an observer might pick out the collection of people around one of the cash desks. There is a marked congregation here, surely a candidate for a group.

But the constituents of this group (if it is to be called a 'group') are not stable. Individuals and small collections of people are constantly breaking away from the cash-desk crowd while others are coalescing with it. Smaller collections of people—families, posses of friends—might be candidates for groups too. (Is an individual a special case of a group?) Zooming out, we might aggregate all the people within the shoe department, and designate them a group on a larger scale. Similarly, we can view all the shoppers on the same storey of the building as a larger group still. But then what about the people on the escalators who are moving between floors? Which large-scale group do they belong to? Perhaps they constitute distinct groups of their own. And what of that child loitering halfway between haberdashery, where his mother is browsing, and the toy department next door? Perhaps he belongs to two groups simultaneously, or to neither.

As far as their individuation is concerned, processes in the brain are like groups in a department store. There is no ultimate answer to how either concept should be defined or individuated. The concept of a brain process is a human invention, and we import it into our scientific ontology insofar as it helps us come to terms with the complexity of what is a very large system of richly interacting components. Accordingly, brain processes can be identified at many levels of organization. A process might be realized by hundreds of neurons (a small patch of visual cortex that responds to motion, say), by hundreds of thousands of neurons (such as the set of neuronal assemblies, distributed across several cortical areas, that respond to a particular face by smiling), or by hundreds of millions of neurons (the set of brain regions implicated in working memory function, for example). But because evolution is not an engineer, the brain is not structured as a strict hierarchy, and there is nothing to prevent the constituents of one process from contributing to another.

Generally speaking, then, processes are resident in the brain at multiple scales, and their boundaries are fluid, indistinct, and overlapping. So the following question arises. Which of these highly numerous multi-scale, fluid,

indistinct, overlapping, spatially distributed processes constitutes the global workspace architecture's set of parallel specialists? Are there, perhaps, just a handful of such specialists at the highest level of brain organization? Or are there tens of thousands of them, busily going about their business in the service of a larger whole like ants in a colony? Indeed, what is meant by a 'specialist' process? A process might possess a narrow domain of expertise, but one that is applicable in many types of situation, such as recognizing faces or parsing speech. A process might realize a learned motor skill whose applicability is restricted to particular contexts, such as peeling fruit or tying shoelaces. A process might have a specific, but large and important, functional role, such as visual pattern recognition, episodic recall, or affective judgement. Which of these senses of specialization is pertinent to the proposed architecture?

In order to answer these questions, let's return to the ethologically inspired architectural sketch of Chapter 2 (see Fig. 2.2). According to this sketch, a simple animal (or a biologically inspired robot) is endowed with a certain repertoire of behaviours, such as foraging, mating, grooming, courtship, and so on. Each behaviour is triggered by a combination of internal deficit and external cue, and contention between behaviours is resolved by a competitive selection mechanism. Each competing behaviour is represented as a distinct process, as is behaviour selection. One option for a hypothetical global workspace architecture would be to identify the set of parallel specialists with a set of behaviours. The remit of each process would then be to realize a single highly specific form of behaviour, and the span of its responsibilities would reach all the way from sensory input to motor output.

However, this would be a poor choice. This is because, for the theory to live up to its billing, there ought be competition for influence on the global workspace *before* the distribution of information to the cohort of processes responsible for different behaviours. To see this, we need only consider any of a variety of phenomena in humans wherein our conscious experience of a given stimulus alternates between two (or more) possibilities. For example, in the paradigm of binocular rivalry, different images are simultaneously presented to each of a subject's eyes—a vertical grating to the left eye and a horizontal grating to the right, say.[7] What subjects typically report seeing is not the criss-cross pattern of both gratings overlaid on each other. Rather, they see one pattern to the exclusion of the other. Moreover, which of the two patterns they see at any given time is not fixed, but typically flips every few seconds, then flips back. Ambiguous images—images that can be interpreted in more than one way, but in only one way at a time—give rise to a related effect. For instance,

[7] Kim & Blake (2005).

the Necker cube can be interpreted as resting above the viewer's horizon or below it, and a subject can usually alter the interpretation at will.[8]

Such examples suggest that, in addition to the competition that takes place among different behaviours, there must be competition at the input end of the path between sensation and action. Indeed, all phenomena of attention lend support to this proposition. Top-down attention, as manifest in our ability to close out all but the conversation we are listening to at a noisy party,[9] bottom-up attention, which is at work when pertinent visual stimuli 'pop out' of a scene at us,[10] and inattentional blindness, wherein our awareness is so taken up with one aspect of a situation that we fail to notice certain events even if they are out of the ordinary[11]—all are evidence of competition for access to consciousness among processes that operate at the sensory level rather than the motor level.[12]

On the other hand, there is much to be said for preserving the sensorimotor emphasis of a behaviour-based architecture. According to the stance of Chapter 2, cognition is to be understood in terms of its role in embodied interaction with the environment, and its proper locale is the sensorimotor loop by means of which the embodied agent is coupled to that environment. So, insofar as the conscious condition is cognitively efficacious, an account of it that pays due respect to its sensorimotor effects is called for. Moreover, according to the contrastive approach of Chapter 3, the conscious/unconscious distinction is elucidated in terms of the subject's awareness of the sensory and motor events which her behaviour comprises. For sure, a theory that adhered strictly to sensation and action would proffer an inadequate portrait of human inner life, the outer signs of which are often more subtle than a twitching of muscles in reaction to a sensorial display. Yet the sensorimotor loop is the only foundation on which a more complete account can be built.

As a step towards accommodating the parallel specialists of a global workspace architecture within a sensorimotor perspective, the behaviour-based architecture might be decomposed in such a way that distinct sensory and motor processes show up (Fig. 4.2). In addition to effecting the requisite separation between input and output processes, this more refined view of the architecture makes explicit the possibility that two motor processes,

[8] Kim & Blake (2005).

[9] Cherry (1953). This is known as the 'cocktail party effect'.

[10] Treisman & Gelade (1980).

[11] Mack & Rock (1998); Simons & Chabris (1999).

[12] For an overview of many of the phenomena listed here and their relationship to the scientific study of consciousness, see Koch (2004), especially Chapter 9.

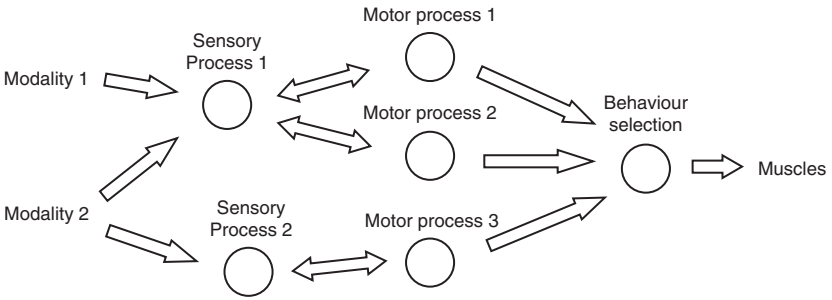

Fig. 4.2 The behaviour-based architecture with separate sensory and motor processes. Modalities 1 and 2 could be vision and touch, for example. The same input process can influence more than one motor process, and information from multiple modalities can be fused by the same sensory process. Note that information flows both ways between sensory and motor processes.

though belonging to different behaviours, might be influenced by the same sensory process. Moreover, a sensory process may fuse information from multiple modalities (the visual and haptic senses, say), although some behaviours may depend on one modality only. As depicted in Fig. 4.2, the information flow between sensory and motor processes is bidirectional, which allows for motor-sensory influences such as top-down attention and expectation.

The decomposition of the behaviour-based architecture into separate input and output processes suggests a viable delineation of parallel specialists for the global workspace blueprint. The revised blueprint (Fig. 4.3) retains a sensorimotor foundation while permitting competition and broadcast mid-way along the sensorimotor arc. That the architecture also underwrites a possible distinction between automatic and consciously mediated behaviour should be clear from the diagram. Automatic behaviour, comprising sensory and motor events of which the subject has no conscious awareness, is mediated by direct connections between sensory and motor processes without exercising any influence on the global workspace, that is to say without giving rise to the broadcast of information about its component sensory and motor events. On the other hand, the possibility of competitive access to, and brain-wide broadcast from, the global workspace, is independently apparent in both the input and output halves of the architecture. This, according to global workspace theory, entails an independent possibility of conscious awareness for both sensory and motor events.

We have arrived at a good first approximation to a plausible architecture. However, the present characterization is somewhat abstract. In the next couple of sections, we shall refine it further, hopefully averting some misconceptions along the way. For example, the architecture presents too sharp a division

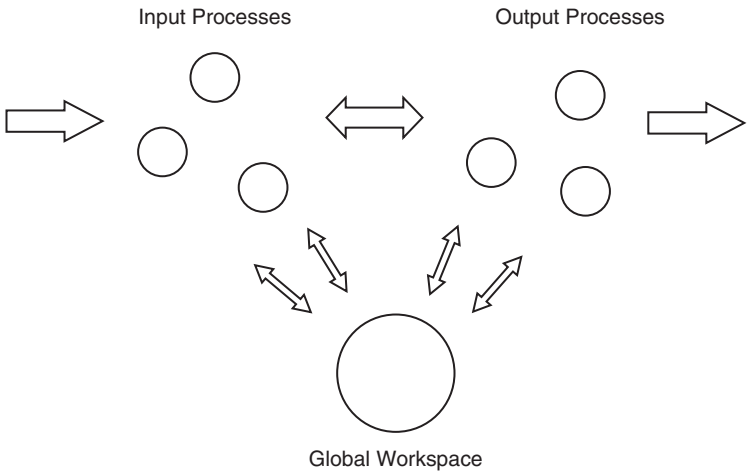

Fig. 4.3 The global workspace architecture with separate sensory and motor processes, and accommodated within a sensorimotor loop. Direct connections between input and output processes permit behaviour without conscious awareness, whereas the conscious condition is mediated by the global workspace.

between input and output processes, and despite the deployment of bidirectional arrows between sensory and motor processes, it remains suggestive of a serial pipeline of computation. Information apparently flows through the system stage by sequential stage, and the imposition of a few backward steps to implement top-down effects does little to mitigate this impression. In the next section, we review the concept of computation in the brain. This will lead to a more nuanced treatment of the architecture's dynamics, in which the concept of coupling plays a crucial part.

4.3 **Neural computation**

The brain is not a computer.[13] That is to say, there is almost nothing about its operation, whether at the level of abstract principle or of underlying substrate, that resembles that of the everyday device we use to send email, to browse the Internet, to store and display photos, to play music, and so on. To begin with, the brain is embodied, whereas an ordinary computer—that is to say a conventional computer running familiar applications—is disembodied. The brain's 'job' is to control a body and direct its interactions with the physical and social environment. An ordinary computer, by contrast, does not have to navigate the physical environment or manipulate the rich variety of objects it

[13] For a relevant discussion on this theme, see Edelman & Tononi (2000), pp. 47–50.

contains, and its severely limited interface with physical reality is through its connection to various static peripherals.

Beside this radical difference of function, there are several fundamental differences in organization. First, the architecture of a conventional computer comprises an active central processor and a passive memory, whereas in the brain there is no such division. Second, the behaviour of a computer running a familiar application is governed by a set of explicitly coded instructions written by a team of software engineers. The brain's dynamics, by contrast, is not programmed but is partly the product of evolution and partly the outcome of adaptation to the environment it finds itself in. Third, the ordinary computer of today is (largely) a serial machine that carries out one operation at a time, whereas the brain is inherently a massively parallel system.[14]

In addition to the differences in function and organization already cited, there are mathematical considerations that separate brains from conventional computers. In particular, an everyday computer is a digital device, whereas the brain is an analogue system. A complete description of the instantaneous state of a computer is possible using a finite set of binary (or natural) numbers, abstracting away from the details of its physical instantiation. The brain, on the other hand, is an analogue system. The membrane potential of a neuron (to pick just one physical property) is a continuous quantity, and its exact value is pertinent to predicting the neuron's behaviour. Theoretically speaking, a complete description of the instantaneous state of a brain is only possible using a set of real numbers—numbers drawn from the continuum, that is. This locates the brain outside the realm of conventional computation from a mathematical point of view, where the realm of conventional computation is defined by the set of functions that can be realized by a Turing machine.[15]

A related point is that ordinary computers are (predominantly) synchronous devices, whereas the brain's operation is asynchronous. That is to say, the entire state of a computer advances to its successor when and only when the computer's centralized clock ticks, so that all its internal events line up in time like a row of soldiers. But events in the brain, such as neuron firings, do not keep time in this orderly way. Although synchronized neural activity is commonplace, there is no centralized clock in the brain to maintain overall temporal discipline. In general, an electrical spike can be emitted by a neuron at any

[14] In fact, there is an increasing trend towards parallelism in contemporary computer engineering, with multi-core processors and highly parallel dedicated graphics processing being the norm. Nevertheless, the parallelism of the biological brain is of an altogether different order and dynamical sophistication.

[15] Siegelmann (2003).

time, where time, of course, is another continuous variable. From a mathematical point of view, this property alone could be enough to push the dynamics of the brain beyond the class of Turing computable functions.

So the brain is very unlike a computer. On the other hand, a computer can be programmed to be very like a brain. For a start, a computer can have an embodied function just as a brain does. It can be programmed to direct the actions of a robot through feedback based control, its input drawn from a variety of sensors, such as cameras or haptic devices, and its output driving effectors such as arms, legs, or drive-wheels. With respect to its organization, a computer can be programmed to implement a *virtual machine* whose principles of operation are entirely different from its own. For example, there are many types of parallel computer architecture, all of which can be emulated on a strictly serial machine using time slicing. The serial computer, so to speak, pretends to be each parallel processor for a short time, doing a little of the work of each in turn. If the serial processor is fast enough, it's impossible to distinguish an emulated parallel computation from the real thing.

One sort of virtual machine that a conventional computer can implement is a network of artificial neurons. The more biologically faithful the artificial neurons are, the narrower the gap becomes between the virtual machine and the brain. Using the differential equations proposed by Hodgkin and Huxley in the 1950s, for example, the spiking behaviour of real neurons can be modelled very accurately.[16] Moreover, a simulated network of Hodgkin–Huxley neurons can be supplemented with a Hebbian learning rule, such as spike-timing dependent plasticity (STDP), leading to a dynamical system that, like the brain, is not programmed but is open to adaptation to its environment.[17] Although the real computer this dynamical system is implemented on is conventionally organized in terms of an active central processor and a passive memory, this is invisible at the level of the virtual neural substrate, whose organization, given the computational power to simulate a sufficient number of neurons, can be made to mimic that of a biological brain.

Of course, no virtual machine can transcend the computational limits of the real machine that is its host. A conventional computer can only ever imperfectly model asynchronous events in a system of continuous variables. But this mathematical limitation may be less significant than it at first seems. Consider any dynamical system of continuous variables. Although it is not possible to represent any state (including the initial state) of the system exactly in a conventional digital computer, it is possible to represent it to an arbitrary

[16] See Izhikevich (2007) for an overview of the Hodgkin–Huxley model and its descendants.

[17] Song, *et al.* (2000); Caporale & Yang (2008).

degree of precision. Likewise, although it's not possible to simulate an exact trajectory through the system's state space on a conventional computer, it is possible, over any given interval, to simulate it to an arbitrary degree of precision. So, unless the system in question is chaotic, this means that a simulation in a digital computer can, in theory, be made to match, state for state, its continuous counterpart up to any degree of precision required.

If the system *is* chaotic—that is to say if a difference in its initial conditions, however small, is amplified over time and becomes arbitrarily large in the limit—then things are not quite so simple. In the chaotic case, because the initial state of the continuous system cannot be represented exactly in a digital computer, imprecisions in the simulation become ever larger over time. However, it is still often possible to simulate typical trajectories through the system, and to extract their statistical properties. So the extent to which the limitations of digital computation are a handicap when it comes to making conventional computers brain-like depends on the role of chaos in neurodynamics. That networks of neurons do indeed exhibit chaotic dynamics is highly likely, as Freeman argued in the 1980s for the example of the olfactory bulb.[18] But it may be the case that functionally equivalent effects—functionally equivalent for the purposes of behaviour and cognition—can be produced in a discrete system that merely simulates such chaotic dynamics.

In sum, although the brain is not like a conventional computer running familiar applications, a conventional computer running the right programme can be made very brain-like. Moreover, neural networks can be made to carry out computation. It has been proved that networks of neurons conforming to a variety of mathematical descriptions, including biologically realistic spiking models, can realize any Turing computable function.[19] Indeed, it has been shown that, with continuously valued synaptic weights, networks of certain types of neuron can also compute functions that are impossible to realize on a Turing machine.[20] Nevertheless, we still have not answered an important question. Is it appropriate to describe the operation of the brain in computational terms?

The brain's dissimilarity from a conventional computer running familiar applications is irrelevant to this question, of course, because we have a more general, more theoretical sense of the concept of computation in mind. We know that neurons can compute. But this too is an irrelevant observation, because it does not entail that mass neuronal activity in the brain is usefully

[18] Skarda & Freeman (1987).

[19] Siegelmann & Sontag (1995); Maass (1996); Carnell & Richardson (2007).

[20] Siegelmann (2003).

thought of in terms of computation. Moreover, our interest in this question pertains most closely to the architectural blueprint we have sketched out. How should we think of the parallel, specialist processes of the global workspace architecture? Are they computational processes? Or are they better characterized in some other way?

The issue is not metaphysical. We are not in pursuit of a claim of the form 'cognition *is* X' where X might be 'computation'. Such philosophically insidious uses of the existential copula are to be banished. We are simply looking for a theoretical vocabulary that has descriptive and explanatory value. From the standpoint of the next section, the behaviour of a set of brain processes is best characterized in terms of their mutual coupling, the trajectories they follow through their combined state space, and the attractors they fall into within that state space. As we shall see, the coupling between processes can be characterized in terms of their influence on one another, which will allow a particular concept of information to be incorporated into the explanatory framework. In the light of the earlier discussion, what results might be called a computational description. But the allusion would not be to the traditional idea of transforming input representations into output representations, and a less conventional paradigm of computation would be in play.

4.4 **Coalitions of coupled processes**

In Chapter 2 we encountered several reasons to favour the foundation of a theory of cognition on the concept of sensorimotor coupling with the environment. An animal's behaviour is a perpetual response to what its environment affords, and the benefit of cognition is to reveal to the animal affordances that were previously hidden. It would not be possible to understand either an animal's behaviour or the processes that gave rise to that behaviour if we observed the animal in isolation from its environment. The twitchings of a dog's nose make no sense without the scent-marked lamppost it is investigating. The manipulations of a squirrel's paws are meaningless in the absence of the nut it is peeling. In such performances, the internal dynamics of the animal's brain are locked in an ongoing embrace with the dynamics of its body and the outer environment. Here we will not only reinforce these intuitions in the context of the global workspace architecture's embodied situation, but will extend them to the internal relations among the brain processes that comprise the global workspace architecture's parallel specialists.

In its present draft, the architectural blueprint we have arrived at (Fig. 4.3) superimposes competition and broadcast (Fig. 4.1) on a behaviour-based architecture that includes distinct input and output processes (Fig. 4.2). Let's focus on the behaviour-based aspect of the architecture first, ignoring the

global workspace, and review the role played by the concept of coupling in our understanding of it. Consider a kingfisher perched delicately on a reed hanging over a stream. The reed bends and sways in the gentle breeze, and the kingfisher's blue-gold body bobs up and down with the rhythm. But its head remains stationary. As if fixed by an invisible bolt, it hangs motionless in space while the rest of the bird moves, a perfectly stable platform for viewing the riverside scene.

The kingfisher's achievement can be modelled as a pair of mutually coupled dynamical systems, one representing the relevant portion of the kingfisher's brain, and the other representing its body. The configuration of its body, including the head, is perturbed by the motion of the reed, which causes its vestibular system to signal its brain accordingly. Thanks to a closed loop control mechanism called the vestibulocollic reflex, these sensory signals modulate the neural activation that governs the kingfisher's neck muscles in such a way that when the rest of the bird's body rises, its head drops, and when the rest of its body drops, its neck stretches. Similar compensatory movements, regulated by vestibular feedback, maintain a fixed head position in the horizontal plane.[21]

It's tempting to think of the relationship between the kingfisher's brain and its body in terms of a staged pipeline of information flow. Indeed it's hard not to describe it in words that invite such an interpretation. But when we speak of the vestibular system sending signals to the brain, and of the brain's response to these signals, the sense of temporal and causal ordering these phrases conjure up is misleading. There is little explanatory benefit in attempting to isolate a discrete chain of events that begins with the vestibular signal, is mediated by the transmission of messages, and leads to a motor response. Rather, there is a continuous reciprocal relationship between two coupled dynamical systems that is best described by a set of differential equations in which neither system has temporal or causal precedence (Fig. 4.4, left). These guarantee that a certain parameter (head position) is kept within certain bounds (almost stationary) despite external perturbation (the motion of the reed).

It's not just to the relationship between brain, body, and environment that these considerations apply. The relations among processes within the nervous system are often best characterized in dynamical systems terms too. For example, the rhythmic motor activity that underlies locomotion in a wide variety of species has been shown to result from the interplay of sets of coupled neural oscillators located in the spinal cord.[22] Bipedal locomotion can be achieved

[21] Zeigler (1993).

[22] See the review by Ijspeert (2008).

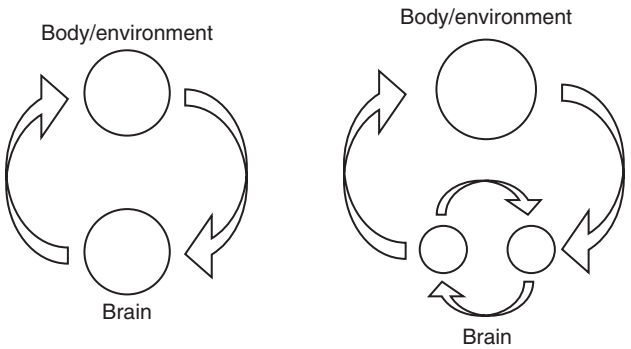

Fig. 4.4 External and internal coupling. The brain is dynamically coupled to the body and the environment (left). Neither partner in this continuous, reciprocal relationship is causally or temporally prior. Processes within the brain can be dynamically coupled in a similar way (right).

with a set of six coupled oscillators, one per leg joint.[23] Each oscillator has a natural frequency, which it adopts in the absence of external influence. When coupled, the oscillators entrain to each other. Oscillators assigned to opposite legs are connected in such a way as to synchronize in anti-phase. The whole system exhibits oscillatory behaviour whether limbs are attached or not, and is an example of what is known as a 'central pattern generator'.[24]

With the attachment of limbs, a further layer of coupling is introduced, with the body and the environment. Suitably tuned, the whole system exhibits robust walking behaviour in spite of disturbances such as variation in terrain or load. As with the kingfisher's head, there is no useful sense in which the internal relations among the oscillators, or the relationship between the oscillators, the legs, and the ground, can be described in terms of chains of discrete events mediated by the passage of information. Rather, these continuous reciprocal relationships are most easily captured by a set of differential equations that describes the ongoing balancing act of keeping several parameters within acceptable bounds so that the animal (or robot) continues moving forward without falling over.

Now let's reconsider the elements of the behaviour-based architecture. Recall that, in the context of global workspace theory, we provisionally separated input processes from output processes, for the compelling reason that there is manifest rivalry between sensory possibilities prior to behaviour

[23] Taga, *et al.* (1991).

[24] Ijspeert (2008). See also Yuste, *et al.* (2005), who propose endogenous pattern generation as a fundamental principle of cortical operation.

selection, as the existence of mechanisms of attention testifies. Our contention shall be that these input and output processes are coupled internally in much the same way that, considered as a single system, they in turn are coupled to the body and the environment (Fig. 4.4, right). The relationship between these processes is, like that between the coupled oscillators that govern locomotion, continuous and reciprocal, and there is no clear temporal or causal precedence between them.

Think of a dog out on a walk in the countryside. Tail wagging excitedly, the dog eschews the centre of the path followed by his owner. Rather, he flits from stump to clump, from rock to puddle, head down and nose flicking here and there, exploring a world of scent. Here we have a combination of sensory and motor processes working closely as a team. Suppose the faint perfume of a rat's carcass reaches the dog's nose. This is the start of a chemical gradient leading to the dead rat itself. The dog ascends this chemical gradient with two lopes and a slight motion of the head, using olfactory feedback to guide his trajectory.

Numerous brain processes cooperate during this brief behavioural episode. The dog's snout is constantly in motion, and while exploring a scent he sniffs several times a second to make wafts of air pass over chemical receptors in his nose. As a result of these motor processes, the olfactory bulb is stimulated, a sensory process that generates a unique spatio-temporal pattern of neural activation for each category of smell.[25] This pattern influences the motor activity that governs where the dog goes. His head moves as he snuffles, and his four legs produce a varying gait depending on the speed of locomotion, which in turn depends on whether he has found the source of an enticing scent, is proceeding towards one, or is seeking out a new object of interest. For the duration of the behaviour, this coalition of processes is active and its members are continuously, reciprocally engaged. When the behaviour ends, the coalition breaks apart. Its component processes cease to operate in concert, and many or all of them become quiescent.

Now we can take a fresh look at behaviour selection. As recognized by many ethologists and biologically inspired roboticists, if input and output processes are dynamically coupled, then the unit of selection is not a motor process but a combination of sensory and motor processes. (Indeed, this is the rationale for using the term 'behaviour selection' rather than 'action selection' here, although the latter term is often used in the literature.) Arriving at the rat's corpse, the dog begins to investigate, his snout roving over the carrion surface like the gaze of an art-lover moving over a sculpture. Then he hears the familiar bark of another dog, and his attention is drawn from the dead rodent.

[25] Skarda & Freeman (1987); Freeman (1999), Chapter 4.

He looks up. It's a neighbour's spaniel with whom he is on intimate terms. Abandoning the rat, he strides towards his friend.

Here we see the outcome of a competition among coalitions of processes, in which one winner is superseded by another. In dynamical systems terms, the coalition of active processes realizing olfactory exploration can be regarded as an attractor in a very large neural state space. But the continuing existence of this attractor depends on the presence of a sensory cue (here emanating from the dead rat), which exceeds all its rivals in salience. When a contending stimulus arrives, namely the bark of another dog, the attractor landscape alters. If the contender is sufficiently salient, a phase shift occurs, wherein the old attractor disappears and is supplanted by its competitor. The upshot is that a new coalition of active sensory and motor processes forms and displaces the old one, a coalition that, in this case, realizes social behaviour. Over time, we find an ever-changing pattern of active processes reflecting an ever-shifting attractor landscape (Fig. 4.5). Stable coalitions form, linger for a while, then dissolve. In the periods of transition from one stable coalition to another there is upheaval, while rival attractors compete, until one of them emerges victorious and a new group of active processes crystallizes and becomes dominant.

So far, our neurodynamical description has made no reference to a global workspace. Indeed, a behaviour-based architecture can function perfectly well

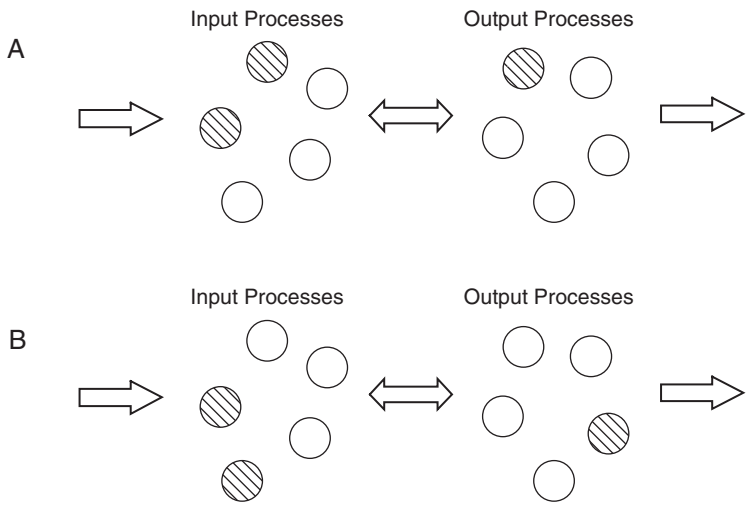

Fig. 4.5 The dynamics of behaviour selection. In response to a sensory cue, a dynamically coupled coalition of active (hashed) input and output processes forms, corresponding to an attractor of the neural state space (A). When a new cue arrives, the attractor landscape shifts, the old coalition breaks up, and a new one with different constituents forms (B).

without one. But a limitation of the unenhanced behaviour-based architecture is the rigidity of the behavioural repertoire it can support. Each possible coalition of processes is effectively a specialist, realizing a single, stereotyped sensorimotor programme tailored to a particular purpose. Every coalition that forms represents a tried-and-tested combination of processes, where 'tried-and-tested' means either innate or acquired through learning by reinforcement. Although the same input process might be able to participate in different active coalitions at different times, the set of combinations of processes that constitute viable coalitions is closed, not open-ended. In a strictly behaviour-based architecture, the set of partners with which a process can dynamically couple is fixed by the evolutionary or developmental context within which it originated. As a consequence, the repertoire of behaviours in such an architecture is also limited to the tried-and-tested.

According to the theory in favour here, increased flexibility is enabled if the influence of every process is allowed to permeate the whole system. Previously unseen process coalitions are then able to form, permitting a blending together of established sensorimotor patterns and thereby generating novel behaviour. This is where global workspace theory re-enters the picture. For the idea of a global workspace, though originally cast in architectural terms, is comfortably recast in terms of a communications infrastructure that achieves the required dynamics by allowing localized neuronal activity to exercise widespread influence. Indeed, unless recast this way, the putative workspace might be mistaken for a dedicated brain region, something akin to the Cartesian Theatre ridiculed by Dennett, a place in the brain where 'it all comes together and consciousness happens'.[26]

The most likely substrate for the sort of connective infrastructure that would realize such a global *neuronal* workspace, as Dehaene and colleagues call it,[27] is the web of long-range white matter pathways linking geographically separate cortical and sub-cortical regions. We shall return to the issue of neuroanatomy in due course, and we'll see how the brain's connectivity promotes the kind of neurodynamics required for systemic broadcast and open-ended coalition formation. In the mean time, equipped with a clearer conception of the dynamical milieu within which a global workspace would function, and guided by the just introduced notion that the global workspace operates by disseminating influence, we can articulate the idea, much alluded to already, that the conscious condition is *integrative*.

[26] Dennett (1991), p. 39.

[27] Dehaene, *et al.* (1998); Dehaene & Naccache (2001). Biologically realistic computer models of aspects of the proposed global neuronal workspace are described by Dehaene, *et al.* (1998; 2003), Dehaene & Changeux (2005), and Shanahan (2008a).

4.5 Integration and the conscious condition

Authors frequently argue for some form of intimate relationship between consciousness and integration, at the cognitive level,[28] at the neurological level,[29] or at the more abstract level of information theory.[30] How is integration to be construed in the context of global workspace theory? How does a global workspace realize integration, so construed? How does integration, so realized, subserve cognition, and how does it relate to phenomenology? The essential insight that answers each of these questions is this. Perfect integration occurs when the *being as a whole* is brought to bear on the ongoing situation. According to our scientific understanding, the 'being as a whole'—the complete person, we might say, or the complete animal—is a living brain, embodied and ecologically situated. Insofar as the brain is understood as a system of processes, the 'being as a whole' includes the *whole system* with its full cohort of processes. When the whole system, the brain with all its resources, bears on the ongoing situation, the response it offers is an *integrated* response. We might venture to say it is the response of a *unified subject*. The conscious condition, thanks to the global workspace, promotes integration in this sense.

If the whole system is brought to bear on the ongoing situation then various conditions conducive to cognition—conducive, therefore, to an animal's survival and well-being—will be met. For example, the system's full cohort of processes will be eligible to participate in the competition to determine its response. Moreover, every process in the full cohort will have the opportunity to contribute to the overall coherence of the system's response. Also, the various ways in which the ongoing situation might exercise deferred influence on the system's behaviour—through motor or declarative learning, for example, or through working memory or through episodic memory—will be simultaneously enabled.

Violations of these enabling conditions are apparent during lapses of attention. When a person absentmindedly pulls out his own front door key when arriving at the house of a friend,[31] habit has prevented the possibility of knocking on the door from taking part in the competition to determine his actions. When a distracted amateur plumber removes the U-bend from her sink and makes a mess on the floor by pouring its contents down the very same sink, her distraction has denied the process that would have predicted this outcome its

[28] Baars (1988; 1997).

[29] Varela, *et al.* (2001).

[30] Tononi, *et al.* (1998); Tononi (2008); Balduzzi & Tononi (2008). See also the review by Seth, *et al.* (2008).

[31] This example is lifted from James (1890/1950, vol. 1, p. 115).

powers of veto. When someone tidying a room while thinking of his plans for the following day unconsciously moves his spectacles onto a shelf, and is later unable to recall their whereabouts, inattention to the task at hand has disabled the memory processes that would otherwise have come to his aid. In each of these cases we are outside the conscious condition. What is brought to bear on the ongoing situation is less than the whole system, and in a sense what we encounter is less than the complete person.

To offer an integrated response, wherein the whole system is brought to bear on the ongoing situation, all potentially relevant processes in the system must be subject to the influence of that situation, and the system's response to it must take account of the activity of all potentially relevant processes. Dennett has likened the conscious condition to 'fame in the brain'[32] and his metaphor is compatible with the present conception. First, it conveys the idea that the major distinction between the conscious and unconscious processing of a sensory event is the difference between merely local and fully global influence. A consciously perceived stimulus is 'famous' in the brain because, so to speak, 'everyone has heard about it'—all the potentially relevant brain processes, whatever their provenance—working memory, episodic memory, language, affect, and so on.

Second, Dennett's metaphor alludes to the fact that the conscious condition can only be ascribed to an event in retrospect, in much the same way that a noteworthy achievement (such as an archaeological discovery) can only be considered famous when news of it gets around and people start talking about it, writing about it, and so on. Something could occur after the event in question that prevents it from becoming famous (the sensory event is masked, the archaeologist dies before revealing his finds to the world), so only hindsight can make the proper judgement.[33] In a similar vein, system-wide influence is a property that can only be attributed retrospectively. Time has to elapse before any potential influence can be realized, and if events intervene it might never be realized.

Moreover, even if every process that is potentially relevant to the ongoing situation is subject to its influence, and even if the system's response to that situation takes account of the activity of every such process, only a fraction of those processes will ever actually have an effect on its outward behaviour. But this raises an interesting question. Do those processes that are potentially relevant to the ongoing situation, subject to its influence, and duly taken account

[32] Dennett (2001).

[33] The metaphor of fame is not perfect. There is only a short, limited window for a sensory event to make it to consciousness, whereas achievements like archaeological discoveries can become famous decades after they take place.

of, but that never actually perturb the system's outward behaviour in the smallest degree, nevertheless contribute to a person's phenomenology? After all, we seem to be claiming that their involvement is constitutive of the conscious condition, so contribute they surely must. Yet it's hard to see what difference it would make if they were simply deleted. Does it really count as fame if lots of people hear of an achievement, but nobody talks about it, writes about it, or even thinks about it, ever again?

To sharpen the point, let's consider a scenario, of the sort much loved by philosophers, in which a person sees a patch of red. Perhaps she is standing before a painting by Rothko, her gaze fixed on a tiny portion of the canvas. The sight of the red patch, especially within the larger context of a gallery visit, engages a large repertoire of potential responses. It might cause her to turn away in boredom, or to step back and view the canvas from further away. It might remind her of menstrual blood, or of a traffic light. Later, the image lodged in her mind, she might photograph the sunset, or read more about modern art. But only a tiny minority of the processes that have the potential to influence her behaviour actually get to exercise that influence. Not only does she not in fact turn away in boredom (her companion is an admirer of Rothko), the idea never even occurs to her. The relevant processes make no contribution to her behaviour, and we might ask what difference it would make to her phenomenology if they were absent altogether.

Yet according to Tononi's 'information integration' theory, not only do relevant but ineffectual processes contribute to phenomenology, even wholly irrelevant processes make a contribution.[34] One of the reasons our art lover enjoys a rich experience of colour in her moment of reverie—unlike, say, a red-sensitive photodiode sited in front of the same painted patch—is that her brain can discriminate this situation from a multitude of others, and has a huge and delicately nuanced repertoire of responses to match. In seeing this particular shade of red, she is also, as it were, seeing not-crimson, not-scarlet, and so on, and her repertoire of potential responses is sensitive to these differences. (Her experience is surely richer, thanks to her capacity to discriminate these shades, than that of someone who cannot tell one shade of red from another.) On a coarser level, her brain can discriminate a glimpse of Rothko from the sound of thunder, the taste of strawberries, a tingling in the feet, and so on, and offers a distinctive range of responses for each sensation. All this, so the argument goes, is included in the totality of experiencing the red patch.

Is it possible to accommodate this holistic intuition with the theory under development here? What does the present claim—the claim that the conscious

[34] Tononi (2008); Balduzzi & Tononi (2008).

condition enables the system as a whole to be brought to bear on the ongoing situation—tell us about phenomenology? To begin with, we must reject the overly metaphysical conception of consciousness that insists on a precisely definable content to a subject's consciousness at any given time, and that the question of that content always has an answer. Instead, we must accept that there is something of the refrigerator light illusion about our inner lives.[35] Whenever we ask ourselves whether we are aware of what a stimulus is not, whether we are aware of the possibility of responding to it in a particular way, or whether we are aware of a given association with it, lo and behold these things are before us. But it is the very act of asking the question that brings them forth. It seems as if the full range of possibility is present to consciousness all at once, indeed as if this plenitude is constitutive of the conscious condition. But this is misleading. A feeling of plenitude is not a plenitude of feeling.

If anything can instructively be said to be 'constitutive' of the conscious condition, it is the means by which the illusion is realized, the mechanism that switches on the light whenever the refrigerator door is opened, so to speak. That is to say, it is the means by which all those possibilities for thought or action are made available even when they are not actualized. According to the present proposal, in a system that comprises numerous interacting processes, this is achieved by a connective infrastructure that enables a process or coalition of processes to have a systemic influence on the whole, while the system as a whole can exercise influence on each of its constituent processes.[36] This is what we are calling the global workspace. The conscious condition is integrative thanks to the existence of this infrastructure, and it is thanks to this infrastructure that, in the conscious condition, the whole person is brought to bear on the situation at hand, whether confronting a dangerous predator, fashioning a tool, or reflecting on what it means to be human.

4.6 **Influence and information**

The pertinent question to ask next is what sort of connective infrastructure is best suited to disseminating influence in the way we have envisaged. That is to say, what is the topology of the network linking together the system's component processes? But in order to address this question, we need to be more precise about what is meant by 'influence', and how it relates to connectivity. To begin with, the concept of the influence one system (or process) has on

[35] We already encountered this metaphor in Section 3.3, but its use here is somewhat different.

[36] This is an example of what Wheeler (2005, Chapter 10), after Clark (1997, Chapter 8), calls 'continuous reciprocal causation'.

another should be contrasted with the idea of a *message* one system (or process) might send to another. The latter term is deliberately not used here because it has semantic overtones that are absent from the former term.[37] To understand this choice, let's rejoin the art-lover standing before the Rothko canvas, absorbed in her awareness of a red patch.

Now, it's tempting to speak here of the 'content' of her consciousness during the episode in question, content in which redness figures large. After all, isn't she supposed to be conscious *of* red? Surely our theory ought to account for this 'of', for the *intentionality* of consciousness, in other words. But it would be a mistake to succumb to this way of thinking. There is no need to accept (or to reject) the philosophically difficult notion of 'content' in order to account for the conscious condition. According to global workspace theory, the red patch participates in the art-lover's conscious condition and contributes to her inner life because the whole system that comprises her living brain, embodied and ecologically situated, is open to influence by the sensory processes that are responding to the red visual stimulus, and reciprocally, the system as a whole is able to influence the various motor processes that govern her potential responses, now or later, to that visual stimulus. Nothing pertinent is omitted from this initial characterization.

However, elsewhere we have spoken informally of the traffic of *information* around the global workspace architecture. Is it not the case that the term 'information' alludes to semantics in the same way as the rejected term 'message'? Well, as we shall now see, the term 'information' helps to enrich the concept of influence, and its usage can be clarified while keeping the potential for philosophical controversy to a manageable level. According to Bateson's useful aphorism, information is 'a difference that makes a difference'.[38] The first sense of 'difference' here—the difference that does the making rather than the difference that is made—is that of a *distinction*. The simplest possible distinction is a binary one, the distinction between Yes and No or between 0 and 1. This is a notion that sits well with Shannon's mathematical theory (which quantifies the amount of information in a signal in terms of bits).[39]

The second sense of 'difference'—the difference that is made—is causal, and relates directly to the notion of influence we have been employing. A signal carried in a wire that goes nowhere and is connected to nothing is without influence, like a 'flower born to blush unseen'. It is a difference that makes no difference. On the other hand, a train of spikes that issues from the primary

[37] This is not so in computer science, where the term 'message' is more neutral.

[38] Bateson (1972), p. 459.

[39] Shannon & Weaver (1949).

visual cortex of a driver's brain and causes him to put his foot on the brake is a difference that makes a difference. In what follows, we shall say that there is a (direct) connection from a process A to a process B if A can influence B without the mediation of any third process. The simplest way one process can influence another is to switch it on or off, to make it active or inactive. But in general, A's influence on B is more subtle. Process A might modulate B's activity by degrees, or it might nudge it into any one of a large repertoire of attractors or metastable states. In these cases, the connection between A and B conveys more than a single bit. It is a channel for information in the sense just characterized, a channel for a range of distinct signals that can make a range of differences to B's activity.

Recall the team of coupled sensory and motor processes at work in the brain of the dog on a country walk. The signals passing from his olfactory cortex to his motor areas convey a great deal of information. A hint of dog-scent here, the moist grass, the hawthorn blossom, a waft of carrion there—each of these has its own characteristic signature, its own combination of spatial and temporal features, reflecting the range of distinctions the signals can bear. Moreover, distinctive signals can exercise distinctive influences. The scent of another dog makes him look up, the waft of carrion determines the direction of his stride. Dynamically speaking, what we have is a system of coupled processes. But the coupling is usefully thought of as effected by the exchange of information, as long as we are careful not to taint our conception of information with semantics.

We previously characterized the conscious condition in terms of the reciprocal influence between the system as a whole and its constituent processes. Now, with the notion of influence upgraded to accommodate the exchange of information, we can do justice to the phenomenological platitude that our consciousness is richly contentful, without becoming mired in the sort of philosophical controversy that an overly metaphysical attitude towards that observation brings about. The next step in this direction is to extend the enriched idea of one process's influence on another to that of its influence (or potential influence) over multiple processes, and by extension to its influence over the whole system of processes of which it is a part. According to our enriched conception, the systemic influence in question will be mediated by the transmission of information, and this is as much explanation as we need of the fact that the conscious condition is rich.

So far so good. But there's a problem. Surely, we might say, in this conscious condition, the *same* information is broadcast to every recipient process. If a person experiences, say, a tickling sensation, then what the conscious condition simultaneously enables that person to talk about, act on, remember, and so on, is surely, in each case, *that* tickling sensation. But what meaning can we attach to this 'same'? What are the identity criteria for information, under

our construal? We might venture an answer that appeals to the identity of the signals sent to all the processes. The information is the same because the signals are the same. But this would be a mistake. There is no reason to suppose that a brain area sends the same signals to all the other brain areas to which it is connected. Moreover, our original conception of influence was pleasingly neutral, and even the enriched version presupposes nothing more than signals that are capable of conveying a variety of distinctions. It would be good to preserve this neutrality as we extend the idea to multiple simultaneous influences.

The important thing to remember here is that there is no meaning in the signal itself, only spatiotemporal structure. What matters is the way the signal is received, how that spatiotemporal structure affects the activity of the receiving process. The urge to impose semantics on signals is what tempts us to think we need to establish identity criteria for the information that mediates the influence of one process on another. But we must not be seduced by the workspace metaphor into thinking of processes as homunculi that speak and hear and stand in need of a *lingua franca*. If a process A influences two other process B and C then the nature of this influence can be thought of as mediated by information if 1) a variety of signals can pass from A to B and from A to C, and 2) the responses of B and C are sensitive to this variety. We might expect a given pattern of activation in A to influence B the same way on one occasion as on another, and equally to influence C the same way on one occasion as another. But, the signals going to B and C from A do not have to be the same. There is no need for further stipulations. The specific structure of the signals drops out of the equation. All that counts, for characterizing the conscious condition, is the various influences they have.

4.7 **The right connections**

With the relevant senses of influence and information duly clarified, we can address the question of connective infrastructure. The appropriate technical vocabulary for this job comes from the study of complex networks, and a short introduction to the relevant mathematical concepts will be needed before we can proceed. A network is a set of *nodes* joined by *arcs*. We shall deem a node to be a brain process at the lowest level of organization that is explanatorily pertinent. There's no need to commit to exactly where the lowest pertinent level lies, but it is clearly above that of the individual neuron and below that of a brain region. An arc is a connection between processes, that is to say a channel that mediates the direct influence of one process on another, a channel for information, suitably construed. To simplify the presentation, we'll assume that arcs are undirected and unweighted, even though the influence between processes sometimes goes one way only and in general admits of degree.

The formal concepts to be deployed are easily generalized to the directed and the weighted cases.

Just as brain processes can be identified at many levels of organization, from handfuls of neurons to large-scale cortical structures, so connections between brain processes can be identified at many levels of organization. But as processes are to nodes, so connections are to arcs, which is to say an arc in the network is deemed to be a connection at the lowest explanatorily pertinent level of organization. We need not take a stand on the underlying mechanism of connection. A physical pathway—a fibre tract—between two brain regions is evidence of a connection. But the existence of a physical pathway is no guarantee that it mediates influence. Correlated activity, as evinced by fMRI data, is a more direct index of active influence, but it is no indicator of potential influence. However, these empirical difficulties can be set aside in the present discussion.[40]

A feature common to many complex networks, both natural and man-made, is the *small-world* property. As we shall see, small-world connectivity also promotes the kind of integrative dynamics we are interested in here. Intuitively, a network is small-world if 1) it is densely connected at a local level, 2) it is sparsely connected at a global level, and 3) it is (typically) possible to move from any given node to any other given node in just a few hops. The link structure of the world wide web enjoys this property, as do human social networks (hence the phrase 'six degrees of separation'—six being the average number of steps in the global social network allegedly required to connect any two people anywhere in the world).

The small-world property was given a precise mathematical characterization by Watts and Strogatz in the 1990s.[41] Consider a network G comprising a set of nodes and arcs. The *path length* between any pair of nodes in G is the number of arcs in the shortest path between those nodes, and G's *mean path length* is the path length averaged over every pair of nodes in G. The *clustering coefficient* of a node P in G is the fraction of the set of all pairs of immediate neighbours of P that are joined by an arc, and the *clustering coefficient* of the whole network G is the clustering coefficient averaged over the set of all nodes in G. A small-world network is one that is sparsely connected overall, but has a low mean path length and a high clustering coefficient. More precisely, its mean path length should be close to that of a comparable random network, but its clustering coefficient should be significantly higher, where a 'comparable

[40] See Bassett & Bullmore (2006) and Bullmore & Sporns (2009) for surveys of established structural and functional connectivity findings – the latter paper in particular for a discussion of the extent to which they match.

[41] Watts & Strogatz (1998).

random network' is one with the same number of nodes and arcs, but where the arcs are randomly assigned with uniform probability.

As well as providing this formal characterization, Watts and Strogatz described a method for constructing small-world networks which is an aid to both mathematical study and intuitive understanding. Their procedure is as follows (Fig. 4.6, left). First, a 'ring lattice' is constructed, that is to say a set of nodes arranged in a circle, each of which is joined by an arc to all of its neighbours that are k nodes or fewer away. Then, each of these arcs is 'rewired' with probability p. To rewire an arc means to unfasten one of its ends and reconnect it to a random node anywhere in the ring. Even if p is small, the few rewirings that result are sufficient to confer the small-world property on the network. To see this, consider the effect of a single long-range rewiring on the ring lattice. This will have negligible impact on the network's overall clustering coefficient, as it only reduces the clustering coefficient for a single node. Yet by introducing the possibility of making a giant hop across the network, it will reduce the shortest path between many pairs of nodes, in some cases significantly.

Certain attributes of a small-world network should be apparent from this exemplar, attributes that are potentially beneficial whatever real-world system the network represents. A low mean path length promotes the rapid global spread of indirect influence, and a high clustering coefficient is a prerequisite for localized activity, whereas sparse overall connectivity is required to keep

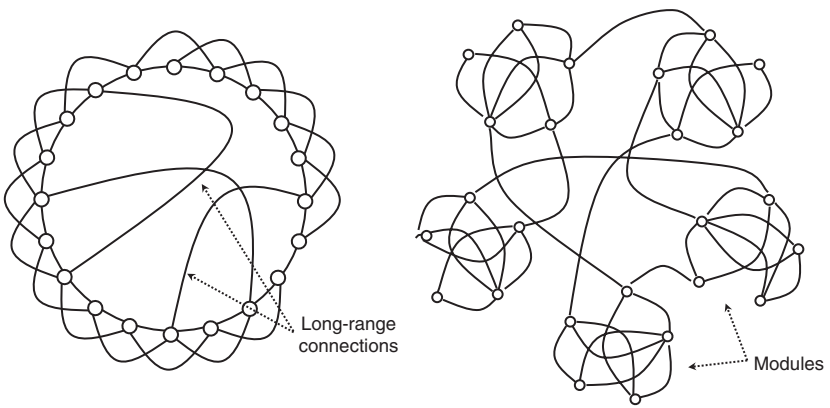

Fig. 4.6 Two kinds of small-world network. Left: a network constructed using the Watts–Strogatz procedure. Starting with a ring lattice, a number of long-range rewirings are made. Each rewired connection provides a shortcut across the lattice, reducing the average path length without significantly reducing the overall clustering coefficient. Right: a modular small-world network. A number of densely intra-connected modules are connected to each other with a sparse set of inter-module links.

down wiring cost. However, the network topology generated by the Watts–Strogatz procedure is only one of many that exhibit small-world properties, and the homogeneous connectivity of its underlying lattice structure is not well suited to our present enquiries.

A different kind of small-world network is shown on the right of Fig. 4.6. This network can be constructed in two phases.[42] First, a set of nodes is partitioned into distinct clusters (or 'modules' or 'communities'), such that each cluster is densely interconnected but isolated from all other clusters. This (disconnected) network will have a high clustering coefficient, despite the isolation of the clusters, thanks to the internal connectivity of the clusters alone. Second, the clusters are connected to each other with just a few randomly chosen cluster-to-cluster connections. These will be sufficient to guarantee a short average path length in the network, and to confer the small-world property on it. The result is a small-world network with *modular* or *community* structure.[43]

Not all modular networks have the small-world property. Here's a counterexample. Imagine a set of modules arranged in a line, like a string of pearls, so that each module is externally connected only to its neighbours. Suppose there are N modules (N pearls on the string). Then, for all pairs of nodes, the average number of trans-modular hops required to get from one node to the other is $N/2$. So it's easy to construct a network that has a high mean path length, because we can make N as large as we like without compromising the network's modular structure. However, many modular networks do inherit the small-world property. Consider any modular network G, and imagine extracting a higher-level network H, whose nodes are the modules of G, such that there is an arc between two nodes P_1 and P_2 in H if and only if there is an arc between a member of P_1 and a member of P_2 in G. Now, if H has a low mean path length and each of the modules in G is a small-world network, then G itself is a small-world network. Less formally, for a modular network to be small world, it should be possible, on average, to go from any module to any module in just a few hops.

If many paths from nodes in one module to nodes in another pass through the same intermediate node then that node is known as a *connector hub*. Using the method outlined earlier, there is no statistical reason connector hubs should emerge, and there are none in the example network on the right of Fig. 4.6. By contrast, the network on the right of Fig. 4.7 has a connector hub in each module. Non-local traffic in the network is obliged to pass through the

[42] The method described here is that of Shanahan (2008b).

[43] Girvan & Newman (2002); Newman (2006); Müller-Linow, *et al.* (2008).

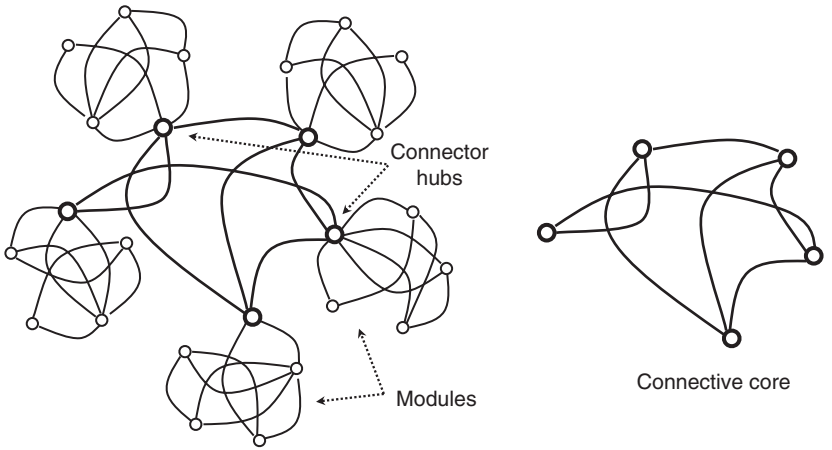

Fig. 4.7 A modular small-world network with connector hubs. Connector hubs (outlined in bold) are nodes that have high trans-modular connectivity. The 'connective core' is the connector hubs plus the arcs joining them.

connector hubs, which can be likened to major junctions in a road network. By extracting the hub nodes and their interconnections from the rest of the network (the junctions plus the motorways), we isolate its *connective core* (Fig. 4.7, right). The connective core of the human brain's structural network will be a major focus of interest in the material to come.

A further refinement of modular structure allows for hierarchical organization. Fig. 4.8 shows three levels of hierarchy, but obviously this can be extended to any number. In a hierarchically modular network, hub nodes within a module—that is to say nodes that join sub-modules to each other—are known as *provincial hubs*. Like flat modular networks, hierarchically modular networks have the small-world property if their inter-module connectivity meets the right conditions. Dense local connections ensure a high clustering coefficient. But in a network with hub nodes, a short path between any two nodes exists involving a few hops—out of a sub-module to the nearest provincial hub, then to a connector hub and across to the destination module, and finally to the target node via another provincial hub. The story is the same if additional levels feature. Thanks to the hierarchical structure, mean path length increases logarithmically as node count goes up.

4.8 **The anatomy of a global workspace**

Recall that we were seeking a connective infrastructure that would support a particular kind of integrative dynamics. It should enable a process or coalition

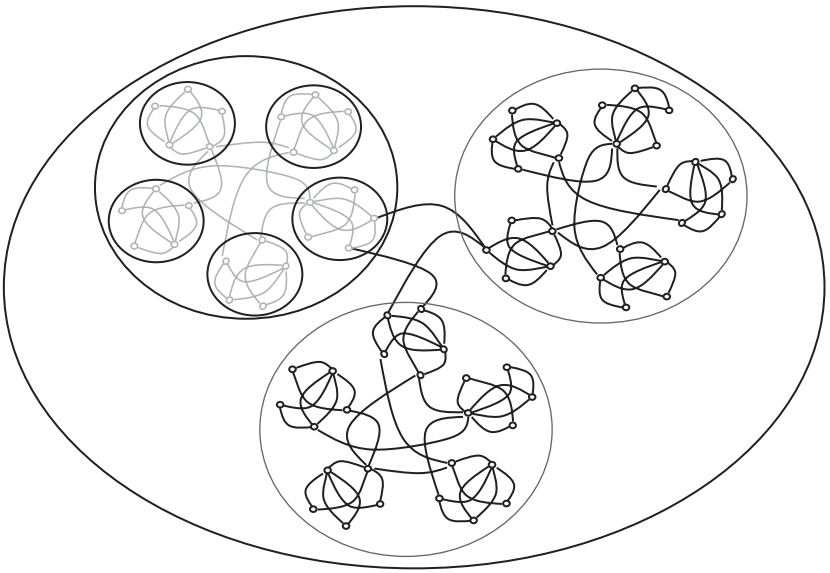

Fig. 4.8 A small-world network with three levels of hierarchical modular organization. Ellipses are drawn around modules to emphasize their hierarchical structure.

of processes to exercise influence on the whole system, and at the same time it should enable the system as a whole to influence each of its component processes. We now have a clearer language for framing relevant empirical questions. In particular, we should be on the lookout for brain networks with a short mean path length, one of the three defining characteristics of a small-world network, as this is conducive to the efficient system-wide propagation of influence and information. Sparse overall connectivity, another defining characteristic of a small-world network, can be taken for granted. This is because any network in which connections are costly is bound to be sparsely connected (unless it contains trivially few nodes). The wiring costs of an animal's nervous system are numerous, and include the space needed to house it and the energy required both to use it and carry it around.[44]

What about the third defining characteristic of a small-world network, namely a high clustering coefficient? Well, the expectation of clustered connectivity is implicit in the very idea of the brain as a system comprising processes at multiple levels of organization. To see this, consider what it means to be a process in such a system. Although, according to the theory we are pursuing, a process must be amenable to the influence of the system of which

[44] Striedter (2005), Chapter 7; Bassett & Bullmore (2006).

it is a part, the notion of a process as an identifiable, separate entity entails a countervailing tendency to behave independently from the rest of the system.

This balance of integrated and segregated activity should be apparent at every level of organization. Processes at the lowest level of organization are the nodes of our network, so they behave independently by definition. But if a process is constituted by a large number of sub-processes, then, to be capable of independent behaviour, that set of sub-processes must be significantly more tightly coupled internally than externally. That is to say, the channels of influence among sub-processes belonging to the same process at a higher level of organization should be more densely interconnected than the connections among sub-processes belonging to different processes at that higher level of organization. It follows directly from this, not only that the communications infrastructure we're looking for should have the small-world property, but also that it should be hierarchically modular.

The next item on the agenda is the relevant empirical data. Is there evidence for the existence of a communications network in the brain with the right topological properties? The answer is yes. The literature is extensive. It begins with Watts and Strogatz, who proved that the nervous system of the nematode worm *C.elegans* is a small-world network.[45] Despite its small size, this network even exhibits a degree of modularity.[46] But more significantly for our investigation, the brains of mammals have been shown to enjoy the small-world property at multiple scales. At the low end of the scale, the neuron-to-neuron connectivity of a small patch of cortex is small-world.[47] However, it is the region-to-region connectivity of the mammalian brain as a whole that is most interesting in the present context.[48]

To get to grips with the details requires a short review of the relevant neuroanatomy. The cerebral hemispheres of the mammalian brain comprise both grey matter and white matter. The grey matter of the cortical surface is a convoluted, laminated sheet of neurons connected by local, short-range dendritic and axonal fibres. The cortical sheet can be subdivided into numerous distinct

[45] Watts & Strogatz (1998).

[46] Reigl, *et al.* (2004).

[47] Sporns & Zwi (2004) proved that networks built according to empirical data collected by Braitenberg & Shüz (1998) and Hellwig (2000) are small world. Moreover, Shefi, *et al.* (2002) reported that neurons cultured *in vitro* tend to self-organize into small-world networks. A similar effect has been demonstrated in computer models (Rubinov, *et al.*, 2009).

[48] Sporns & Zwi (2004) used connectivity matrices acquired by Felleman & Van Essen (1991), Scannell & Young (1993) and others to demonstrate the small-world property for macaque and cat cortex.

regions on the basis of anatomical and histologically identifiable boundaries. The white matter, on the other hand, contains no neurons, but comprises dense bundles of long-range axonal connections joining distant brain regions. These long, white matter axons are coated with a sheath of myelin, which gives them a lighter appearance under the scalpel than the surrounding grey matter. This myelin sheath ensures that electrical spikes travel much faster in the white matter than in the short, unmyelinated axons of the grey matter, permitting rapid communication even between remote cortical regions.[49]

There are three major classes of human white matter connection.[50] First, the *corpus callosum* carries traffic between the two cerebral hemispheres. Second, the *corona radiata* relays signals to cortex from the thalamus and back to the thalamus from cortex (Fig. 4.9, right). Third, several fibre tracts directly connect remote parts of cortex to each other (Fig. 4.9, left). The major cortico-cortical tracts include the superior longitudinal fasciculus, which bridges the occipital and frontal lobes, the inferior longitudinal fasciculus, which bridges the occipital and temporal lobes, the superior and inferior fronto-occipital fasciculi, which (obviously) join the occipital and frontal lobes, and finally the

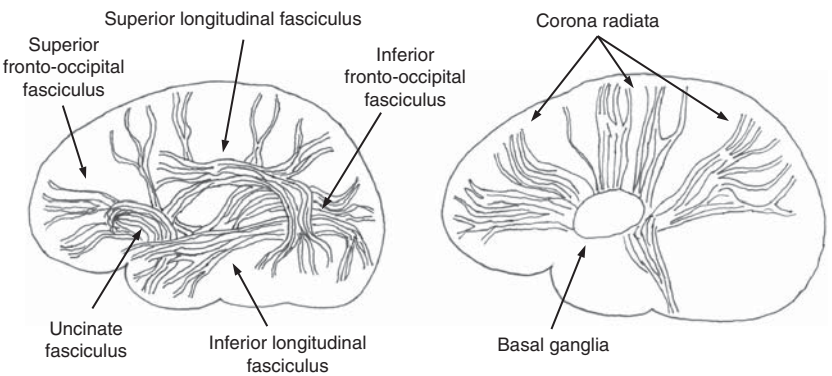

Fig. 4.9 White matter tracts (loosely based on Wakana, *et al*. (2004)). Direct cortico-cortical pathways provide rich connectivity between the frontal, temporal, and occipital lobes, and are a likely substrate for a global neuronal workspace (left). Thalamocortical pathways relay traffic to and from different parts of cortex, and may also play a role in the putative workspace (right). (The thalamus is not shown, but is behind the basal ganglia.)

[49] See Schmahmann & Pandya (2006) for a comprehensive overview of the history of the study of white matter in neuroscience, and for a detailed tracer study of white matter anatomy in the rhesus monkey. For a view of human white matter functionality, see Fields (2008).

[50] Wakana, *et al*. (2004).

uncinate fasciculus, which arches between the temporal and frontal lobes. Taken together with numerous smaller tracts, all this constitutes a significant amount of wiring with the potential to link all parts of the brain together in a single wide-area network, a kind of pan-cortical web.

Using white matter connectivity matrices established by traditional neuro-anatomical methods, it has been shown that the cortices of cats and macaques are small-world networks.[51] For humans, diffusion-based imaging techniques enable neuroscientists to build detailed white matter atlases *in vivo*, and the topological properties of the resulting connectivity matrices are then open to analysis. Not only do they exhibit the expected small-world properties, they also manifest a modular organization.[52] In one study, by Hagmann and colleagues, six modules were revealed with clear anatomical loci, each comprising some subset of 66 anatomical sub-regions. Connections between sub-regions were weighted according to the density of fibres connecting them (averaged over several subjects), and the resulting network was partitioned into highly clustered sub-networks. The set of modules found comprised a frontal module for each hemisphere, a mainly posterior module in each hemisphere that included a small number of more frontal sub-regions, a bilateral posterior medial module, and a central bilateral module (Fig. 4.10).[53]

In the same study, a set of 12 connector hubs were also identified, and these were found to lie along a medial cortical axis running from front to rear and comprising many of the most highly weighted arcs in the network (Fig. 4.11, left). An amendment to the definition of a network's 'connective core' that

[51] Sporns & Zwi (2004); Bassett & Bullmore (2006).

[52] Hagmann, *et al.* (2008); Iturria-Medina, *et al.* (2008); Gong, *et al.* (2009); Bullmore & Sporns (2009). Prior to the structural connectivity results reported in those papers, Eguíluz, *et al.* (2005) revealed a network of functional brain connections, using fMRI, that conform to the power law characteristic of a scale-free, small-world network, and Achard, *et al.* (2006) provided a connectivity map of the associated cortical hub nodes. Preliminary results on human structural connectivity were obtained by He, *et al.* (2007), who correlated measures of cortical thickness to establish the likely presence of a white matter pathway. Chen, *et al.* (2008) used the same method to establish modularity. Hierarchically modular structure has been uncovered in functional connectivity studies (Ferrarini, *et al.* (2009)). For a detailed overview of these and similar findings, see Sporns (2010).

[53] Hagmann, *et al.* (2008). This study leaves out certain important sub-cortical structures, so must be considered provisional. Notable omissions include the thalamus, which is accorded a significant role in cortico-cortical communication by Sherman and Guillery (2005), the basal ganglia, which play an important part in cortical competition according to Redgrave, *et al.* (1999), and the amygdala. The first two structures were included in the study of Iturria-Medina, *et al.* (2008), who found that the putamen (part of the basal ganglia) had a significant connective role.

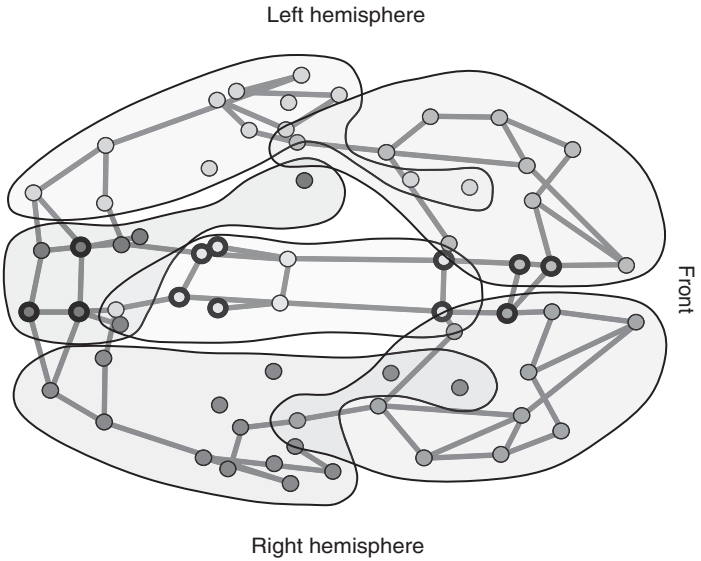

Fig. 4.10 Structural connectivity in the human brain (adapted from Hagmann, *et al*. (2008)). Each node corresponds to a cortical sub-region. Nodes are grouped into six colour-coded modules, where a module is a set of nodes that is more densely interconnected internally than externally. The blue lines are among the most prominent connections within and between modules. Note the pronounced medial axis. Connector hubs are outlined in bold. (See Plate 3).

takes account of weighted arcs can be used to characterize this medial axis.[54] Let *G* be a weighted modular network with hub nodes (such as the one we have here). A node is included in *G*'s connective core either if it is one of *G*'s hub nodes or if it is connected to one of *G*'s hub nodes by an arc whose weight is above a given threshold θ. There is an arc between two nodes in *G*'s connective core if there is an arc between those nodes in *G* whose weight is above θ. Now, having had a glimpse, thanks to neuroscience, of what may be the human brain's connective core, we can advance a two-part hypothesis about the possible anatomical locus of a global neuronal workspace.

The hypothesis is that 1) the brain embeds a network with a pronounced connective core that is capable of globally disseminating the influence of a process or coalition of processes, and 2) only one coalition of processes at a

[54] In the course of identifying the brain's 'structural core', Hagmann, *et al*. computed the network's 'connectivity backbone', defined as the maximal spanning tree (maximizing arc weight) plus a set of arcs whose weights exceeded a given threshold (Hidalgo, *et al*., 2007). The 'connective core', according to the present definition, is a subset of the 'connectivity backbone' defined by Hagmann, *et al*.

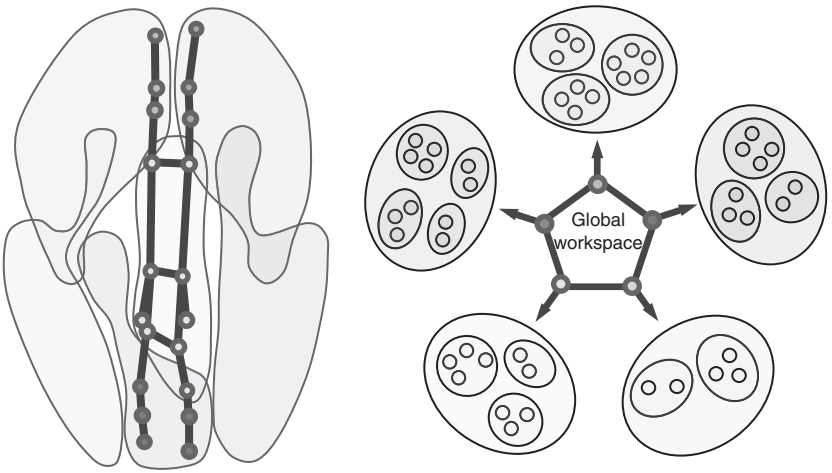

Fig. 4.11 The brain's connective core as the possible locus of a global neuronal workspace. The 'connective core' of a modular network comprises its connector hubs and the major arcs associated with them. There is evidence that human cortex has a connective core that runs along its medial axis (left). The connective core of a hierarchically modular network is topologically well placed to realize both broadcast (because influence funnels into and fans out from the centre) and competition (because it acts as a limited capacity bottleneck). So it is an ideal candidate for the anatomical locus of a global workspace (right). Note that the modules are indicative only, and that the modules on the right do not map directly to those on the left. (See Plate 4).

time can take over the connective core, to the exclusion of its rivals. It should be clear that, in general, the connective core of a network is topologically well-placed to act as a medium of both broadcast and competition (Fig. 4.11, right). A network with a marked connective core has an inherently radial structure. Influence and information funnels in from the periphery and fans out again from the centre. Moreover, the connective core acts as a bottleneck, which enforces competition. Even allowing for a filigree of minor connections overlaid on the network's major arteries, the bulk of long-range network traffic is likely to traverse the core, whose capacity is limited. Note, however, that lesions to parts of the connective core, although likely to compromise its functionality, are not always catastrophic (i.e., causing coma or loss of consciousness), because there are typically multiple pathways between any two nodes.

The connective core revealed in the Hagmann study and depicted in Fig. 4.10 is broadly consistent with several other structural connectivity studies in which human white matter connectivity was established using diffusion-based imaging then analysed in terms of network theory.[55] Small-world properties are a

[55] Iturria-Medina, *et al.* (2008); Gong, *et al.* (2009).

robust finding, and topologically central sets of hub nodes and arcs are invariably found. Moreover, certain sub-regions along the medial cortical axis feature prominently in each of these studies. One such region is the *precuneus*, which is one of the most richly and centrally connected sub-regions of cortex. Notably, the precuneus has also been implicated by functional imaging studies in a wide range of high-level cognitive tasks, involving mental imagery, episodic memory retrieval, and first-personal perspective taking.[56] Cavanna and Trimble venture to comment that 'converging evidence therefore suggests that the precuneus may be involved in the integration of multiple neural systems producing a conscious self-percept'.[57]

The precuneus, along with portions of posterior cingulate cortex, medial frontal cortex, and inferior temporal cortex, is also a component of the brain's *default mode network*. This is a set of brain regions whose activity is low during task performance, but which exhibit correlated activity when the subject is resting and is not performing a cognitively demanding task.[58] Under these conditions the subject's mind is free to wander, and the brain's dynamics is predominantly internally driven and spontaneous.[59] The brain's default mode network is complemented by a *cognitive control network*, a set of regions—including portions of prefrontal cortex, anterior cingulate cortex, and several parietal regions—that exhibit correlated activity during novel or unfamiliar tasks that require attention and control.[60] The members of both these functional networks have high global functional connectivity, which is consistent with their participation in the conscious condition according to the present theory.[61] Indeed, they appear to represent two distinct modes of conscious activity. Process coalitions that are active during mind wandering or task independent thought draw their membership from the regions of the default mode network, whereas process coalitions that form to deal with difficult or novel situations draw their membership from the regions of the cognitive control network.[62]

[56] Cavanna & Trimble (2006); Buckner & Carroll (2007); Spreng, *et al.* (2008).

[57] Cavanna & Trimble (2006), p. 579. What they mean by 'a conscious self-percept' is not entirely clear.

[58] Raichle, *et al.* (2001); Greicius, *et al.* (2003); Damoiseaux & Greicius (2009); van den Heuvel, *et al.* (2009); Raichle (2010); Vanhaudenhuyse, *et al.* (2010). However, see also Margulies, *et al.* (2009), whose findings are based on a detailed anatomical subdivision of the precuneus.

[59] Mason, *et al.* (2007).

[60] Fox, *et al.* (2005); Cole & Schneider (2007).

[61] Cole, *et al.* (2010).

[62] There's no reason to suppose there are only two such modes. A more complete list might include dreaming, meditation, mystical experience, and other altered states of

All these data are highly suggestive. It may be the case that the human brain's connective core runs along the medial cortical axis (perhaps also taking in certain key sub-cortical structures that were excluded from the Hagmann study), and the precuneus may well be an especially significant component of the connective core. But it would be imprudent to assign too much importance to any single structure. The global workspace architecture is, more than anything else, a *distributed* architecture. The conscious/unconscious distinction it underwrites plays on the contrast between systemic and localized influences. So it is inherent in the theory that no single structure is the locus of the conscious condition. It would be equally unwise to load too much weight on a specific structural connectivity matrix, especially when the techniques for producing these matrices are still relatively new. Moreover, structural connectivity is best understood in the context of functional networks, such as the default mode network and the cognitive control network. For sure, a functional connection can only exist if the structural connections are there to support it. Yet functional connections can exist where there are only indirect structural links, and a direct structural connection might have a negligible functional role.

In due course, the human connectome (the underlying blueprint for human brain structural and functional connectivity) will be thoroughly mapped. But it's imperative that we abstract away from such empirical specifics if we are to arrive at a deep understanding of the conscious/unconscious distinction and the various properties of those two conditions. As discussed in Chapter 3, corvids exhibit many of the cognitive attributes associated with the conscious condition, such as the apparent ability to combine expertise from different domains to solve unfamiliar variants of a problem. Yet avian neuroanatomy presents a very different organization to that of the mammal.[63] Rather than stratified like the cortex of a mammal, the homologous portion of the avian brain is nucleated. Similarly, there is plentiful evidence for high-level cognition in the octopus, whose neuroanatomy diverges even further from that of a primate.[64] Currently there are no whole brain connectivity studies for birds or octopuses comparable to those recently carried out with humans. But there is no reason to expect detailed correspondences. Rather, whatever the species, we should be on the lookout for the right topological profile—a hierarchically modular organization with small-world properties and a pronounced connective core.

consciousness. As James (1902/1985) wrote: '… our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different. … No account of the universe in its totality can be final which leaves these other forms of consciousness quite disregarded' (p. 388).

[63] Emery & Clayton (2005); Güntürkün (2005).

[64] Edelman, *et al.* (2005); Edelman & Seth (2009).

Chapter 5

Neurodynamics

To complement the previous chapter's focus on connectivity, this chapter looks at global workspace dynamics. The challenge is to pin down the dynamical signature of the conscious condition against a backdrop of the continual formation and break-up of coalitions of coupled brain process. The characteristic of this signature is episodic broadcast punctuated by bursts of competition, where the competition in question is between rival process coalitions for control of the global workspace. But the key to the conscious condition, and the reason it subserves integration, is that it facilitates the generation of an open-ended repertoire of process coalitions thanks to the brain's connective topology. The chapter concludes with a short survey of empirical work favouring synchronous oscillation as the neural mechanism of coalition formation and the exchange of information among coalition members.

5.1 From connectivity to behaviour

Our purview here is the space of possible minds. What we seek is a theory that embraces human, corvid, and octopus, a theory that would even apply, in principle, to an extraterrestrial or an intelligent robot. In this chapter we return to the issue of dynamics, initially without committing to the low-level biological details, although the chapter concludes with an overview of some relevant neuroscientific evidence. Recall the earlier characterization of behaviour selection as the formation of coalitions of coupled input and output processes that are in turn coupled with the environment. Having established the empirical plausibility of a broadcast mechanism in the brain, we are in a position to accommodate the idea of broadcast within the dynamics of coalition formation and break-up. Most importantly, with a means of disseminating influence and information in place, the repertoire of possible coalitions is no longer limited to the merely tried and tested.

To illustrate this enhanced dynamics of coalitions in flux, let's imagine ourselves in the company of a commuter who has just arrived at Victoria Station in London, and is walking with the crowd across the station foyer. The commuter is on autopilot. This is the same journey he undertakes every day of his

working life. Every sight and sound along the way is familiar —pigeons poking at discarded scraps, announcements on the public address system, people in bright tee-shirts handing out leaflets —and he notices none of it. His actions are entirely habitual. Without thought, he adjusts his trajectory so that he passes to the left of a newspaper kiosk rather than the right, because it's less crowded there. At one point he brushes another commuter coming in the opposite direction and offers a quick apology. With hundreds of others he arrives at a set of steps and descends into the Underground.

If the omnipotent psychologist from Chapter 3 detains the commuter at this point and subjects him to interview, she will find him unable to report very much about his journey from the train to the Underground station. (Thanks to her special powers, she can command his attention, unlike the pusher of leaflets.) No, he says, he cannot recall which side of the newspaper kiosk he passed. In fact he doesn't know what newspaper kiosk she means. Nor does he remember apologizing to anyone about anything. There are always people handing out leaflets, but he doesn't remember speeding up to avoid one today, however bright her tee shirt was. The commuter by now is becoming impatient. He has a report to finish. So the omnipotent psychologist rewinds time, and replays the scene up to the moment the commuter sidesteps the newspaper kiosk. She intends to disrupt his journey.

This time, the crowd on the other side of the kiosk is uncharacteristically large and static. Something is clearly afoot. Slowing down, the commuter realizes that the Underground station is shut and that no one is being allowed down to the platforms because of some unspecified emergency. The commuter is irritated. He has a report to finish. What a waste of his precious time! The commuter considers his options. Should he wait? Or should he get a coffee somewhere and work on his laptop until the situation goes back to normal. While he's ruminating, someone pushes a leaflet towards him. She is rather attractive, he notices, and she is wearing a very bright tee shirt. He takes the leaflet, which is offering him cheap coffee. This clinches the decision. He opts for the coffee shop, and proceeds towards it, leaflet in hand.

In the coffee shop, just as the commuter is about to get his laptop out, the omnipotent psychologist conducts her interview. Yes, of course he remembers the large crowd outside the Underground station. The station was closed. Yes, he can describe the person who handed him the leaflet. It was a young woman with dark hair. Her tee shirt was bright yellow with a company slogan on it. Now, if the psychologist would excuse him, he has a report to finish.

Back in her laboratory, the omnipotent psychologist can review events. Thanks to a recent grant, her equipment has been upgraded. As before, through repeated replays and interventions of different sorts, she has been able to build

up a contrastive data set showing which of the sensory and motor events that contributed to the commuter's behaviour were conscious and which were unconscious (see Chapter 3). But now she also has a view of her subject's brain processes. As events unfold, she can see the extent to which each brain process is active, identify coalitions of coupled processes, and sample the traffic of information through the connective core.

Let's assume the psychologist (with our enthusiastic endorsement) has adopted the architectural blueprint in Fig. 5.1, which accords with the hypothesis that the brain's processes are organized as a hierarchically modular network whose connective core is the global workspace. (See also Fig. 4.11 (right)). A basic tenet of the theory is that behaviour, cognition, and the conscious condition can only be understood in terms of *anatomically distributed* coalitions of brain processes. But despite their anatomical distribution, coalitions are assumed to draw their membership from a small number of regionally defined super-modules that can also be given crude functional labels. So, at the topmost level of organization, we have five super-modules: a sensory module,

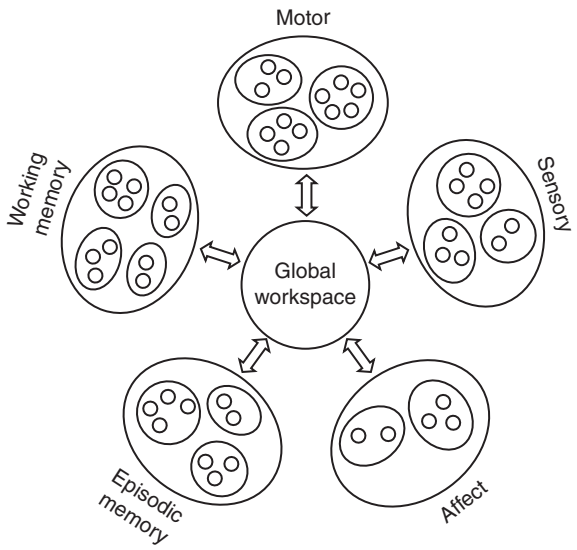

Fig. 5.1 A speculative functional decomposition that broadly conforms both to known anatomical divisions and to the modular structure revealed by Hagmann, *et al*. (2008). The lateralization prominent in the Hagmann study has been ignored, and a further module has been assumed that corresponds to the sub-cortical structures absent from their study (see Iturria-Medina, *et al*. (2008)). (Compare Friston (2003), Fig. 1.)

a motor module, an affect module,[1] an episodic memory module, and a working memory module. Rough anatomical assignments for these modules might be as follows: sensory module → occipito-temporal regions; motor module → fronto-parietal regions; affect module → sub-cortical structures including the amygdala and basal ganglia, plus orbito-frontal cortex; episodic memory → medial regions including entorhinal cortex and the hippocampus; working memory → prefrontal cortex. However, these assignments are not to be taken too seriously. Real networks, especially very large ones, have too much rich and subtle structure to be rendered so simply.

Now, when she inspects the commuter's brain, the psychologist will witness competitive, episodic coalition formation and break-up, plus the systemic broadcast of the influence of some (but not all) of the winning coalitions. Fig. 5.2 presents a representative series of snapshots based on the blueprint of Fig. 5.1. Each module includes a number of processes. Processes with common colours belong to the same coalition. The intensity of the colour indicates how active the process is. When the influence of one coalition begins to dominate the connective core—when that coalition wins access to the global workspace—the central circle takes on the corresponding colour. The outlying modules then take on that colour as the pattern of activation in the connective core starts to exercise systemic influence.

What the psychologist sees is a mixture of conscious and unconscious information processing. In Snapshot 1, three coalitions have formed, two of which (green and blue) are competing to control the global workspace. The third coalition (red) is not in competition with any others, and is not making a bid to access the workspace. This is the sort of picture the psychologist might see when the commuter is mid-way across the station foyer and nothing out of the ordinary has yet occurred. His behaviour is automatic. The coalition of red processes is guiding his actions. One of its member processes is keeping an eye on the usual landmarks in the scene that guide him towards the Underground station—the line of shops to the right, the departures board overhead, the Underground logo in the distance. Another process senses the ground every time he takes a step. Yet another process is keeping track of nearby people and their movements, while another process makes small adjustments to his trajectory to avoid collisions. The whole set of processes is mutually coupled and working smoothly in concert. The coalition is in turn coupled with the

[1] The role of affect in the present architecture is critical, and the subject of emotion demands a far lengthier treatment than given here. For a detailed overview of the relevant neuroscience, see Rolls (2005). (In the same book, Rolls also advances a theory of consciousness that emphasizes language and higher-order thought in a way that the present theory does not. But his treatment of emotion appears compatible with the present theory too.)

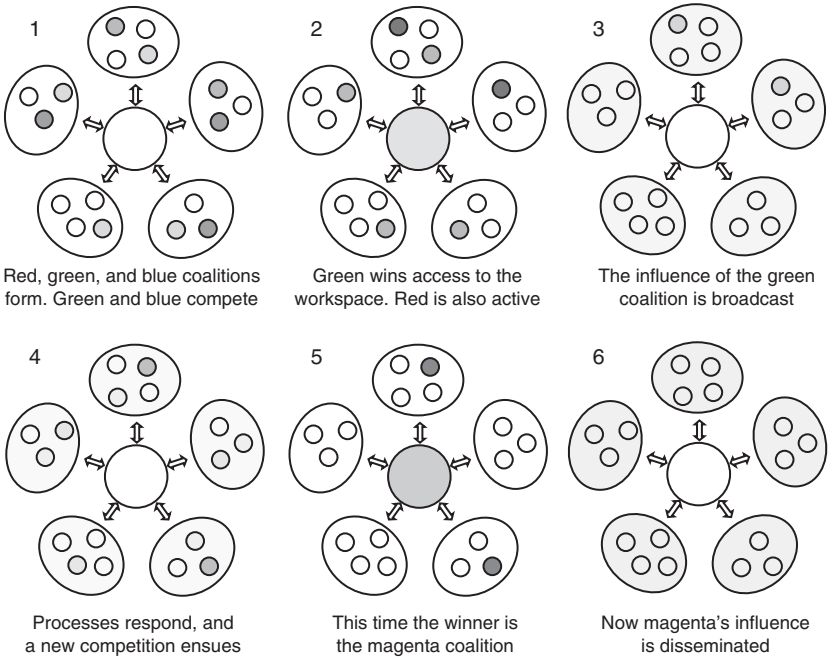

Fig. 5.2 The dynamics of competitive coalition formation and broadcast in the context of a modular network with a connective core (the global workspace). The figure shows a series of snapshots from top left to bottom right. Epochs of broadcast are punctuated by episodes of competition during which rival coalitions attempt to form. The winning coalition stabilizes and gains access to the global workspace (in the centre of each snapshot), from where its influence is disseminated to the whole system. This disrupts the dominant coalition and instigates a new competitive episode. (See Plate 5).

environment, and the upshot is the commuter's habitual journey across the station.

At the same time, the commuter has a report to complete. At the back of his mind, this irritating piece of unfinished business is nagging at him. This is the blue coalition, whose constituent processes compose draft sentences, generate fragments of imagery for the figures, or recall parts of the report that have already been written. (It would be no surprise if many members of this coalition were drawn from regions associated with the brain's default mode network.) But the blue coalition is losing out to the green coalition, which is also trying to compose sentences and generate imagery. For the commuter had a disagreement with his wife over the breakfast table, and the words that were exchanged keep coming back to him, along with look of anger in her face. The green coalition triumphs, and in Snapshot 2 we see that its influence has

invaded the connective core, the global workspace, excluding its rivals. By Snapshot 3, its influence has percolated outwards and can be seen throughout the brain. However, all this while the red coalition is still active, working in the background to keep the commuter walking (distractedly) towards the Underground.

The consequences of broadcasting the green coalition's influence are considerable. It has no immediate impact on the commuter's behaviour. He's still making his way across the station. But all the resources of his brain can now be brought to bear on the deferred problem of his domestic dispute. As recollective memory processes reconstruct the breakfast scene and language processes construct imaginary dialogue, both past and future, affective processes are colouring his mood, and will inform the response of other processes. He is regretful. His words were harsh. He entertains the possibility of apology, and feels better. The words form in his head. He resolves to phone home when he gets to the office. His resolution is fixed, and the very words he intends to use are retained by working memory processes, along with a trace in episodic memory of the whole sequence of internal events. All this is possible despite the absence of direct connections between the relevant processes, thanks to the systemic dissemination of influence and information via the global workspace.

A number of new and important themes are introduced at the end of this brief tale. We have so far taken only a preliminary step towards a proper account of internal speech, mental imagery, and the wider pageant of human inner life in terms of global workspace theory. We'll return to these topics in the next chapter. In the mean time, let's suppose the commuter's soul-searching has all taken place by Snapshot 4. No doubt numerous other steps will have occurred, each involving its own admixture of localized and system-wide processing. But we shall ignore these. In Snapshot 4, other processes are clamouring for attention (and 'attention', in its technical sense, is precisely the right word here).[2] The light blue and yellow coalitions are remnants of processing pertaining to his domestic dispute. But the magenta coalition has arisen in

[2] Everyone knows what James said everyone knows attention is, namely 'the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought' (James, 1890/1950, vol. 1, pp. 403–404). Under the present conception, it is inappropriate to speak of attention as if it were a dedicated facility whose mechanisms could be isolated from the rest of the brain's dynamics. Everything that influences the competition for dominance of the connective core contributes to attention in James's informal sense, and there is no more technical sense in which the term deserves currency.

response to an unexpected contingency. His way is about to be blocked. There is a crowd outside the entrance to the Underground.

The commuter is now no longer able to proceed on autopilot. The familiar routines encoded in the tried-and-tested red coalition are inapplicable to the current situation. The magenta coalition, whose members include the visual process that has recognized the crowd and the motor process that is slowing the commuter's stride, easily takes over the global workspace (Snapshot 5), and its influence soon permeates the whole system (Snapshot 6). This will inaugurate a further competition among potential coalitions (whose membership might well be drawn from regions associated with the brain's cognitive control network), which will not only alight on a course of action tailored for this novel situation, thus immediately impacting on behaviour, but will also lay down traces in both working memory and episodic memory, enabling the commuter later to issue a verbal report of the incident to the psychologist.

We have been examining just one among many possible variants of this episode, but in our science fiction scenario the omnipotent psychologist builds up a densely layered contrastive data set by replaying the same scene over and over again, making a range of different interventions. In addition to contrasting behaviour in the conscious and unconscious conditions, she can now contrast the internal brain dynamics that accompanies them. It's clear what she should expect according to the theory being championed here. In the unconscious condition, sensory and motor events should be accompanied by localized processing that impacts on behaviour but has little or no influence on the connective core. In the conscious condition, sensory and motor events should correlate with active process coalitions that dominate the connective core and thereby disseminate their influence throughout the brain.

5.2 **Dynamics in focus**

By now, a picture should be emerging of the sort of dynamics predicted by the theory being put forward. It results from the interplay of coalition formation, break-up, and broadcast. The overall effect resembles a stadium full of noisy football supporters.[3] During a lull, several different chants may be initiated in different parts of the home crowd. Each chant recruits singers from nearby, and each becomes louder as it spreads around the stand. Soon the chants are in competition. One is a particular favourite, and as more and more supporters join in, the others fade away until the whole stadium is singing with one voice. But this condition is only temporary. The supporters soon tire of the same song.

[3] Of course 'football' means soccer, not the game played in North America where the sort of tribal behaviour described here is perhaps less common.

There is another lull, and alternatives vie for popularity. Or perhaps events on the field intervene, causing widespread cheers or boos.

On the football stand, as in the brain, we see a mixture of spontaneous dynamics and external perturbation. We also see episodes of broadcast punctuated by bursts of competition among rival coalitions. (The medium of open air is not quite the same as that of a modular, small-world network with connector hubs, but both support a combination of local and global interactions.) A similar regime is posited in Edelman and Tononi's *dynamic core hypothesis*.[4] The 'dynamic core' is a 'distributed functional cluster' of groups of neurons.[5] The groups of neurons in question correspond to brain processes in the present account, and the distributed functional cluster corresponds to a coalition. The 'dynamic core' is dynamic because its composition constantly varies and its membership is determined by competition, in much the same way that access to our connective core (a structural rather than functional construct) is determined by competition among rival coalitions. Edelman and Tononi liken the dynamic core to a tangle of coupled springs.[6] A perturbation to any one spring quickly reverberates through the whole. (Elsewhere, they compare it to a 'riotous parliament' and contrast this to Baars's theatre analogy.)[7]

This blizzard of metaphors is all very entertaining (football crowds, theatres, riotous parliaments). But to effect a proper comparison between different accounts they need to be put on a firmer, that is to say more rigorous, footing. In mathematical terms, such systems can be characterized in terms of *attractors*. Recall that an attractor of a dynamical system is a stable point or set of points in its state space to which the system converges. There are three types of attractor—point attractors, limit cycles, and chaotic attractors. Think of a ball revolving around the inside of a cone. If the ball's energy is decreasing, then it eventually falls to the bottom of the cone and remains there forever. This is a point attractor. If the ball's energy remains constant (the cone is frictionless) and it turns around the inside of the cone at the same height forever, then it is in a limit cycle. If the ball's energy gently fluctuates (for reasons unknown), and its overall trajectory is confined to a narrow band, but divergent trajectories issue from nearby points within that band, then (a few details notwithstanding) it is in a chaotic attractor. If a system is in an attractor, of any type, then it is insensitive to modest perturbations, always returning to the attracting state.

[4] Edelman & Tononi (2000). See also Seth & Baars (2005).

[5] Edelman & Tononi (2000), pp.143–144.

[6] Edelman & Tononi (2000), p.172.

[7] Edelman & Tononi (2000), pp.245–246. In fact, as the present section hopefully makes clear, global workspace theory and the dynamic core hypothesis are not incompatible.

The collection of a system's attractors along with their surrounding basins of attraction are called its *attractor landscape*. Like a real landscape, we can think of an attractor landscape as comprising hills and valleys, where the valleys are the attractors (and the hills are 'repellors'). Imagine a giant ball rolling around in such a landscape. Depending on where the ball starts off, it will eventually end up in one valley or another. In this metaphor, the attractor landscape is a convoluted two-dimensional surface, representing a two-dimensional state space. But in general, the state space of a complex system can have many more than two dimensions. This is certainly the case with the sort of system we are envisaging here. If we suppose that the state of each brain process (including its level of activity) is a single dimension, then the system as a whole has as many dimensions as it has processes (active or dormant). Similarly, in the football stadium metaphor, the state of each supporter (his current chant, not his level of inebriation) might be considered a single dimension. But the concepts of an attractor and an attractor landscape are the same for state spaces with very many dimensions as for those with just two.

One way to see competition in a dynamical system is in terms of the opposing pull of various attractors. In the present case, if the repertoire of coalitions were fixed, then we might elect to make the level of activity of each coalition a dimension in the pertinent state space. In this case, the giant ball rolling around the attractor landscape would feel the pull towards several valleys at once—several coalitions—the level of activity in rival coalitions waxing and waning, until eventually the system settles into one, the winning coalition. Matters are made considerably more complicated by the requirement for open-ended coalition formation. This entails that every viable combination of processes is a potential attractor, and the forces tugging at the giant ball are therefore all the more subtle. But the basic picture is the same—for one coalition to become dominant, excluding its rivals, is for the system to succumb to the pull of the corresponding attractor.

However, this basic picture is incomplete. First, it must be elaborated to allow for the co-existence of coalitions that are not in competition with each other (as we saw in Snapshot 2 of Fig. 5.2). This is not much of a problem, as the system's state space can simply be partitioned into subsets that are effectively independent. Second, and more significantly, we have to face up to the fact that the system never actually resides in the same state for very long. Like Odysseus, its fate is to wander restlessly among the attractors, none of which has sufficient pull to become its permanent home.[8] The triumph of one

[8] Of course, Odysseus got home in the end, whereas the brain is restless until it goes into coma or dies. No metaphor is perfect.

coalition may usher in a period of relative calm, as the system remains in, or close to, the same state for a while. But this is only a temporary condition. Before long, thanks either to evoked dynamics (external perturbation) or to spontaneous (internal) activity, the system will shift away from the 'attractor' that previously held it captive, so that the resulting dynamics, like a bird's life, 'seems to be made of an alternation of flights and perchings'.[9]

These two cases—external perturbation and spontaneous activity—need to be treated differently. When then the system is perturbed by some external influence, such as the sight of a crowd outside the Underground station, its attractor landscape alters. It is as if some god-like being had grasped the landscape by one corner and tugged at it so that its folds are re-arranged. Some valleys become hills and some hills become valleys. The giant ball begins to roll once more, pulled towards this slope and pushed away from that, until it finds its way in due course to a new low point. However, if the system's move away from a region of state space where it previously lingered is solely due to internal activity, then it's incorrect to speak of the attractor landscape changing. Rather, we must acknowledge that the 'attractors' in question are not true attractors. They are so-called *quasi-attractors*, that is to say regions of state space in which the system dwells for statistically significant periods without being trapped in them permanently.[10] The system is then said to exhibit *metastability*, and while it resides in a quasi-attractor it is said to be in a metastable state.[11]

Thanks to its ecological situation, the dynamics of the living brain is neither wholly spontaneous nor wholly evoked, but a blend of the two. This entails that even evoked dynamics is best thought of in terms of quasi-attractors rather than true attractors. What takes place when a stimulus is received is then a reshaping of the associated quasi-attractor landscape, and the brain's state is always metastable. As argued by Kelso and others, a metastable system, because it never completely relaxes into a stable condition and is more easily nudged away from an attracting region of its state space, is well placed to respond rapidly and fluidly to the ongoing situation. In short, metastability is a necessity for the machinations of off-line cognition, and a benefit when responding to incoming stimuli.[12]

[9] James (1890/1950, vol. 1, p. 243). James uses the metaphor of the bird to evoke phenomenology rather than neurodynamics. But the neurodynamics we are here trying to evoke surely underlies the phenomenology that James was alluding to. For a neurodynamical account compatible with the present story, see Gros (2009).

[10] Amit (1989), Chapter 5; Haken (2006).

[11] Kelso (1995); Bressler & Kelso (2001); Werner (2007).

[12] Kelso (1995); Bressler & Kelso (2001). Bressler & Kelso write: 'metastable dynamics … endow cognitive functions with the capacity for rapid and fluid change, without ever relaxing into stable states' (p.34).

However, metastability is not enough in itself. A system that flips back and forth between two attractors (we'll drop the 'quasi-' prefix from now on and take it as read), tarrying awhile in each but never settling down, is metastable, but uninterestingly so. A brain capable of a large repertoire of responses, each tailored to the subtleties of the ongoing situation for the animal, requires a non-trivial repertoire of attractors. Indeed, as the attractors in question are brain-wide coalitions of processes, whose composition is open-ended not just tried and tested, this repertoire is more than non-trivial. It is exponentially large.[13] So the question arises of how the system is to navigate among them. We shall address this question in two stages. First, we'll point out some likely characteristics of the trajectory a system will follow as it moves among the attractors. Second, we'll identify the likely mechanism by means of which this trajectory is realized.

5.3 **Wandering among the attractors**

For reasons to be spelled out shortly, the system's movement from attractor to attractor is likely to be a form of *chaotic itinerancy*.[14] To visualize the concept of chaotic itinerancy, rather than a ball in a cone, imagine a fly (the state of the system) temporarily trapped in a tube (an attractor). The fly buzzes around inside the tube until by chance its erratic trajectory takes it to the rim and it escapes. Having escaped, its trajectory, though still erratic, is no longer confined by the tube, so it gets to explore the whole room. But now suppose the room contains numerous other tubes (other attractors). Before long it will find itself trapped once again, in a different tube. Over time, the fly will visit many different tubes, in no particular order, sometimes revisiting the same tube more than once, spending a little while in each one. The fly's trajectory is chaotic (hard to predict) both when it's in a tube and when it's free. But it is *more chaotic* (even harder to predict) when it is free than when it is in a tube.

This notion of degrees of chaos can be made mathematically precise in terms of *Lyapunov exponents*, which quantify the rate of divergence (or convergence) of trajectories issuing from nearby points in a system's state space. The notion can be given an intuitive gloss using the fly example. Within a tube, trajectories emanating from nearby points diverge over time. But they remain confined to the tube (until the moment of the fly's escape), so their rate of divergence is not great. Trajectories emanating from nearby points in free space, on the other hand, are not so constrained and diverge much more dramatically,

[13] In other words, it scales exponentially with the number of brain processes. Assuming the brain processes are themselves numerous then, even allowing for the various incompatibilities that rule out most combinations, the number of possible coalitions will be astronomical.

[14] Kaneko & Tsuda (2003).

a condition that persists until the fly becomes trapped in a tube once more. It is characteristic of a chaotically itinerant system, such as our fly, that it has a fluctuating Lyapunov exponent.[15] Within an attractor, the Lyapunov exponent is close to zero, but when the system is in transit between attractors its Lyapunov exponent is much higher.[16]

Not every system with a large repertoire of metastable states is chaotically itinerant. Such a system might visit each of its (numerous) attractors in a periodic and orderly fashion which is unaffected by small perturbations. So why might the system we are interested in exhibit chaotic itinerancy? One reason is that when an animal is at rest, pausing, or performing an undemanding task, the dynamics of its brain can support an elevation in spontaneous activity. Chaotic itinerancy is one way—perhaps the only way in a biological rather than silicon substrate—to take advantage of this and to effect an off-line exploration of the space of possible affordances. To see this, recall that a combinatorial tree of possibilities branches out from every situation the animal might find itself in (Fig. 2.1). In order to search this tree effectively, it must be possible both to foresee the outcome of several actions or behaviours chained together and to investigate multiple possibilities branching away from a single state. Although this allows for a degree of parallelism, the chaining together of hypothetical actions necessitates a certain amount of serial processing.

The dynamical regime presently under discussion exhibits just the right combination of serial and parallel processing for the job. The procession of attractor states whose influence is disseminated by the global workspace is serial, yet each state-to-state transition results from the sifting and blending of massively many parallel computations.[17] However, to search through a combinatorially structured space of behavioural sequences, the system must be capable of revisiting a state it has already seen and generating a *different* successor. In a conventional computer this can be achieved by maintaining a data structure called a stack, which keeps a record of unexplored alternatives. Brains do not have stacks. However, if the dynamics of generating the successor to the

[15] Sauer (2003).

[16] For this to make sense, we must consider finite-time Lyapunov exponents rather than the Lyapunov exponent in the limit (Abarbanel, *et al.*, 1991; Sauer, 2003; Tsuda & Umemura, 2003).

[17] For an insightful discussion of the issue of serial versus parallel processing in relation to consciousness, see Dennett (1991), pp. 209–226. Sackur & Dehaene (2009) present neuroscientific evidence for such a combination of serial and parallel processing in a simple arithmetic task, and supply an interpretation of their results in terms of global workspace theory. Their findings suggest a degree of pipelining, wherein serial operations partly overlap. This is consistent with the present discussion.

current attractor is sensitive to small perturbations—if it is chaotic in other words—then a stack is not needed. In a dynamically rich milieu, the same attractor will have different successors on different occasions, and search is made possible.[18]

So the advantages of being able to carry out off-line forays into the space of possible affordances, thereby exposing affordances that were previously hidden from the animal, favour chaotic itinerancy. Very well. But there are important questions still to be asked about the statistical properties of the system's wandering trajectory. For example, how long should we expect it to dwell in each attractor compared to the time it spends wandering between them? Some chaotically itinerant trajectories feature long transients between fleeting attractor visits. But this would be disadvantageous to an animal who needs to take action quickly, and for whom behaviour selection depends on the establishment of a winning coalition. So rapid transitions between attractors are in order. So much is obvious.

A trickier question is this. What does the probability matrix for transitions from attractor to attractor look like? We have already established that it should comprise more than just ones and zeros, because attractors have non-unique successors. The simplest distribution would be uniform—whatever attractor the system is in, every other attractor has an equal chance of becoming its successor. But this is not a plausible answer. The tree of affordances is highly structured. Only certain actions and behaviours are possible in any given situation, and this guarantees a zero for most cells in the probability matrix. But what about the rest? An enormous number of actions and behaviours are possible in most situations. Imagine someone turning up for work at the office. She arrives at her desk, puts down her coat, and she is faced with a choice. She could do something commonplace, like fetching a cup of coffee or switching on her computer. But she could do something unusual, like deliberately tipping a cup of coffee into the computer or dancing a jig on her desk.

On the one hand, it would be wasteful to spend much time exploring exotic regions of the tree of affordances, regions filled with unlikely events and pointless actions. On the other hand, only stereotyped behaviour will result if off-line exploration is confined to the commonplace. For an animal capable of invention, of opening up new vistas of affordance, a balance must be struck. Most of the time its brain must go down familiar paths in the attractor landscape. Considering that many such paths emanate from most attractors, this still leaves much to explore. But sometimes its brain should follow a faint path, and sometimes it should open up a path where there was none before.

[18] Nara & Davis (1992); Tani (1996).

We should not be surprised, then, if the probability matrix for transitions among attractors contained a large spread of numbers.[19] (Of course, the very act of exploration is liable to alter the matrix, further complicating the picture.)

The emphasis of this discussion has been off-line exploration. But the set of issues that arises is similar when the space of affordances is subject to on-line exploration, that is to say when the animal playfully interacts with the environment and thereby reveals affordances that were not apparent to it beforehand. The attractors then in question are not the product of internal coupling only, but arise from the brain's coupling with the body and its environment and the resulting interplay of internal and external dynamics. Recall the kingfisher bobbing on the reed, but imagine an infant with a box of new toys. As in the off-line case, to effect a useful exploration, the system must be capable of visiting the same attractor more than once, and moving on to a different successor on each occasion, so chaotic itinerancy is to be expected. Following Ikegami, we might term this *embodied chaotic itinerancy*, noting that the interplay of on-line activity and off-line exploration is itself likely to follow a complex, itinerant pattern.[20]

5.4 **Dynamical complexity**

Now, if the attractor landscape in question were simple, we might let the matter rest there. However, recall that for the cognitive high-fliers of the animal kingdom (humans and perhaps others), the space of possible affordances is not only combinatorially structured but also open-ended. The set of potential actions or behaviours that are executable in a situation and meaningful for the animal is neither fixed in advance nor bounded. In terms of the dynamics we are investigating this goes hand-in-hand with the open-endedness of the repertoire of process coalitions that are eligible to respond to that situation. So the following question arises. What sort of mechanism is capable, not merely of selecting from ranks of known attractors, but of surveying an endless panorama of combinatorial possibility and actually engendering just those very few coalitions that might be worthy successors to the incumbent attractor, and then choosing the best among them as its heir?[21]

Drawing on their empirical findings with rabbit olfaction, Skarda and Freeman (in a classic 1987 paper) conjectured that 'chaotic activity provides a way of exercising neurons that is guaranteed not to lead to cyclic entrainment

[19] For an example of chaotic itinerancy with this property, see Dias, *et al.* (2008).

[20] Ikegami (2007).

[21] As we shall see in the next chapter, this question relates to the so-called frame problem.

[and] allows rapid and unbiased access to every limit cycle attractor'.[22] Their conjecture, originally formulated with specific brain regions in mind, might be generalized to the dynamics of the whole brain, and applied both to spontaneous and evoked activity.[23] According to a generalized conjecture, the chaotic nature of the spontaneous activity forming the backdrop to an incoming stimulus would grant rapid access to the system's full repertoire of coalitions. This would facilitate not merely swift and appropriate behaviour selection, but also the immediate assembly of novel behaviours tailored to meet the situation at hand, as well as ensuring that off-line exploration of the space of affordances was open-ended.

As we have already conceded that nearby trajectories in the transitional phase between attractors can diverge quickly enough to yield different successors on different occasions, then if the generalized Skarda–Freeman conjecture is true, perhaps no more needs to be said. Chaos will confer easy access to every attractor, to the full repertoire of process coalitions. Unfortunately though, irrespective of whether Skarda and Freeman's original conjecture is true, the generalized version surely is not (which is no discredit to them, of course). Chaotic dynamics may help, but it is not enough. To restate the difficulty, what we seek is a mechanism that can bring into existence coalitions that have never been generated by the system before, as well as overseeing a competition among them.

The conjecture to be defended here (which in fact sits comfortably with Freeman's more recent work)[24] is that it is not chaos that confers access to the full repertoire of coalitions, but *dynamical complexity*, wherein a balance is struck between the antagonistic forces of *segregation* and *integration*.[25] Now, the term 'integration' has already been given a special meaning, so we must take care to distinguish it from the present sense of dynamical integration, which is a different, though intimately related concept. In the former sense, integration is achieved when the whole subject (including the full resources of that subject's brain) is brought to bear on the ongoing situation. Integration in

[22] Skarda & Freeman (1987). The point has been reaffirmed by many other authors since the Skarda & Freeman paper (eg: Rabinovich & Abarbanel (1998); Dias, *et al.* (2008)). See also the review by Korn & Faure (2003). The hypothesis has also been vindicated in modelling work, such as Torres, *et al.* (2008).

[23] See the remarks of Tsuda (2001; 2009), for example.

[24] Freeman embraces the concept of chaotic itinerancy, which is compatible with his EEG findings (Freeman, 2003; 2006). See also the overview in Kozma & Freeman (2009).

[25] Tononi, *et al.* (1998); Seth, *et al.* (2008); Shanahan (2008b). As discussed in these papers, several mathematically precise measures have been proposed for quantifying dynamical complexity.

this cognitive/behavioural sense is subserved by integration in the latter, dynamical sense, which arises when the activity of each of the system's parts is under the influence of the system as a whole. But the former concept only makes sense in the context of a behaving animal, whereas the latter concept can be applied to any dynamical system of interacting components.

Why is a balance of segregated and integrated activity important in the present context? Well, the need for segregation should be uncontroversial. In a system that comprises multiple components, such as an architecture comprising many parallel specialists, each of the components must be free to go about its own, independent business. If there were no segregation of activity, and every part of the system did the same thing, then there would be no parallelism, no specialists (and no need for separate components). Such a system would be maximally integrated, but it would be paralysed, its state frozen.

In a system that is overly segregated, on the other hand, each part goes about its business oblivious to the activity of all the other parts. In such a system it would be impossible for a previously unseen coalition to form, wherein the expertise of one specialist is blended with the expertise of another with whom it had never before co-operated, because the specialists in question would be unable to exercise any influence on each other. A system is dynamically integrated when the activity of its parts is influenced by the activity of the whole, and it is dynamically complex when this influence is not too great, when the activity of its parts is not *dictated* by the activity of the whole. Only when the route the system takes through its attractor landscape is governed by such a balance of segregated and integrated activity, when there is a two-way flow of influence and information between the parts of the system and the whole, is it possible for arbitrary coalitions to emerge.

It has been argued, notably by Bak and Chialvo, that the brain should exist in a critical state, poised at the edge of a phase transition from order to disorder.[26] As Chialvo points out, certain systems that are known to fit this description, such as the 2D Ising model of ferromagnetism, generate a large repertoire of metastable states when (and only when) they are at the critical point. Moreover, in the Ising model, the ordered condition is highly segregated whereas the disordered condition is highly integrated, in the very sense at issue. So at the critical point, when integration and segregation are in balance, we have dynamical complexity. Given the striking similarities then, it makes sense to ask whether the dynamics we are trying to pin down here is also critical, that is to say on the edge of a phase transition.

[26] Bak (1997); Chialvo (2004; 2008); Fraiman, *et al.* (2009). Turing (1950) anticipated the thought in his seminal paper in the journal *Mind*.

The answer appears to be yes, but not trivially so. For the idea of criticality to be applicable, there need to be two qualitatively distinct conditions (for example, solid and liquid) such that the transition from one to the other, brought about by continuous change in some variable (such as temperature), is abrupt. Nothing we have said about our system so far is suggestive of this abruptness. For sure, we have the requisite opposed regimes—highly segregated and highly integrated. But could the transition from one to the other not be gradual rather than sudden? Well, evidence from computer models of modular neural networks, where the continuously varied parameter governs the density of intra-modular connections, suggests otherwise.[27] The onset of integration in this model is very sudden, and the decline into too much integration, though slower, is also very rapid. Similarly, in a related model of synchronous oscillators organized into a modular network, the ability to generate a large repertoire of synchronized coalitions occurs only in a narrow parameter range bracketed by overly segregated and overly integrated regimes.[28] Both computer models mimic the connectivity and the dynamics we have been investigating, so we can tentatively conclude that it too is poised at a phase transition, and add criticality to the list of significant properties we are compiling.[29]

Work with computer models suggests that the promotion of dynamical complexity is a generic property of modular small-world networks.[30] This makes intuitive sense. It seems natural for segregation to result from a high clustering coefficient, and for integration to result from a short mean path length. This appears to be the case whether or not the modular small-world network in question has connector hubs and a pronounced connective core. But when the network does have a pronounced connective core, as we are proposing here, there are further implications. Specifically, the two-way flow of influence and information that mediates the transition from one attractor to the next has to be channelled through it. This entails that the connective core is not only the locus of broadcast, but also the arena for competition among competing attractors and the medium of coupling among coalition members.

In a system capable of generating an open-ended repertoire of attractors, this is exactly what we should expect. How else could two processes that have

[27] Shanahan (2008b).

[28] Shanahan (2010). A similar phenomenon is demonstrated by Kitzbichler, *et al.* (2009), following the work of Kuramoto (1984, Chapter 5), though not in the context of modular small-world connectivity.

[29] However, it is less clear what can be said about various other phenomena that are frequently discussed in the context of criticality, such as self-organization and the prevalence of power laws.

[30] Sporns, *et al.* (2000); Shanahan (2008b).

never co-operated before, between which there are no pre-existing dedicated pathways, establish a coupling except via an open-access, public communications infrastructure, which is exactly what the connective core is? The connective core has to be the medium of coupling. Moreover, as it has a limited capacity, there is bound to be competition for its narrow bandwidth. Not every pair of processes that attempts to establish a coupling via the connective core will succeed. Not every process that attempts to join a coalition can be accepted. A winning coalition crystallizes when a dominant set of couplings emerges and their rivals melt away. In short, the fight to establish (novel) couplings is the competition for access to the connective core, to be victorious in this fight is to take up the connective core's limited capacity, and domination of the connective core is broadcast (Fig. 5.3).[31]

One final piece is required to complete the jigsaw puzzle. In the dynamical regime we are targeting, the balance of segregated and integrated activity is not static when viewed at fine timescale. If we imagine a thin line of perfect balance, dividing order from disorder (high integration from high segregation), then an itinerant trajectory from one metastable state to another, from one dominant global coalition to another, corresponds to a periodic wobble from

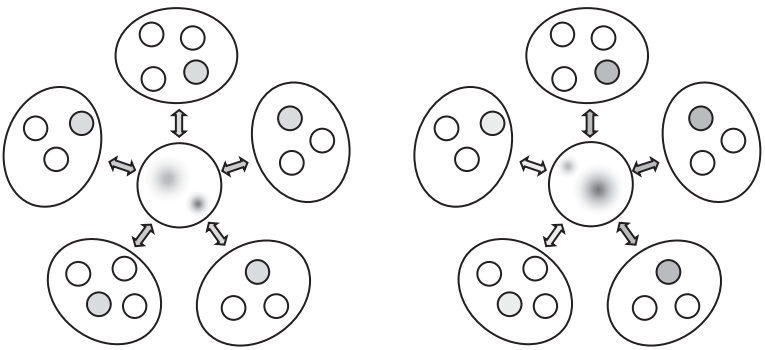

Fig. 5.3 Competing for the global workspace. The global workspace is not only the locus of broadcast, but also the arena for competition between rival coalitions (central colours) and the medium of coupling for the members of those coalitions (arrow colours). Here we see a contest between the red and green coalitions. Initially the green coalition is dominant. But the red coalition manages to recruit a former member of the green coalition, and begins to supplant it in the workspace. (See Plate 6).

[31] This is not to imply that the connective core is the *only* medium by which two processes can become coupled. It is assumed that a web of direct connections also exists, that allows coalitions to form without implicating the connective core. But it follows from the present proposal that such direct connections lack the configurability necessary to support *novel* coalitions.

one side of the line to the other. When one coalition is dominant, and its pattern of activation is being globally broadcast, the system is temporarily resident in a small margin lying on the integration side of the line. But the system drifts back over the line into the more segregated region during the upheaval that precedes a new coalition becoming dominant. In other words, the system repeatedly makes phase transitions from order to disorder, from integration to segregation, and back again, albeit remaining forever close to the dividing line between the two.[32]

To summarize, complexity, in the sense of a balance of integrated and segregated activity, is a signature of the dynamical milieu within which the competitive formation of arbitrary coalitions drawn from an open-ended repertoire is possible. The important point, of course, is that the global workspace architecture promotes dynamical complexity. Or rather, the network topology we have advocated as the substrate for a global workspace architecture promotes dynamical complexity. In fact, we are now in a position to formulate a fairly precise hypothesis: *A small-world, modular network with a pronounced connective core can support a dynamical milieu that 1) promotes complexity, a (fluctuating) balance of integrated and segregated activity, facilitating chaotic itinerancy among an open-ended repertoire of metastable states (each comprising a coalition of coupled processes), and 2) is characterized by episodes of broadcast punctuated with bursts of competition*. In Chapter 4, we saw that the structural connectivity of the human brain conforms to the required topological prescription. It now remains to provide some evidence that the human brain also conforms to the hypothesized dynamical description.

5.5 **Fireflies of the mind**

Although the brain is frequently alluded to in the preceding presentation, there is nothing in the basic theory that is specific to biology. A theory couched in terms of network topology and dynamical systems is surely closer to biology than the more abstract style of architectural description that opened Chapter 4. But in principle, its component processes could be realized not only by familiar forms of neural wetware, but also by some barely imaginable product of extraterrestrial evolution, or by artificially manufactured circuitry, analogue or digital. This is as it should be if our sights are set on a high-level theory expressed in terms of deep principles. In computer science terms, what we have provided is analogous to the specification of a virtual machine amenable

[32] In a similar vein, Glazebrook and Wallace (2009) characterize broadcast from a global workspace in terms of a phase transition wherein a 'giant component' emerges, that is to say a dominant network of interactions among processes.

to hardware or wetware implementation in any number of ways.[33] Nevertheless, the biological brain is our touchstone, the only source we have of empirical data that could support or discredit our hypotheses. So, in this section and the next, we survey findings supportive of the foregoing dynamical description, and suggestive of the possibility that synchronous neural oscillation is the low-level mechanism that instantiates it in the human brain.[34]

Synchronization is a commonplace phenomenon in Nature.[35] One of the most enchanting examples is the firefly.[36] At night, in many Southeast Asian countries, entire swarms of male fireflies distributed about the branches of a tree will sometimes flash rhythmically and simultaneously, keeping perfect time with each other. If a barrier is interposed between two fireflies, so that they cannot see each other's (or anyone else's) displays, their flashes will desynchronize, indicating that the processes involved are mutually coupled. Like a tree laden with fireflies, some of whom can see each other and some of whom cannot, the brain throbs and pulses with electrical rhythms at various frequencies, and many of these rhythms are synchronized.[37]

Although various brain rhythms have been known about since the 1940s thanks to EEG technology, it was the 1980s work of Singer and colleagues on the visual cortex of the cat that brought about the discovery of synchronized gamma-band oscillations.[38] (The gamma band ranges from 30 to 70 Hz.) In a sense, the presence of synchronous oscillation in the brain is hardly surprising. When large numbers of dynamically interacting components are connected together, the emergence of all sorts of rhythmic patterns is almost unavoidable. Think of a tree waving in the wind. The tips of the branches wave back and forth in a periodic fashion. Some of them synchronize for a while, especially if they share common branches, then desynchronize. Individual leaves flutter rhythmically at higher frequencies. The whole bough rocks to and fro at a low frequency. This is all very pleasant to watch on a breezy day, but it tells us nothing of significance about the biology of the tree.

The question, then, is whether synchronous oscillations in the brain are any different. Are they merely an intriguing emergent phenomenon, or do they

[33] Sloman & Chrisley (2003).

[34] Long-range coherent oscillation was first posited as a means for realizing a global neuronal workspace by Deheane, Changeux, and colleagues (Dehaene, *et al.* (1998); Dehaene & Naccache (2001)).

[35] Pikovsky, *et al.* (2001).

[36] Buck (1938).

[37] Buzsáki (2006).

[38] Gray, *et al.* (1989).

have a functional role? A popular line of enquiry in the 1990s was inspired by the hypothesis that neural synchrony addresses the so-called binding problem.[39] The binding problem, in the sense intended here, is highlighted when an animal simultaneously perceives multiple relations among stimuli. For example, suppose a subject is looking at a screen containing both a red circle and a blue triangle. It is reasonable to assume that the redness of the circle will excite a set of neurons specialized for recognizing red, and the blueness of the triangle will excite the corresponding set of blue neurons. At the same time, the roundness of the circle will excite a set of neurons specialized for recognizing round shapes and the presence of the triangle will excite the corresponding triangle neurons. If all these neurons are active at once, the puzzle is to understand how redness is associated with (bound to) the circle and *not* to the triangle, whereas blueness is associated with the triangle and *not* to the circle.

According to the 'binding by synchrony' hypothesis, two or more features are associated with each other when the relevant neural populations fire in synchrony. Conversely, features do not become associated when they are not supposed to be, even when they are all perceptually present, because the relevant neural populations fire out of phase with each other. In our example, the red neurons would fire in synchrony with the circle neurons and the blue neurons would fire in synchrony with the triangle neurons, but the former pair of neural populations would fire out of synchrony with the latter pair of neural populations. Because the oscillations in question are fast, it's possible to time-slice rapidly between one relational percept and the other, enabling both to be processed 'simultaneously', in much the same way that multi-tasking is achieved on a computer using a single processor without the user (who operates at a slower timescale) being aware of it.

The binding by synchrony hypothesis enjoys a degree of empirical support, and postulates a clear functional role for synchronization. But a different, though compatible, functional role is posited by the 'communication through coherence' hypothesis, advanced by Fries, and this is what really interests us here (Fig. 5.4, left).[40] The essence of this hypothesis is that the selective passage of information between two groups of neurons is facilitated if 1) there is synchronous oscillation within each group, and 2) this synchronous activity is coherent (phase-locked) between the two groups.[41] The reasons for this

[39] Singer & Gray (1995); Von der Malsburg (1995).

[40] Fries (2005; 2009).

[41] Coherence is a more liberal notion than synchrony. Two periodic systems are coherent if they maintain a stable phase relationship. They are synchronous if they are in phase. In fact, we shall mostly be concerned here with synchronous oscillations. But we shall follow the naming convention used by Fries.

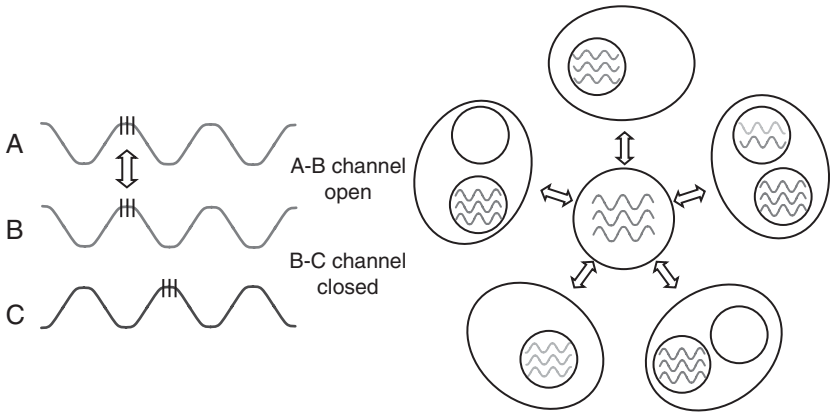

Fig. 5.4 Left: The basis of the 'communication through coherence' hypothesis (adapted from Fries (2009)). Populations A and B exhibit coherent oscillations with a phase relationship that allows information to pass between them in both directions. No such channel is open between populations B and C, however, because A's spikes are timed to arrive when C is least excitable. Right: Coalition formation through coherent oscillation in the context of the global workspace architecture. There is traffic of information among the three red populations, which have formed a dominant coalition. The green and blue populations are unable to communicate because they are out of phase. The cyan and magenta population is not synchronous, and is unable to join any coalition. (See Plate 7).

are twofold. The first reason is this. Suppose A is a sending group of neurons and B is a receiving group. A pattern of spikes sent from A is more likely to make any given neuron in B fire if those spikes are closely timed, quickly ratcheting up the membrane potential of the receiving neuron, than if they are spread out giving the membrane potential time to slip back between incoming spikes. If A is characterized by synchronous activity, then the spikes it sends out will be concentrated in peaks of activity, and will have a correspondingly greater impact on their targets.

The second reason concerns the receiving population, and assumes that the synchrony in B is due to a rhythmic alternation between periods of excitability and periods of inhibition.[42] In effect, this regular inhibitory activity modulates the gain of B's excitatory population, ensuring that it is most sensitive to incoming spikes during the inhibitory troughs. By contrast, during peaks of inhibition, its gain is low, and B is effectively desensitized to input. The combination of these two periodic effects—the greater influence of a synchronized

[42] Relevant neuronal mechanisms are described by Hasentraub, *et al.* (2005) and Dupret, *et al.* (2008).

sending population and the increased sensitivity of a synchronized receiving population—enables the rhythmic opening and closing of a channel allowing the transmission of information from A to B.

However, for this to occur, the phase relationship between the two populations has to be right. Specifically, spikes have to arrive at the receiving population during an inhibitory trough. If the conduction delay is small enough compared to the cycle time of the oscillations in question, it suffices for the two populations to be nearly in phase. Indeed, if this is the case, the channel that is opened up in peaks of (excitatory) activity allows for two-way communication—the transmission of information from A to B and from B to A. Note that the mechanism described is capable of closing a channel of communication between two internally synchronized populations as well as opening one. To open a channel the synchronous activity in one population needs to be in phase, or nearly in phase, with the synchronous activity in the other. To close a channel, it suffices to disrupt this phase relationship.

Now, the (putative) phenomenon of communication through coherence is of particular interest in the context of a global workspace architecture because it suggests a precise neural mechanism both for the competitive establishment of coalitions of brain processes, and for the global dissemination of the winning coalition's influence (Fig. 5.4, right). First, it posits a reconfigurable network of switchable channels, which is sufficient to sustain arbitrary coalitions of neural groups coupled through the mutual exchange of influence and information. Second, as Fries explains, it proposes a means to realize a winner-takes-all competition between two neural populations for influence on a third.[43] Suppose neural groups A and B supply converging input to a third group C. A and B can compete to influence C if, despite being internally synchronized, they are out of phase with each other. Group C will tend to entrain to either A or B, but will be unable to entrain to both. Any preference it shows for one group over another—possibly thanks to a closer initial phase relationship to one than another, or because one is less synchronous than the other—will quickly be amplified, ensuring that the influence of one group dominates while that of the other is excluded.

Thanks to this winner-takes-all mechanism, nascent coalitions could competitively bid to recruit groups of neurons into their membership, at the same time as competing for influence on any neutral target groups, including those sited in the connective core of the brain's structural network. As they are richly interconnected, it's reasonable to suppose that these groups would tend to entrain to each other. So any coalition that wins control of a dominant set of

[43] Fries (2009).

these groups would be well placed to take over the connective core altogether. The consequence of such a takeover is that the winning coalition would get to disseminate its influence throughout the brain (according to the proposal of Section 4.8). Now the very same trick of exploiting coherence to gate the flow of influence and information between processes becomes the means of dissemination. Driven by the connective core into coherent oscillation with the winning coalition, outlying processes would become receptive to whatever pattern of activation that coalition wishes to deliver to them.

However, no coalition could remain in the ascendant for long. Although re-entrant pathways in the connective core will promote reverberation, enabling self-sustaining patterns of activation (attractors) to linger there for some time after the excitation that initiated them has faded, in due course some rival pattern, initiated either by an incoming stimulus or by one of the very outlying processes that was stimulated by the connective core's present pattern, will inevitably nudge the core out of its old attractor and into a new one. Over a longer time period, if this story is right, we would expect to see episodes of broadcast punctuated by bursts of competition, resulting in a chaotically itinerant series of visits to different attractors. This is a speculative description. But it does suggest that the hypothesized mechanism of cortical communication through coherence would be capable of supporting the sort of global workspace dynamics we have envisioned.

5.6 **Evident coherence**

So is there any empirical evidence of such a mechanism at work, producing the hypothesized dynamics? Preliminary evidence in favour of the communication through coherence hypothesis has been obtained by Womelsdorf and colleagues.[44] Using data gathered by intracranial recording in the visual cortices of awake cats and monkeys, they showed that gamma-band power (the prevalence of gamma oscillations) was better correlated between groups of neurons that had a close phase relationship than between those that did not. Moreover, they showed that increased correlation in gamma-band power between well-correlated groups was preceded by increased phase coherence, suggesting causal precedence between the former and the latter. Both results are supportive of the hypothesis.

But what evidence is there of the distinctive dynamics proposed here? If the present proposal is on the mark, we should expect to find evidence of long-range synchronization, not just synchronization within a single region such as visual cortex. In fact, evidence of such a phenomenon has been accumulating

[44] Womelsdorf, *et al.* (2007).

for a number of years. Varela and colleagues supplied the earliest result.[45] They gathered EEG data from subjects who were presented with very high contrast black-and-white images of faces. The type of image in question is easy to recognize as a face when presented in an upright orientation, but appears meaningless when presented upside-down. Thirty electrodes were placed across each subject's scalp, and the EEG data for the face-perceived (upright) condition was compared with that for the meaningless-image (inverted) condition. In particular, the phase synchrony between pairs of electrodes was measured, using wavelet filtering. The electrical activity at the sites of two electrodes was considered synchronous if the phase lag between the corresponding signals remained constant throughout all trials.

In the face-perceived condition, averaged over trials, electrodes, and subjects, phase synchrony rose significantly to a peak approximately 250 ms after presentation of the stimulus, then fell sharply to a trough at approximately 500 ms, then rose again to a new plateau at around 800 ms.[46] By contrast, in the meaningless-image condition, phase synchrony remained close to zero until around 600 ms after presentation of the stimulus when it began to rise, reaching a plateau at around 800 ms. The authors speculated that the alternation between coherence and decoherence denoted 'a transition between two distinct cognitive acts … punctuated by a transient stage of undoing the preceding synchrony and allowing for the emergence of a new ensemble'.[47]

Similar results were reported by Freeman and colleagues, based on two sets of EEG experiments, one with rabbits and cats and another with humans.[48] In the animal experiments, high-density arrays of electrodes were placed intracranially on several cortical regions—visual, auditory, somatomotor, and entorhinal. Subjects were trained to discriminate two auditory and two visual conditioned stimuli, one with reinforcement and one without in each case. The EEG data obtained showed that the overall level of intercortical synchrony was significantly higher than that of a 'shuffled' control signal. However, this intercortical synchrony was not spread evenly over time. Rather, it was divided into epochs of high synchrony punctuated by peaks of decoherence.[49]

---

45 Rodriguez, *et al.* (1999); Varela, *et al.* (2001).

46 See Fig. 2 on p. 431 of Rodriguez, *et al.* (1999).

47 Rodriguez, *et al.* (1999), p. 432.

48 Freeman & Rogers (2003).

49 The intervals between synchronization peaks varied across trials, but the peaks tended to be more frequent in cats than in rabbits, who also displayed less intercortical synchronization overall. See fig. 2 on p. 2872 of (Freeman & Rogers, 2003).

In the human experiments, a curvilinear array of 64 electrodes was placed on each subject's forehead, and EEG data was collected while the subject was in a state of relaxation, first with eyes closed then with eyes open. The results were similar to those in the animal experiment. The phase difference of each of the 64 signals was plotted, and this revealed epochs of coherence lasting 100–200 ms punctuated by brief peaks of decoherence occurring almost simultaneously across multiple, distant electrode sites.[50] The rate of recurrence of phase stabilization was, perhaps surprisingly, faster in humans than in cats or rabbits. When subjects' eyes were closed and they were asked to relax, the recurrence rates tended to be in the alpha range (7–12 Hz), and when they opened their eyes, this shifted to the theta range (3–7 Hz).[51] Freeman interprets these findings as evidence that neocortex processes information in a series of movie-like frames that denote 'recurring episodes of exchange and sharing of perceptual information among multiple sensory cortices'.[52]

Further evidence of long-range gamma-band synchrony in humans was obtained by Doesburg and colleagues. In their first experiment, EEG data were gathered from subjects participating in a visual attention task.[53] The task involved the presentation of stimuli either to the left or the right half of the visual field while the subject's gaze was fixed on the centre. Their results showed an increase in gamma synchrony between visual areas contralateral to the presented stimulus and several widely distributed cortical regions. Moreover, echoing the findings of Varela's and Freeman's groups, they found that episodes of synchrony were punctuated by periods of desynchronization, and that this modulation of gamma synchrony occurred at a frequency in the theta range.

[50] See fig.12 on p. 2879 of Freeman & Rogers (2003).

[51] In Freeman (2004a; 2004b), EEG data from an earlier experiment are revisited in the light of these discoveries. The original data were collected from small square arrays of 64 electrodes attached to the primary sensory cortices of rabbits trained to respond to a variety of visual, auditory, and tactile stimuli (Barrie, *et al.* 1996). When re-analysed using newly defined indices of temporal synchrony and spatial stability, the data revealed a similar pattern of episodic activity, featuring epochs of high coherence separated by peaks of instability.

[52] Freeman (2004a), p. 2077. See also the figure in electronic-only attachment 6 to that paper. The idea is expanded on in Freeman (2006). In a similar vein, Llinás and his colleagues have conjectured that 'consciousness is a non-continuous event determined by synchronous activity in the thalamocortical system' (Llinás, *et al.*, 1998, p. 1845).

[53] Doesburg, *et al.* (2008).

A second study elicited similar findings with a binocular rivalry paradigm.[54] Recall that the phenomenon of binocular rivalry occurs when different images are presented to each eye. Subjects report seeing one or other of the images (not both at once), but which image they see changes every second or so. Doesburg and colleagues collected EEG data from 59 electrodes across the cranium from subjects who reported which of the two images they were experiencing with button presses. Their analysis revealed that button presses, indicating the onset of a change of percept, were preceded by increased local gamma-band activity in widely separated cortical areas (including the precuneus), most notably in prefrontal and parietal regions, and that there was significant long-range synchronization of these oscillations. Moreover, they found that long-range gamma synchrony preceding a button press waxed and waned at intervals corresponding to the theta band. Explicitly evoking global workspace theory as an explanatory framework for their findings, they propose that the transient formation of large-scale neural coalitions is the basis for conscious experience.[55]

Compatible views are expressed by Gaillard, Dehaene and colleagues, who report the findings of a study using a visual masking paradigm.[56] Their study compared a conscious (unmasked) condition, wherein a word is presented to the subject long enough to be reportable, with an unconscious (masked) condition, wherein a word is presented too briefly to be reportable but for long enough to elicit priming effects. Because their subjects were neurosurgical patients, they were able to make intracranial EEG recordings, which enjoy a higher signal-to-noise ratio than the more common extracranial set-up.

In the unmasked condition, they saw a period of elevated gamma activation (averaged over 147 electrodes) beginning some 200 ms after the presentation of the word and lasting for around 200 ms, an effect that was lacking in the masked condition. They also found an increase in long-range synchrony during the same period, in the masked condition only, but in the beta band (13–30Hz) rather than the gamma band. Appealing to global workspace theory, the authors of the study interpret their findings as supportive of the

[54] Doesburg, *et al.* (2009).

[55] They claim that their findings support the view that 'consciousness emerges as a product of large-scale brain integration implemented by synchronization of relevant neural populations in the gamma band. We interpret this as reflecting the selective integration of information represented in relevant cortical regions into a large scale assembly that constitutes a global workspace for consciousness. Periodic activation of, and integration within, this network would thus correspond to the formation of a new large scale assembly defining conscious contents …' (Doesburg, *et al.* (2009), p.7).

[56] Gaillard, *et al.* (2009).

hypothesis that, in the conscious condition, a brain-scale reverberating state is ignited enabling an 'assembly of workspace neurons [to] distribute its contents to a great variety of other brain processors, thus making this information globally available'.[57]

To summarize, data from studies using a variety of experimental paradigms are suggestive of a distinctive increase in gamma synchrony during conscious perception, as well as long-range coherence (in the gamma or beta bands) between neural populations that are remote from each other. Although the evidence is sparse and preliminary, and is open to alternative interpretation, it is consistent with the proposal that there is widespread dissemination of influence and information during the conscious condition mediated by long-range neural synchrony. Moreover, there is also evidence (subject to the same caveats) that the conscious condition waxes and wanes at a frequency within the theta range, and that this correlates with an attendant fluctuation in both local synchrony and long-range coherence. All of this fits perfectly with the description of global workspace dynamics we have been promoting, on the assumption that what holds a coalition of brain processes together is synchronous oscillation, as postulated by the communication through coherence hypothesis.

[57] Gaillard, *et al.* (2009), p. 0473.

Chapter 6

The inner life

This chapter is dedicated to exploring the implications of the theoretical foundations laid down in Chapters 3 to 5. The focus of Chapter 3 was the conscious/unconscious distinction in the context of immediate perception and action. But a person's thoughts and feelings can stretch back into the past, can reach forward into the future, and can roam worlds of pure imagination. The neurodynamical account of Chapters 4 and 5 paved the way for a scientific approach of these issues by accommodating both externally driven and spontaneously driven activity within the global workspace. It's the latter that concerns us now. The chapter argues that an internal sensorimotor loop operating through the global workspace can account for foresight and planning, as well as episodic memory and other facets of the imagination. Along the way, certain theoretical positions in adjacent fields are appropriated, notably the view that conceptual blending is the foundation of abstract thought.

6.1 The simulation hypothesis

The material in Chapters 3 to 5 is intended, first and foremost, to establish an empirical basis for the distinction between consciously and unconsciously mediated *behaviour*. It is in the context of behaviour that the scientist gains the firmest grip on the conscious/unconscious distinction. But what does the theory have to say about thoughts and feelings that unfold internally, about those things that make up the *inner* life of a human being? In the global workspace architecture, according to the dynamics proposed in Chapters 4 and 5, the transition from one coalition of brain processes to the next is partially motivated by external perturbation (counting both incoming stimuli, such as the sight of a predator or the smell of food, and bodily sensations such as thirst, pain, or sexual desire). But the proposal also emphasizes internally driven activity, and allows for the spontaneous transition from one attractor to another. It's surely here, in the swirls and eddies of spontaneous activity, that we shall find the neuronal basis of thought that does not immediately issue in action.

The viewpoint to be developed here is based on the *simulation hypothesis*, according to which conscious thought can be accounted for in terms of simulated

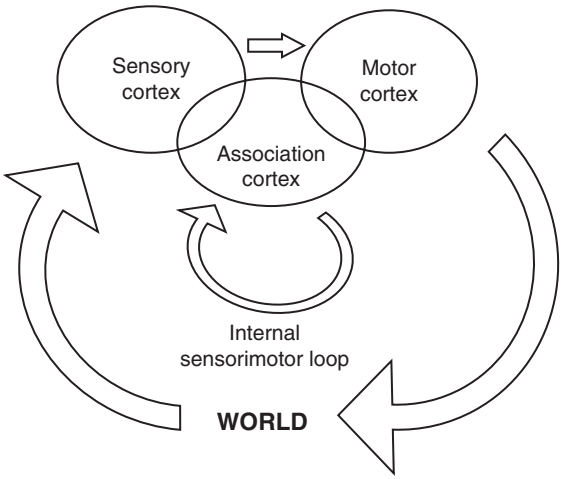

Fig. 6.1 The blueprint for an internally closed sensorimotor loop. Motor-cortical areas are assumed to be capable of generating activity that does not result in overt movement, and sensory-cortical areas are assumed to be capable of generating activity in the absence of an external stimulus. Finally, an internal loop realized by associative mechanisms ensures that internal motor activity elicits an internal sensory response. The upshot is an inner loop whose activity mimics that of the outer loop that is closed via the world.

interaction with the world.[1] The obvious benefit of internal simulation is that it allows an animal to rehearse its actions without the risks or costs that come with actual performance. From a neuroscience perspective, the central tenet of the hypothesis is that the brain's sensory and motor circuitry is capable of operating in an off-line mode, resulting in an internally closed sensorimotor loop (Fig. 6.1). Hesslow breaks this down into three assumptions.[2] First, the brain can generate motor-cortical activity without the production of overt movement. Second, the brain can generate sensory-cortical activity without the presence of external stimuli. Third, associative mechanisms exist whereby

[1] Hesslow (1994; 2002); Cotterill (1998). (We shall take care to avoid the claim that conscious thought *is* simulated interaction with the environment, with its philosophically provocative use of the word 'is'.) Near relatives of the simulation hypothesis have been articulated by other authors. It was anticipated by Craik (1943), for example, who considers the hypothesis that the brain 'imitates or models external processes' (p. 53). More recently, related ideas have been advanced by Grush (2004), and Barsalou (1999; 2009), among others. The present proposal is what Grush would call an 'emulation theory'.

[2] Hesslow (2002).

internal motor activity can loop back and cause sensory activity resembling that elicited by actual behaviour.

A more liberal formulation of the simulation hypothesis is obtained if we relax the requirement that the very same circuitry is activated during both simulated and real interactions with the environment, and admit the possibility that some or all of this circuitry is instead replicated to obtain an internalized sensorimotor system that operates in parallel with the outer loop. According to this formulation, the circuitry underlying simulated interaction only shares a blueprint with the circuitry underlying real interaction, and may or may not share any amount of wetware. At one end of the spectrum of possibilities allowed for by the liberal version of the simulation hypothesis, we have what we might term the *reuse* version, which insists on maximal wetware overlap. At the other end, we have what we might term the *replication* version, which insists on zero wetware overlap. In between we have a continuous range of intermediate possibilities.

One advantage of the replication version over the reuse version is that it allows for the dissociation of imagery and perception. It has been shown that brain lesions can compromise a subject's perceptual capacities while their visual imagery remains intact.[3] Another advantage is that it helps to account for what Nichols and Stich call *cognitive quarantine*, the ability to keep fantasy and reality separate.[4] If there is commonality of design but negligible actual overlap between the machinery of the inner loop and the outer loop then this separation needs little further explanation. The disadvantage of the replication version is that it makes it hard to explain apparent breakdowns of cognitive quarantine, as in the auditory hallucinations of schizophrenic patients. Fortunately, we need only embrace the hypothesis in its most liberal guise to proceed.

Evidence that favours the simulation hypothesis, liberally construed, has been obtained using various paradigms. First, studies using chronometric methods based on the mental rotation experiments pioneered by Shepard and Metzler support the thesis that the mechanisms underlying imagined spatial operations are temporally constrained in the same way as their physical counterparts.[5] Second, imaging studies reveal substantial overlap between the regions of the brain that are active during the performance of a task in actual and imagined conditions. For example, in one fMRI motor imagery study, researchers examined fronto-parietal activation in pianists, who were asked

[3] Bartolomeo (2008).

[4] Nichols & Stich (2000). Nichols and Stich are primarily concerned with pretend play.

[5] For example, see Borst & Kosslyn (2008).

both to imagine playing a certain piece, and actually to play it (on a silent keyboard).[6] They found significant overlap between the regions displaying increased activation in the two conditions. Similar findings have been reported in fMRI visual imagery studies.[7]

The simulation hypothesis covers not only interaction with the physical environment, but also interaction with the social environment. The ability to rehearse an encounter with a peer is no less beneficial than the ability to rehearse the manipulation of an object. However, the behaviour of another animal is not governed by simple physics like the behaviour of an inanimate object. Instead, it depends on what the other animal wants and what it knows. If it is hungry and has seen food, for example, it will move towards that food. Indeed it will do so even if it has to move uphill, in defiance of gravity, which clearly marks its behaviour out from that of a rock. On the other hand, if the animal has not seen the food it will go about other business, and it's advantageous to a rival to be able predict this difference.[8]

One way to internally simulate the behaviour of another animal is to take its perspective, to imagine being 'in its shoes', and then to carry out an egocentric rehearsal. This can be regarded as the social aspect of the simulation hypothesis. The proposal that our understanding of others is the result of such an ability is known as the *simulation theory* of mind-reading, and it stands in opposition to the so-called *theory theory* which posits a more classically representational substrate.[9] The discovery in monkeys of so-called mirror neurons, which are active both when the monkey performs a particular action and when it witnesses another monkey perform the same action, is evidence of the sort of overlap in brain areas that the social aspect of the simulation hypothesis predicts.[10]

Hopefully the claim that the *brain* engages in 'simulated interaction with the environment' is by now fairly clear. But little of what we have so far said obviously concerns consciousness. What exactly is the simulation hypothesis claiming about the inner life, about phenomenology, and what are the grounds for that claim? By way of counterpoint, let's briefly consider the mental imagery debate between Kosslyn and Pylyshyn.[11] The battleground here is the

[6] Meister, *et al.* (2004). The keyboard was silent because all metal components had to be removed from it so that it could be used inside the MRI scanner.

[7] See Ganis, *et al.* (2004), for example.

[8] For example, a series of experiments by Hare, *et al.* (2001) show that a chimpanzee's behaviour is sensitive to what it has seen that a rival has seen.

[9] Gordon (1986).

[10] Decety & Grèzes (2006); Galesse (2007).

[11] This debate has been ongoing for decades. See Kosslyn (1981) versus Pylyshyn (1981) for an early skirmish. More recently we have Pylyshyn (2003) versus Kosslyn, *et al.* (2003).

supposed 'format' of the 'representations that give rise to the experience' of visual imagery. Drawing on neuroscientific evidence of the sort cited earlier, Kosslyn ventures that these representations are, in a specific sense, picture-like. Pylyshyn dismisses this. Making a sharp distinction between the conscious experience and the mechanisms that might underlie it, he argues that 'far from supporting the picture theory, the results of imagery experiments tell us nothing about the format of images'.[12]

It's important to see that the simulation hypothesis, in the way it has been characterized here, does not take sides in this debate. It entails nothing about the 'format of the representations' underlying mental imagery, or about the 'format' of the presumed images themselves. It speaks only of simulated interaction, and makes no mention of representations or their alleged format. It does make the assumption that similar neurological mechanisms underlie both actual and simulated interactions with the world, and its chief claim is that conscious thought arises from the latter. But it says nothing about what those mechanisms might be. Moreover, there is no implication that the neurodynamics of simulated interaction faithfully reproduces the neurodynamics of real interaction. On the contrary, we might expect substantial differences to show up between sensorimotor activity that is externally caused and sensorimotor activity that is internally driven. And if conscious thought really arises from the latter, it will be no surprise if subjects (under careful interrogation) report their experiences of imagery, rehearsal, and inner speech to be fragmentary, hazy, and so on.

Our aim now is to defend the claim that conscious thought arises from simulated interaction with the environment, and the strategy will be first to reconcile the proposed internal sensorimotor loop with the global workspace architecture, and second to extrapolate global workspace theory's account of the conscious/unconscious distinction for overt sensory and motor events to the case of internally generated sensorimotor activity. We'll return to the first issue shortly. In the mean time, we need to establish the methodological credibility of the second goal. In particular, how could the scientist get a sufficient grip on the phenomenology of thinking to verify the extrapolated account empirically? What are the measurable, objective correlates of conscious thought if it does not issue in action?

In the case of external stimulation, introspective reports can be verified by marrying them with observable events. A subject who reports hearing a tone through headphones when and only when there is in fact a tone can be considered to be providing a reliable window into his consciousness. No such

[12] Pylyshyn (2003), p. 117.

paradigm is available to a scientist who wants to study conscious thought (not even the omnipotent psychologist). Other paradigms are at her disposal, though. A subject can be shown a drawing and shortly afterwards told to imagine it, or asked to learn a piano piece and then to imagine playing it (if he is a pianist, that is), or instructed to learn a passage of text and then to recite it inwardly and silently. In each case, the hope is that exercise of the imagination will reliably correlate with neurological markers obtained from whatever imaging and/or recording methods are available.

For these paradigms to work, it must be taken for granted that the subject carries out the experimenter's instructions faithfully, and that the subject's conscious, inner life is perturbed accordingly. But there are no unconscious influences on behaviour, because there is no behaviour, so the target for the scientist now is not a conscious/unconscious distinction as it was before. Rather, the aim is to extrapolate from the theory already established. This theory (we are supposing) will have identified the brain's global neuronal workspace and confirmed that reportable sensory and motor events match up with significant workspace activation, whereas unreportable events that nevertheless influence behaviour match up with local activity that fails to impact on the workspace. The theory will also have established mappings between particular workspace patterns and the sensory and motor events they correspond to.

Now, each way of exercising the imagination will be expected to correlate with a distinctive pattern of workspace activation, confirming the role of the workspace in consciousness even in the absence of outward behaviour. Moreover, to the extent that the patterns in question resemble those of the corresponding sensory and motor events, support will be lent to the simulation hypothesis. It will then be a reasonable step to extrapolate from patterns of workspace activation that correlate with observable sensory and motor events, and patterns of workspace activation that correlate with cues that are assumed to elicit related sensorimotor imagery, to spontaneously evolving patterns of workspace activation of the same kind, the presumed neurological signature of unfolding conscious thought.

The sceptic at this point might question whether the scientist can really conclude anything at all about the subjective, inner life of her subject from these objective markers. After all, the inner life is private. But the sceptic would be in the grip of dualism. Nothing is hidden. The subject is not in a *metaphysically* privileged position with respect to his inner life compared to fellow inhabitants of his shared world—the scientist, say, who can observe his brain and record what he says. There is no better theory than one that accounts for whatever is public, for whatever is manifest in the world, whether through language or through neurological phenomena that can be detected and measured with a scientific instrument. Is something left unaccounted for by such a theory?

Or is nothing left unaccounted for? This is not a dichotomy we should care to address. We need only remind ourselves that a nothing would serve as well as a something about which nothing can be said.

6.2 Simulation through a global workspace

The question now is how to accommodate an internally closed sensorimotor loop within the global workspace architecture.[13] This is necessary to extend the reach of global workspace theory, and to fulfil a prerequisite for the research programme envisioned in the previous section. The basic move is to equate the activity of the inner sensorimotor loop with the spontaneous, internally driven movement from attractor to attractor posited in the previous chapter. Or, to put it another way, the challenge is to see how this spontaneous activity might be modulated and shaped in such a way as to realize internally simulated interaction with the world.

Thankfully this is a fairly straightforward job. There is ample connectivity within the network topology already proposed to allow influence and information to circulate continuously between sensory and motor areas via the connective core (Fig. 6.2). To see how this might work, let's consider what might take place in the brain of an animal—a rook or a crow, perhaps, but there's no need to specify—who is confronted with a simplified version of the trap-tube apparatus (Fig. 6.3). There's a food item lodged inside the transparent tube, and a plunger that can be pulled left or right to move it. But there's no trap, only a bung in one end of the tube, which means that the animal has to pull the plunger to the right rather than the left to gain the reward. An animal that had already learned this task through trial and error would be able to obtain the food without unduly exercising its cognitive faculties. But let's suppose the animal has never seen the apparatus before, and resorts to internal rehearsal to anticipate the outcome of its actions prior to actually carrying them out.

The animal's first glimpse of the tube and its contents results in sensory-cortical activity (Fig. 6.3, left).[14] This pattern of activation achieves global

[13] The marriage of these two architectural elements was described in Shanahan (2006), which presents a computer model conforming to both global workspace theory and the simulation hypothesis. The resulting blueprint is similar to the one drawn up by Carruthers (2006), although Carruthers assigns a more significant role to language in his proposal.

[14] The pictorial icons used to indicate activity in sensory and motor cortex in Fig. 6.3 should not be taken to imply anything about the character of that activity. No commitment to picture-like representations is intended. What arises in the sensory and motor regions shown is a spatiotemporal pattern of neural activation, and exactly how this codes sensory and motor information is left as an open question.

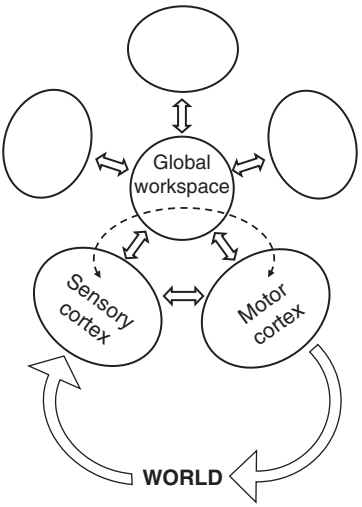

Fig. 6.2 Accommodating an internal sensorimotor loop within the global workspace architecture. The architecture's connectivity permits the continuous circulation of influence and information between sensory and motor areas through the global workspace, without overt behaviour. Direct connections (not mediated by the global workspace) are also present.

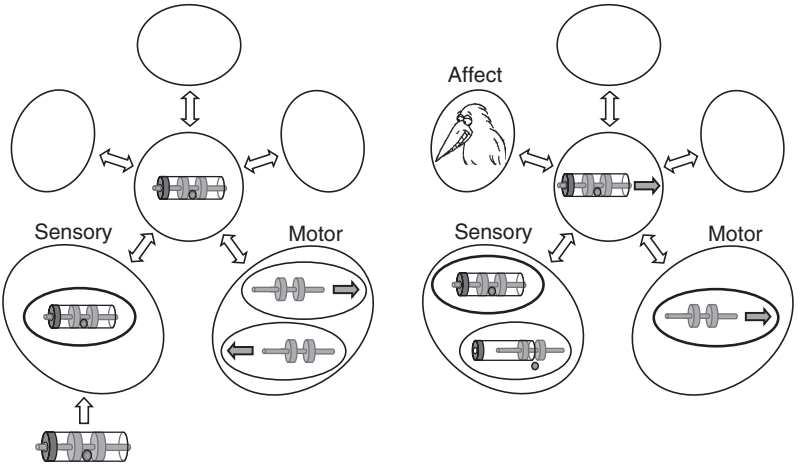

Fig. 6.3 Internal rehearsal through the global workspace. Left: the animal sees the tube with the food item lodged inside, and (the influence of) this visual stimulus is broadcast via the global workspace. Right: A sensorimotor coalition involving pulling to the right dominates the workspace. But it doesn't result in overt action. Rather, it causes sensory-cortical activity corresponding to the expected outcome of the action.

influence thanks to the workspace, reaching (among other brain regions) motor cortex. (A more familiar stimulus might never have made it to the workspace, and elicited a habitual motor-cortical response by a more direct route. But this is a novel situation.) Two rival motor processes respond to the broadcast, corresponding to the immediate affordances of the plunger, namely pull left and pull right. A competition between these processes ensues. The pull-right process wins and, together with the sensory process responding to the visual stimulus, forms a coalition that dominates the workspace (Fig. 6.3, right). In other circumstances, this might result immediately in motor output. But in this instance, the motor-cortical activity remains internal, and elicits an internalized sensory-cortical response resembling the activity that would arise if the animal had actually performed the action. A subsequent snapshot of the architecture might show the influence of this internally generated sensory-cortical activity being disseminated via the workspace, and this in turn might elicit an affective response to the expected outcome. In this case, the outcome is good, the affective response would be positive, and the animal would be motivated actually to execute the rehearsed action.

It should be clear that this example conforms to the dynamical description of the previous chapter. Coalitions of sensory and motor processes are the (quasi-) attractors of the system, and for the system to realize an internally closed sensorimotor loop, these processes must be detached from the outside world. The system is itinerant, visiting a series of attractors corresponding to waypoints along a trajectory through sensorimotor space. One coalition (attractor) will dominate the workspace for a while. Then, without external perturbation, thanks to the system's internal dynamics only, this coalition will break up, and a new one will take over the workspace. The transitions from one attractor to another, from one sensorimotor waypoint to another, are governed by associations formed during actual behaviour. An association, in this light, is nothing more than the tendency for one pattern of activation to follow another, for one coalition of sensory and motor processes to succeed another. If these tendencies are acquired by interacting with the environment (through neural plasticity) then they will recapitulate the sensorimotor trajectories that occurred during that interaction, and simulated interaction will occur when the brain's dynamics is allowed to freewheel.

The benefits of allowing the brain to freewheel in this controlled way are considerable. If an animal can hold back from immediate action, and anticipate the likely outcome of a behaviour before selecting it, then it can avoid dangers it would otherwise have been prey to and take opportunities it would otherwise have missed. The survival value of such an ability is obvious (although the animal has to avoid getting stuck in a thinking rut, carefully weighing its

options while a predator rapidly approaches).[15] More generally, the ability to rehearse interaction with the environment permits the animal to explore its space of affordances, filling in blindspots and exposing regions of affordance that were previously hidden, which is one of the primary functions of cognition (as set out in Chapter 2).

We have been concentrating on the conjunction of an internally closed sensorimotor loop with a global workspace. But the latter isn't a requirement for the former. An internal sensorimotor loop might be realized by direct connections between sensory and motor processes, without involving a broadcast mechanism. So what are the ramifications of the combination? First, according to global workspace theory, only rehearsal that proceeds via the global workspace can contribute to phenomenology. Internally simulated interaction with the environment that operated by direct reciprocal connections between sensory and motor areas would (if it took place at all) be unconscious. Second, according to the dynamical characterization of global workspace architecture on offer here, only an exploration that proceeds via the workspace is capable, not merely of exposing hidden regions of the space of possible affordances, but of actually extending it, of opening out whole new vistas of affordance by exploiting its open-endedness.

6.3 **Open-ended affordance**

The reason a global workspace so dramatically extends the functionality of an internal sensorimotor loop is that it facilitates the exchange of information and influence among processes that would otherwise have no means of communication, processes that arose to meet a specific demand in the evolution or the lifetime of the animal and whose usual coalition partners were fixed accordingly. Such specialists can transcend their specialism and form new partnerships if they have access to a communications backbone capable of flexibly opening and closing a channel between any two peers, like a telephone exchange. According to the present theory, the global workspace fulfils this function as well as that of broadcast. Indeed, these two functions go hand-in-hand in a network conforming to the topological constraints that have been proposed, one in which information and influence funnel into and fan out from a limited capacity connective core. Such a system has the means to enact, and therefore to simulate, behaviours that are entirely new to the animal.

Let's consider the example from Chapter 2 (Section 2.2) in which a schoolboy realizes that he can use a pile of building bricks to manufacture a missile for his catapult (Fig. 2.1). This unusual affordance of the pile of bricks is not

[15] See Dawkins (1976), pp. 57–60.

immediately apparent. He sees the pile of bricks in front of him, and the influence of this visual stimulus is duly disseminated via the global workspace (Fig. 6.4, left). Various motor processes are aroused (just two of which are shown in the figure), corresponding to known affordances of the bricks. A competition between them ensues, which is won by the make-s-shape process. However, being especially lazy, the schoolboy doesn't execute this action right away as it fails to elicit a sufficiently positive affective reaction. Instead he continues to imagine. The consequences of building the s-shaped assembly are anticipated in sensory cortex (Fig. 6.4, right), and the resulting pattern of activation gains global influence thanks to the workspace. (This episode of broadcast is not shown in the figure.)

Now, the s-shaped assembly is also familiar. Its various affordances (just one of which is shown) might lead to the construction of a wall and then to a house. But this prospect is unable to produce a significant affective response, and the schoolboy remains inactive. However, another motor process has also

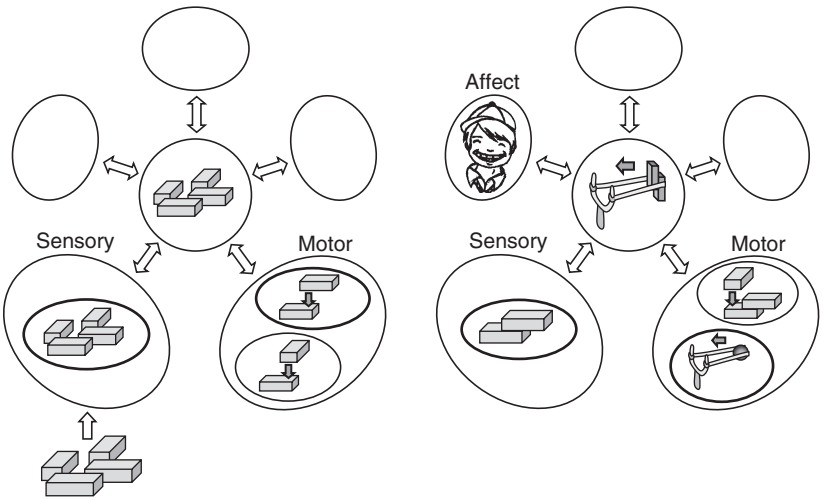

Fig. 6.4 Opening out the space of affordances. Left: the schoolboy is confronted with a set of building bricks. These have various common affordances, reflected in competing motor responses (two are shown). Right: after one round of rehearsal, the schoolboy imagines an s-shaped assembly of bricks. This has further obvious affordances (one is shown). But because information about the imagined assembly was broadcast, and because it has approximately the right size and shape, a load–catapult motor process is also aroused. If the imagined assembly is substituted for the ball usually implicated in the load–catapult process, the schoolboy will come to see the bricks as a potential missile, opening up a whole new region of affordance (no doubt with unwelcome results).

responded, namely load–catapult. This process is normally mobilized in a coalition with various visual and tactile processes specialized for the recognition and handling of the small (safe) foam balls that were provided for his new toy by the manufacturer. But thanks to the broadcast mechanism of the global workspace, the load–catapult process gets the opportunity to evaluate its own relevance to the situation at hand (or, in this case, the situation in mind). The s-shaped assembly is not dissimilar to the foam balls. It's just the right size, and its shape is a good enough approximation—neither too long nor too thin, neither too knobbly nor too spiky. It will fit the catapult, and will serve as a missile.

This, of course, is not a judgement at the personal level, but simply the merging of two compatible patterns of neuronal activation, one in sensory cortex and one in motor cortex. These processes are spatially separated and have never before worked in concert, but they can become coupled thanks to the connective core. If the sensorimotor merge is successful, the coalition that results will come to dominate that core, at least transiently, and will therefore be able to exercise influence at the conscious level. Also, the patterns of activity that first seed the merge (in this case sensory activity corresponding to the imagined two-brick assembly) must attain global influence, so that all potentially relevant processes can respond. So they too must be conscious. But the machinery that first brings about the merge operates in the background. Moreover, once the new coalition becomes established through repeated invocation, direct connections among the component processes are likely to form. It will then be able to work at the unconscious level, where it can contribute to later sensorimotor merges of ever greater complexity and abstraction.

Back in the schoolboy's head, the load–catapult process's level of activation is increasing. We can think of it as making a bid for access to the global workspace. Because it has little competition—or rather, because its competition has such a low level of affective support—it's in a good position to recruit allies. The result will be a new coalition, a sensorimotor blend whose ingredients will include processes for recognizing, tracking, and manipulating assemblies of building bricks, as well as processes for grasping, loading, aiming, and releasing catapults. The resulting coalition wins control of the connective core and its influence is disseminated accordingly (Fig. 6.4, right). The affective response is highly positive, and the schoolboy finally springs into action, much to the displeasure of the household cat.

The basis for the sensorimotor blend in this example is the substitution of one object for another. The assembly of bricks substitutes for the foam ball in the load–catapult process, and the schoolboy is thereby able to see the bricks as a missile. This capacity for 'seeing as', for seeing one kind of object as another by allowing it to fulfil the other's role in some behaviour (or whole nexus of behaviours), is what permits two processes to form an alliance even though

their original specialisms are distinct. However, for this to be possible, a degree of compatibility between the processes is required. The substitution has to be feasible. Even the naughtiest schoolboy wouldn't entertain the possibility of using a garden rake as a missile in a toy catapult. The size of the rake renders it ineligible as a substitute for the foam balls (although there is nothing to stop the boy from imagining a very large catapult to suit).

A prerequisite for a system that can effect a sensorimotor merge is that its constituent processes must be amenable to recombination. That is to say, the processes that become coupled when a coalition forms must also be capable of being active independently, of being decoupled from each other. In the naughty schoolboy example, we are envisaging a load–catapult process that normally couples with processes specialized for finding, recognizing, grasping, and manipulating foam balls in a particular way. (These processes are assumed to operate by recruiting and modulating even more basic processes for visual search, reaching, and so on.) If these were not separate processes, no decoupling of the act of loading and the object being loaded would be possible, and there would be no way to effect the substitution of the assembly of bricks for the foam ball.

A female cricket who uses phonotaxis to steer towards a potential partner cannot learn to use the same process to steer towards a food item that emits a distinctive noise, because the sensory activity that detects the sound of a male is indivisible from the motor activity that reacts to that sound. Behaviour must be the product of multiple processes working in concert for the very idea of recombination to be applicable. For recombination to be feasible in a particular instance, it must be possible to plug the candidate processes together. This pluggability constraint is met, for example, when one process realizes an action while another process procures and makes available the object to be acted upon. Similarly, if one process realizes an action in which one object (such as a stick or a hammer stone) is applied to another (such as a food pellet or a flint core), then potential flexibility is maximized when the processes for handling both objects are pluggable in this way.

6.4 **Conceptual blending**

The sensorimotor blend that takes place in the schoolboy's head and the slight opening out of his space of affordances that results hardly qualify as a conceptual breakthrough. Nevertheless, a convincing case can be made that the neural mechanisms implicated in the schoolboy's feat are no different in kind from those that led to such cultural innovations as religious ritual, money, mathematics, and poetry. For a shaman to see himself as an animal, for a shopper to see a 5-pound note as desirable, for a child learning algebra to see a variable as

a container, for a budding poet to see life as a journey—each of these achievements can be thought of in terms of what Fauconnier and Turner call *conceptual blending*. Taking their cue from the seminal work of Lakoff and Johnson in the 1980s, Fauconnier and Turner conceive of the whole edifice of human thought as constructed, layer upon layer, out of mappings (blends) between 'mental spaces'.[16]

To illustrate Fauconnier and Turner's framework, and to compare it with the account on offer here, let's consider how a creature might learn to retrieve food from a trap tube in such a way as to be able to transfer what it has learned to a novel variation of the apparatus (Fig. 6.5). As before, it doesn't matter what animal we have in mind. Nor does it matter that the task in question could perhaps be solved in a more pedestrian way. Our imaginary subject will solve it by 'conceptual blending'. Specifically, it will combine its expertise in two micro-domains—pulling with sticks and dropping in holes. Perhaps its expertise in these domains is innate, or perhaps it has been learned. The important thing for the illustration is that the animal hasn't previously encountered a situation in which both sticks and holes need to be dealt with at the same time. Nor does it have the benefit of a generic facility for physical simulation, which it could simply run on the configuration of objects before it.

Now, suppose we are watching at a point when the animal undergoes a jump in competence with the original version of the task. We're not interested in actual performances on novel variations. All we're interested in is the understanding it acquires at this moment, an understanding that is transferable to variants but just now is being applied to the original apparatus. Our subject sees the tube. The influence of the (still novel and challenging) visual stimulus is duly disseminated via the global workspace (Fig. 6.5, left). The plunger affords pulling left and pulling right, and motor processes corresponding to both possibilities will be aroused, as in the earlier example. But neither process has further associations at present (and neither is depicted in the figure). However, two other motor processes also exhibit a response. The trap part of the trap tube is reminiscent of the sort of hole the animal sometimes drops food into (sometimes by accident and sometimes for caching purposes), and the plunger resembles the sort of twig the animal sometimes uses to extract insects from otherwise inaccessible crevices.

As these processes are compatible, they are eligible to co-operate rather than compete, and a novel coalition arises in much the same way as it did in the schoolboy scenario (Fig. 6.5, right). In Fauconnier and Turner's terms, the animal effects a conceptual blend between three 'mental spaces'—a holes

[16] Fauconnier & Turner (2002). See also Lakoff & Johnson (1980) and Turner (1996).

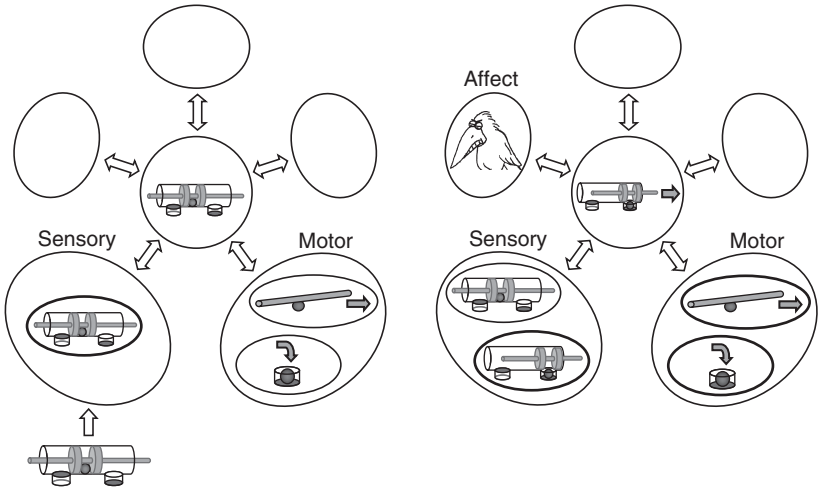

Fig. 6.5 Conceptual blending in the framework of global workspace theory. Left: the influence of the unfamiliar visual stimulus of the trap tube is broadcast, and elicits responses from two motor processes – one that normally deals with sticks and one that normally deals with holes. Right: The influence of these two compatible processes is blended, resulting in internally generated sensory activity resembling the anticipated disagreeable effect of pulling the plunger to the right.

mental space, a sticks mental space, and the mental space of the trap tube. The blend requires the stick in the sticks mental space to be mapped onto the plunger in the trap-tube mental space, the hole in the holes mental space to be mapped onto the trap in the trap-tube mental space, and the food items in all three mental spaces to be mapped to each other. In the emergent structure of the blended mental space, the plunger acts like a stick, and the trap acts like a hole, so that when the blend is run and the plunger is pulled to the right it drags the food into the trap where, as the animal has already learned to its cost, it is irretrievably lost. Running the blend therefore enables the animal correctly to predict the unfavourable outcome and avoid the wrong action. Moreover, the blend is transferable. It enables the animal to make similar predictions in different but analogous circumstances, unlike a purely associative rule, yet it manages this without simulating the physics of the situation in full.

It should be clear from this example that there are no glaring incompatibilities between the global workspace account we have been developing and Fauconnier and Turner's conceptual blending framework. (We might say that a conceptual blend between these two meta-level mental spaces is straightforwardly produced.) In fact, they can be regarded as mutually enriching ideas. The present combination of global workspace theory with the simulation

hypothesis complements the theory of conceptual blending by supplying detail at the level of neural connectivity and neurodynamics, and by assimilating it within an account of consciousness.[17] The theory of conceptual blending complements the present account by showing how its reach might extend to every corner of human culture.

Despite their affinities, there are differences of emphasis between the two explanatory frameworks. In particular, the present theory assigns a prominent role to consciousness. Both accounts are in agreement that the machinations of blending (the formation of novel coalitions) are not accessible to consciousness. That is to say, subjects are not able to report or reflect on their workings. Nevertheless, as Fauconnier and Turner affirm, we 'live in the blend'.

> Consciousness is dedicated to the products of blending. Once we have the blend of money or the [wrist]watch or social action or ritual, we are not consciously aware of the different input spaces and the projections across the network. In the blend, the money has its value, the watch shows the time, and this is what we are aware of.[18]

But according to the present account, consciousness—or more precisely the connectivity and dynamics that underlie the conscious/unconscious distinction and the capacity for integration that goes hand-in-hand with it—is the necessary means for realizing the sort of conceptual blending that, according to Fauconnier and Turner, has such far-reaching implications. The connective core, we might say, is the 'blender'. This is not to say that blending itself is a conscious operation. Rather, it is to claim that conscious awareness of the ingredients of the blend is required before the processes that will contribute to the novel coalitions implicated in the blend can be brought together in the brain. This, indeed, is a major evolutionary advantage conferred by those mechanisms.

6.5 **Cognitive fluidity and the frame problem**

One of the grand challenges in archaeology is to understand the explosion in human technology and culture that took place in the transition from the Lower Palaeolithic to the Upper Palaeolithic period, some 30,000 to 40,000 years ago. In this relatively short interval—the blink of an eye in evolutionary terms—there is a dramatic increase in the sophistication of human tools, and at the same time the first irruptions of art and ritual. Why did this occur? One place to look for an answer is the human brain. According to Mithen, an evolutionary

[17] As far as the likely neural substrate of conceptual blending is concerned, Fauconnier and Turner (2002) assert that: '… elements in mental spaces correspond to activated neuronal assemblies and linking between elements corresponds to some kind of neurobiological binding, such as co-activation.' (p. 102).

[18] Fauconnier & Turner (2002), p.391.

adjustment to the architecture of the hominid brain took place at this time that seeded the development of modern humans and their distinctive mental capacities.[19] The effect of this adjustment was to facilitate what Mithen calls *cognitive fluidity*, which enables 'thoughts and knowledge generated by specialized intelligences [to] flow freely around the mind', resulting in 'an almost limitless capacity for imagination'.[20] Cognitive fluidity, as Fauconnier and Turner acknowledge, is what permits conceptual blending.[21] It is also one provision of the global neuronal workspace—its integrative facility.[22]

Mithen situates the concept of cognitive fluidity in relation to *modular* theories of the mind, and our understanding of the global neuronal workspace would be incomplete if we didn't follow his example. Modular theories come in many flavours, but each of them supposes that some portion of the mind can be divided into parts (modules), and that the parts are functionally specialized in some way. (The term 'module' is used in a different way here from the way it is used in the theory of networks, although a module in the network sense could realize a module in the present sense.) A recurring theme among modular theorists is the limits of modularity, and there is a widely perceived need for some means of going beyond the constraints of specialization and transcending modular boundaries.

The presentation in Fodor's 1983 book *The Modularity of Mind* was particularly influential.[23] Fodor distinguishes between the mind's *peripheral processes*, specialized modules that handle such operations as low-level vision and parsing, and its *central processes*, which are implicated in higher cognitive functions. Fodor itemizes a number of features of each type of process. One property of peripheral processes is *informational encapsulation*, which means that they draw on information from a fixed set of sources (assumed to be small in number). Central processes, by contrast, are *informationally unencapsulated*, meaning they can draw on information from any source. Analogical reasoning epitomizes informational unencapsulation, as its very essence is the establishment of mappings between domains previously considered unrelated.

Another informationally unencapsulated process, in Fodor's way of thinking, is 'belief update'. When the world changes, there is no way to circumscribe the domain of the set of beliefs that will require modification to match. The concept of belief doesn't feature in the theoretical edifice we have been erecting

[19] Mithen (1996).

[20] Mithen (1996), p. 71.

[21] Fauconnier & Turner (2002), pp. 74–75.

[22] Shanahan & Baars (2005).

[23] Fodor (1983). See also the updated discussion in Fodor (2000).

here, and as a consequence nor does belief update. Instead of belief, we have the dynamics of interacting brain processes and the behaviour they produce. Belief update corresponds to various forms of neural plasticity. Fodor's notion of belief assumes a language-like representational substrate that is similarly absent from the present standpoint.[24] Nevertheless, it's worth temporarily taking up Fodor's perspective in order to engage with the *frame problem*, as it has been conceived by philosophers of mind. This will eventually lead us back to the issue of cognitive fluidity.

The frame problem, in its original formulation, was discovered by artificial intelligence researchers in the late 1960s who were attempting to formalize the effects of actions in mathematical logic.[25] The difficulty for the AI researchers was how to avoid having to represent explicitly everything that does *not* change when an action is performed. Philosophers such as Fodor, Dennett, and Dreyfus interpreted this as an instance of a larger problem, namely how an informationally unencapsulated cognitive process that worked by carrying out computations over representations could ever determine that its job was done, that it had taken into account all the information relevant to its task.[26] For informationally encapsulated operations such as low-level vision, this isn't much of a problem. When all parts of the visual field have been inspected for edges, there is no need to look elsewhere. But for an informationally unencapsulated process, so the argument goes, there is no end to what could be taken into account. In the case of belief update, the question is how to determine when all the relevant revisions to a set of beliefs have been effected.

> … it's *just got to be* possible to determine, with reasonable accuracy, the impact of adopting a new belief on one's prior epistemic commitments without having to survey those commitments in their totality. … The totality of one's epistemic commitments is *vastly* too large a space to have to search …[27]

[24] The present stance might be thought of as an example of *eliminative materialism* (Churchland, 1981). This would be wrong, however, insofar as eliminative materialism is taken to entail the view that the propositional mental states of everyday folk psychology (beliefs, desires, and intentions) do not actually exist. There is no need to discuss what does and does not actually exist. Beliefs, desires, and intentions are not required in our scientific vocabulary, and that is all that needs to be said on the matter.

[25] McCarthy & Hayes (1969); Shanahan (1997).

[26] Fodor (1983); Dennett (1984); Dreyfus (1991), pp. 115–121. See also Fodor (2000), Wheeler (2005; 2008), and Dreyfus (2008).

[27] Fodor (2000), p.31. See also Carruthers (2003). Fodor regards this difficulty as fatal for the ambitions of any research programme in cognitive science that is based on the idea of computation over representations, and as he cannot conceive of a viable alternative, he delivers a dire prognosis for the project of understanding the mind in scientific terms: 'it's a mystery, not just a problem, what model of the mind cognitive science ought to try next' (Fodor, 2000, p. 23).

Despite the fact that the paradigm we are working in here is not Fodor's, we cannot altogether escape his challenge. For us there is no computational threat associated with belief update as such. However, there is a sense in which conceptual blending, like its close relative analogical reasoning, is unencapsulated. In principle, any combination of processes might be mutually relevant and eligible to form a new coalition. To paraphrase Fodor, the totality of possible process combinations seems like 'vastly too large' a space to have to search. So even though the building blocks of our theory are brain processes and our theoretical focus is the dynamics of their interaction, the question still arises of how cognitive fluidity is actually realized. In an open-ended repertoire of coalitions, how does the brain single out the relevant ones?

Half the answer is scaffolding and the other half is architecture. We'll deal with scaffolding first. Genuinely spontaneous, creative acts of social significance, such as the production of the first wheel, the invention of paper, or the discovery of general relativity, are historically rare. The society we live in is the product of millennia of gradual development, and the conceptual blends we inhabit are sedimented layer upon layer. But as children we are inducted into a culture that is already there for us. In a few short years, a child has to recapitulate the entire history of conceptual innovation in the society into which it is born, and this involves thousands of individual acts of creativity and moments of insight. Thousands of remarkable conceptual blends must take place in its head if it is to learn how to talk, how to add up, how to tell the time, and how to use a computer. But this is not a journey the child undertakes alone. Our parents, carers, teachers, and peers *scaffold* our learning.[28] By setting us tasks and giving instruction, by using props and demonstrations, by providing encouragement and feedback, they bring the relevant brain processes into juxtaposition for us.

However, scaffolding is useless without the right architecture. No amount of scaffolding will enable a chimpanzee to apply Pythagoras's theorem, even if there are bananas involved. The child's cognitive endowment has to be one that, when presented with good instruction or apt demonstration, allows 'the penny to drop'. Moreover, often the best instruction induces the child to discover things for itself, because the conceptual blends that result tend to have wider application. So we must think of scaffolding as reducing the space of possible process combinations to manageable proportions, not as eliminating the need for search altogether. Despite the reduction in size, this space still appears forbiddingly large if we imagine having to search it one element at a time, checking for relevance to the present problem-solving context

[28] The concept of scaffolding was articulated by Wood, *et al.* (1976), whose characterization was anticipated by Vygotsky (1934/1986).

(whatever that might be) as we go (Fig. 6.6, top).[29] But the architecture of the brain is massively parallel, and thanks to the broadcast mechanism of the global workspace the responsibility for determining which processes are relevant to the problem-solving context can be distributed among the processes themselves (Fig. 6.6, bottom).[30]

So are we done with the frame problem? Not quite. If the business of singling out combinations of brain processes relevant to the ongoing problem-solving context were simply a matter of selecting from a repertoire of tried-and-tested coalitions then no further explanation would be needed. Processes would be alert to the situation at hand (or in mind) thanks to the global workspace, and relevant processes would bid for influence alongside their usual partners. (The circles in Fig. 6.6 (bottom) might then be thought of as complete coalitions.)

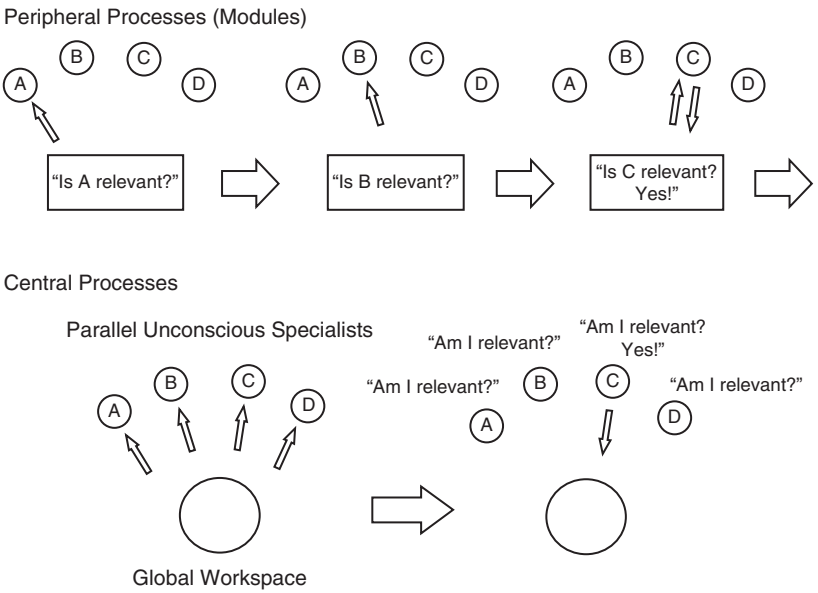

Fig. 6.6 Serial versus parallel approaches to relevance. Top: if serial processing is used to determine relevance then it looks computationally infeasible for informationally unencapsulated cognitive processes. Bottom: a parallel approach that takes advantage of the broadcast facility of the global workspace is more plausible. Although only a handful of processes are shown here, in a real brain they will be massively numerous.

[29] This serial view is encouraged by the robot thought experiment Dennett (1984) uses to highlight the frame problem.

[30] Shanahan & Baars (2005).

But where conceptual blending is concerned, *usual* partners are to be avoided. It's *unusual* partnerships we're looking for. The repertoire of coalitions potentially relevant to the problem-solving situation is open-ended. So the space to be searched is not the space of processes, but the space of combinations of processes, an altogether more daunting prospect.[31]

This is where other features of the global workspace architecture come into their own. Recall the proposal that the global workspace is realized by the bandwidth-limited connective core of a modular, small-world network and that information and influence funnel into and fan out from this connective core (Chapter 4). This network topology promotes dynamical complexity, a balance of segregated and integrated activity, which in turn supports the generation of a large repertoire of metastable states. Recall too that the global workspace is hypothesized to be not only the locus of broadcast, but also the arena for competition among competing process coalitions and the medium of coupling among coalition members (Chapter 5). Now, these architectural features may make an open-ended repertoire of coalitions *available*. But how do they help to isolate the *relevant* coalitions in that repertoire from the rest?

The answer lies in the dynamics of the bursts of competition that punctuate the workspace's episodes of broadcast. Negotiations between potential partners, the recruitment of new coalition members, the suppression of rival coalitions—all this goes on in parallel, until one coalition emerges from the melee as dominant (Fig. 5.3). The discovery of a novel competition is not a matter of trying out combinations, one at a time. Rather, every process that determines itself to be relevant to the ongoing situation, as broadcast, attempts to commandeer the workspace to broadcast its claim and thereby to initiate its own bid for power. There is no consideration by a central executive of the takeover bids of rival consortia, each carefully prepared in advance. Rather, the workspace is a battlefield in which alliances are made and broken on the fly, depending on the ebb and flow of battle, and in which participants give up of their own accord when the fight is going against them.

The deciding factor in this battle is relevance. Relevance to the situation at hand (or in mind) is what determines the strength of a coalition, its ability to recruit allies while subduing its opponents. Whatever longer-term, unconscious incubation may have taken place, wherein processes adjust their internal workings and external sensitivities, when its time arrives, a novel coalition

[31] Wheeler (2008) characterizes this as the 'inter-contextual' dimension to the frame problem, which is the challenge of saying 'how a purely mechanistic system might achieve appropriate, flexible and fluid action in worlds in which adaptation to new contexts is open-ended and in which the number of potential contexts is indeterminate' (p.340). Wheeler's hypothesis that the problem is overcome in systems that exhibit 'continuous reciprocal causation' is close to the present standpoint.

inaugurating a new conceptual blend will crystallize very rapidly, invading the workspace and disseminating its influence in a flash of insight. Far from being the provenance of great poets and mathematical geniuses, the present standpoint entails that such moments of insight are the engine of everyday development and learning.

6.6 **Space, time, and memory**

Integration is achieved when the brain's full battery of resources is brought to bear on the ongoing situation. For a while now, our emphasis has been sensory and motor processes that realize expertise in one micro-domain or another, and whose influence on behaviour or rehearsal is immediate. But the concept of integration also embraces processes whose influence is deferred, that is to say memory processes. Without committing to a sharp distinction, we shall follow the usual practice within cognitive psychology of supplying separate treatments of *working memory* and *episodic memory*. Both types of memory contribute to conceptual blending, and both furnish material to the internal sensorimotor loop. We'll begin with working memory, focusing on what, after Baddeley's influential model, might be called the 'visuospatial sketchpad'.[32] But we shall conceive of this faculty, in more neural terms, as a temporary pattern of activity that persists after the spatially structured stimulus that caused it has faded.[33]

To see a solid object, such as a cube on a table, is not simply to see a flat image, nor even (as we have stereoscopic vision) to see a raised surface with protrusions. It is to perceive everything that the three-dimensional solidity of the object affords—the possibility of picking it up and rotating it in the fingers, for example, or of moving around to the other side of it to see its back.[34] In a similar way, when a person looks at one scene (an office worker surveying the disorderly collection of objects on her desk, say), then turns her head to look at a different scene (the view through a window to the trees outside, say), there is a sense in which she remains aware of the continuing availability of the first scene even while admiring the second. All she needs to do is turn her head again for it to come back into view.

[32] Baddeley & Hitch (1974); Baddeley (2007). The functionality of the 'phonological loop' – another component of Baddeley's model – is subsumed by the internal sensorimotor loop in the present architecture.

[33] Wang (2001). Imaging and lesion studies classically implicate the prefrontal cortex in working memory tasks (Curtis & D'Esposito, 2003). But as D'Esposito (2007) states, 'working memory is not localized to a single brain region but probably is an emergent property of the functional interactions between the [prefrontal cortex] and the rest of the brain' (p.769).

[34] Compare the discussion by Merleau-Ponty (1945/1962), pp.235–239 & 283–347.

But there is an important difference between looking at a small solid object and exploring the wider setting. In the latter case, the affordances are constructed incrementally and on the fly, because they depend on how the local scenery unfolds, rather than being given all at once by the familiar three-dimensionality of the object before the eyes. Consequently, the newly discovered affordances (the possibility of picking up the stapler on the desk, say) have somehow to be retained when they are no longer apparent to the senses (because the office worker is now looking out of the window). This temporary retention of local affordances is one aspect of working memory function.[35] (In fact, the visual 'field' itself is not continuous like a field. Rather, it is constructed by a series of saccades and small head movements. So, although the overall three-dimensionality of a small object can be taken in with a single gaze, at a fine spatial scale even the cube on the table is subject to the same considerations as the wider scenery.)

We must take care not to over-inflate the claim that to see a solid object is to see what the three-dimensionality of that object affords. It would be imprudent to deny that the three-dimensional solidity of a cube contributes to the phenomenology of seeing it. Similarly, we have spoken of the continuing 'awareness' of objects that are no longer in view in terms of the persistence of their affordances. But our stance here does not entail that all the associated affordances are 'before the mind' when we see a solid object or explore our surroundings. We should reject this way of putting things, and recall that a feeling of plenitude is not a plenitude of feeling. All those affordances, all those possibilities for action, are merely *available* (looking behind the solid object, revisiting part of the scenery, and so on). Yet whenever a particular action is called to mind—during inner rehearsal, say, or in response to a question—there it is. Again we find that the refrigerator light illusion is a good metaphor.

In order to carry out useful off-line rehearsals of interaction with the physical world, the activity of the internal sensorimotor loop must respect the spatial structure of the world it is simulating.[36] This means that in rehearsal, as in real interaction, affordances must be constructed on the fly and retained. After an animal has effected a blend between the micro-domain of pulling with sticks and the micro-domain of dropping in holes in order to solve the trap tube (Fig. 6.5), it can only carry out rehearsal with the resulting coalition of processes (and run the blend) if it can keep track of the different parts of the imagined apparatus as they are manipulated in the simulation. In other words, we need to

[35] This is also one way to characterize the grasp of *object permanence* (Piaget, 1954; Baillargeon, 1993).

[36] The issues are somewhat different when it comes to interaction with the social world.

supplement the presuppositions of the simulation hypothesis—and by implication the presuppositions of the proposed marriage of the simulation hypothesis with global workspace theory—with the assumption that working memory, just like sensory and motor cortex, can function while disconnected from the world, and that its activity while off-line mimics its activity when on-line.

It's important to see that, although they enjoy significant interaction, working memory, as the term is used here, and the global workspace are *not the same thing*.[37] Indeed, working memory can be doubly dissociated from conscious experience (for which global workspace involvement is assumed to be a prerequisite). On the one hand, conscious experience is possible in which working memory is not implicated, as in the pain of a simple animal with negligible working memory capacity.[38] In such a case, global workspace theory predicts that the brain's pain centres have taken over its communications infrastructure in order to broadcast an alarm signal.[39] On the other hand, items can be retained in working memory without ongoing conscious involvement. Suppose our office worker notices that the stapler is on her desk (usually it's to be found on someone else's), then crosses to the photocopier where she is occupied for a time. When asked about the stapler by a colleague, she may well be able to report that it is on her desk, despite having had no thought of it while using the photocopier. (It might require the interventions of the omnipotent psychologist to establish this last fact empirically, but the possibility of its being true seems uncontroversial.)

Not only do we see the world as three-dimensional, we also experience a world that is in constant flux. If we see, say, a Frisbee gliding through the air towards us, we speak of an object in motion, not of a series of disjointed snapshots. As both James and Husserl might have said, it is as if our experience of the Frisbee included, at each moment, both a trace of its trajectory from the

[37] Baars & Franklin (2003). Of course, the term 'global workspace theory' here refers to the interpretation and development of Baars's original ideas on offer in the present chapters. Baddeley (2007, Chapter 16) offers a sympathetic treatment of the idea of a global workspace. But he does so by assimilating it to his own concept of working memory, and the result is somewhat different from the present conception of a largely passive communications infrastructure, although Baddeley attributes integrative capabilities to it that are similar to those at the heart of the present account.

[38] Consider fish, for example. A good case for the possibility of pain in fish is made by Braithwaite (2010). Working memory has also been demonstrated in fish (Hughes & Blight, 1999), but it's hard to see why such a spatial capability should be implicated in the capacity to feel pain.

[39] Indeed, the communications infrastructure required to broadcast such an alarm signal is a plausible evolutionary precursor to a full-blown global workspace.

immediate past and a projection of its trajectory into the immediate future.[40] Likewise, if we hear a familiar musical phrase, we are more likely to think of the tune it comes from than to dwell on the individual notes. From a behavioural point of view, we could not run to intercept a flying Frisbee if we had not already observed a fragment of its trajectory, nor whistle the fifth and sixth notes of the tune if we had not heard the preceding four.

Like the previous observations about space, these observations about time are suggestive of the operation of working memory, according to a slightly extended sense of that term as it is used here. Some remnant of the history of a time-varying sensory stimulus must linger briefly after the stimulus itself has gone, perhaps as persistent activity in some reverberating neural circuit. If such lingering patterns are in any way to inform phenomenology, to influence what it is like to be an animal, then they must exercise influence on the global workspace. Much as we can speak of the spatial organization of the information that is disseminated from the global workspace, so we can speak of its temporal organization. In short, the present theory is compatible with the notion that the persistent activity of working memory processes shapes the way we experience the world before us.

Nothing of what has been said here about the spatial and temporal character of the world as we find it is intended to be philosophically challenging. It's not an account of the 'structure of experience' in the manner of Kant or Husserl, but simply a statement of the obvious.[41] As such, if it fails to meet with universal approval then the fault is with its literary presentation. In Chapter 1, a similar theme was discussed, along with allusions to the idea of subjective experience disappearing into the abyss of time. But our purpose there was to dismantle a way of thinking that leads to a metaphysically burdened idea of subjectivity. Our purpose here is quite different. By juxtaposing such descriptions with related neurodynamical explanations we hope to illuminate the relationship between our inner lives and the physical world. While doing so, we strive to maintain our hard-won post-reflective silence.

6.7 **Remembering and reconstructing**

Working memory processes exercise their deferred influence on a relatively short timescale—over seconds, minutes, or perhaps hours. But human phenomenology is richly informed by a narrative of remembered events that spans

[40] Husserl (1911/1991); James (1890/1950), vol. 1, pp.608–610.

[41] Husserl (1911/1991); Kant (1781/1929). In this respect, James's poetically truthful treatment of such topics is to be preferred to the theoretically heavy approach of either of these philosophers.

a lifetime and reaches back into early childhood. This faculty for *episodic memory* is both *autobiographical* and, to use the term favoured by Tulving, *autonoetic*.[42] It is autobiographical because it concerns the events that make up a person's life story, the life story of the person whose memories are in question. But not all autobiographical memories are episodic memories. A person can typically remember where and when she was born without being able recall the actual event. Episodic memory is autonoetic because it involves the conscious replay of past experiences, whether recent, such as last night's party, or distant, such as a childhood visit to the seaside. A number of controversies attend the topic of episodic memory, so construed, and it's instructive to review them in the light of the theory we have been developing.

Because of the notorious difficulties associated with operationalizing the distinction between conscious and unconscious influences on behaviour—difficulties we have already explored in some detail for the case of immediate influence, and which are magnified when the influence is deferred—psychologists have been concerned to isolate purely behavioural criteria for episodic memory. One such criterion is that behaviour involving the exercise of episodic memory should be sensitive to *what*, *where*, and *when* facts about past events.[43] In telling a friend about her day, our office worker might describe what she said to her boss (that he was incompetent), where the pronouncement took place (in his office), as well as when (shortly after lunch). Applying this criterion allows researchers to look for episodic memory in non-human animals. Clayton and colleagues, for example, demonstrated that scrub jays, who cache food items for later recovery and consumption, modulate their recovery behaviour according to how long a food item has been cached and whether or not it is perishable as well as the location of the cache.[44]

But despite demonstrating an animal's sensitivity to the what, the where, and the when of its past actions, these researchers are open to the sceptic's charge that such sensitivity doesn't amount to *truly* episodic memory, because there is no evidence that it involves the conscious recall of a past event. It can equally be explained in terms of the deferred modulating influence of unconscious brain processes. The sceptic's point is a pertinent, because it would be hard to over-estimate the importance to a human of the ability to bring to mind significant events in her life, of being *consciously* in touch with her past.

[42] The distinction between semantic and episodic memory was first articulated by Tulving (1974), who later introduced the concept of autonoetic consciousness (Tulving, 1983). For a more recent overview see Tulving (2002).

[43] See Tulving (2002).

[44] Clayton & Dickinson (1998). Evidence of similar abilities has even been found for voles (Ferkin, *et al.*, 2008).

So animal researchers are caught between the Scylla of being unable to operationalize the conscious/unconscious distinction in non-human animals and the Charybdis of the requirement for autonoetic consciousness in 'truly' episodic memory. For this reason, they have conservatively resorted to characterizing their findings as evidence for 'episodic-like' memory.

Pushing the debate into new territory, Suddendorf and Corballis, citing the so-called Bischof–Köhler hypothesis that 'only humans can flexibly anticipate their own future mental states of need and act in the present to secure them',[45] recast episodic memory as one facet of a generic capacity for *mental time travel*. This capacity 'allows humans to mentally project themselves backwards in time to re-live, or forwards to pre-live, events'.[46] Neuroscientific evidence is supportive of their conception. Both imaging and clinical studies suggest that the ability to recall episodes from the past shares a common neural substrate with the ability to imagine future scenarios,[47] a 'core network' that has much in common with the so-called default mode network discussed in Chapter 4.[48] The default network, in turn, lies in the connective core that we have hypothesized as the substrate of a global neuronal workspace.[49]

Suddendorf and Corballis emphasize the combinatorial reach of mental time travel: 'given a basic vocabulary of actors, objects, and events, we can construct unique episodes in the past and create scenarios to deal with unique contingencies in the future'.[50] Reiterating this point, Schacter and Addis frame the *constructive episodic simulation hypothesis* that 'simulation of future episodes requires a system that can draw on the past in a manner that flexibly extracts and recombines elements of previous experiences'.[51] Like Suddendorf and Corballis, Schacter and Addis take a symmetrical view of the past and the future, treating mental time travel in both directions as a *constructive* process. Episodic memory, on this construal, is a matter of the *reassembly* of the elements of a past episode, as opposed to the *retrieval* of a complete trace of a past experience. This reconstructive view helps to account for the limitations of episodic memory, limitations that lawyers are adept at exposing when questioning a witness in court. However, no specific neural mechanism capable of

45 Suddendorf, *et al.* (2009), p. 1320.

46 Suddendorf & Corballis (2007), p. 299. The term was coined in Suddendorf & Corballis (1997).

47 Schacter, *et al.* (2008).

48 Buckner & Carroll, (2007).

49 Hagmann, *et al.* (2008). See the discussion in Chapter 5.

50 Suddendorf & Corballis (1997), p. 147.

51 Schacter & Addis (2007), p. 773.

putting together scenarios using fragments of past episodes as the raw material has so far been proposed by any of these authors.

Now, it will not have escaped the reader that the attributes of the global neuronal workspace, according to the advocated theory of its connectivity and neurodynamics, make it ideally suited to this job. To replay an episodic memory (on the reconstructive view), or to imagine a future scenario, is to assemble a coalition of active brain processes, a coalition that in this case involves memory processes. The challenge—the same challenge we confronted in Chapter 5—is to account for how such coalitions of arbitrary composition can be formed. That is to say, the repertoire of coalitions of brain (memory) processes involved in constructively imagining a future must be open-ended, not confined to the familiar or tried and tested. As before, our response is that this requirement is met in a parallel architecture with the right connective topology—one whose component processes form a modular, small-world network with a pronounced connective core. Such an architecture promotes metastability, chaotic itinerancy, and a balance of integrated and segregated activity, the hallmarks of a dynamical system capable of generating a large repertoire of coalitions.

Of course, the 'constructive episodic simulation hypothesis' resonates with Hesslow's more succinctly titled 'simulation hypothesis', and not just in name. For the global neuronal workspace to be a viable mechanism for blending the narrative elements of previous experience into novel imaginary future scenarios, it has to work in tandem with an internally closed sensorimotor loop in the way set out earlier in the chapter. Moreover, the simulation hypothesis can accommodate more than projection into the past or the future. Moving away from the notion of mental time travel, the idea of 'constructive simulation' can be widened further to embrace both augmented reality and fantasy.[52] When a child absorbed in pretend play pours imaginary tea from a real plastic teapot, there is no mental time travel involved. The *where* and the *when* of the episode are here and now. Only the *what* has been augmented by the imagination. Likewise, no mental time travel needs to take place when an adult enjoys a sexual fantasy. There may be no particular *when* or *where* to the fantasy, just an overwhelming *what*, and most likely a particular *who*. The combination of a global workspace with an internal sensorimotor loop is applicable to both cases.

[52] Hassabis & Maguire (2008). The term 'augmented reality' comes from the field of computer vision, and refers to a technique in which computer-generated imagery is mixed with real-time video data to produce a composite scene.

In short, episodic memory, alongside mental time travel and internal simulation in general, is naturally accounted for by the combination of a global workspace and an internal sensorimotor loop. A by-product of this accommodation, if we accept the tenets of global workspace theory, is the applicability of the conscious/unconscious distinction for which we have already established a theoretical basis. Thanks to this distinction, we can rehabilitate Tulving's original, intuitive distinction between semantic memory of past events and true episodic memory, the latter involving the conscious replay of the events in question (autonoetic consciousness) whereas the former requires only knowledge of the relevant facts. Because episodic memory, like mental time travel and other forms of internal simulation, involves the dissemination of influence and information via the global workspace, it contributes to the subject's phenomenology.

Armed with a principled way of dealing with the conscious/unconscious distinction, we are also in a position to revisit the question of mental time travel in non-human animals.[53] We cannot settle the debate, of course. That will require a good deal of further empirical work. But we can situate it within a broader theoretical context. For an animal to be capable of true mental time travel, of mental projection backwards or forward in time accompanied by autonoetic consciousness, the right neurological substrate is required. The structural connectivity of its brain should be modular, small-world, and possess a pronounced connective core. The associated dynamics should exhibit episodes of broadcast punctuated by bursts of competition, and the repertoire of process coalitions its brain generates should be open not closed. So in addition to directly addressing the question, we should look for evidence of such a neurological substrate, evidence that might come from imaging studies or from electrophysiological data, as well as from various behavioural paradigms.

6.8 **Talking to ourselves**

The topic of memory is a large one, and it's impossible to do justice to it in a small space. Now we are going to touch on the topic of language, which is larger still. But with language as with memory, the aim is only to highlight points of contact with the theoretical foundation we have been attempting to lay, because it would be cause for concern if these points of contact were missing.

[53] Supplementing existing evidence for episodic-like memory in non-human animals, Raby, *et al.* (2007), for example, showed that scrub-jays are capable of planning for their future needs, based on their past experience, and independently of their current motivational state. But Suddendorf, *et al.* (2009) profess scepticism that the evidence to date for mental time travel in non-human animals is conclusive.

Moreover, to tackle language from the standpoint of our theory is to effect a kind of closure, for language was also our point of departure. By taking to heart Wittgenstein's remarks on the nature of language, we were able to work through metaphysics and emerge in post-reflective silence. Language is something we do. Moreover, it is something we do together. If we want to understand the things people say to each other, we can do no better than to look at the role those things play in people's ordinary affairs, even if the people in question are philosophers and the things they say are strange and disquieting.

In Chapter 1, seeing language this way helped to demystify certain difficult words, such as 'experience', 'sensation', 'truth', and 'concept'. Now this way of seeing language takes its place in our overall scientific theory of cognition and consciousness. In this view, the use of language is just another form of interaction with the world—a world that not only contains the physical paraphernalia of everyday life, but that is also shared with a community of fellow language users. So if, as the simulation hypothesis proposes, thought arises from simulated interaction with the environment, then inner speech—thinking in words—arises through simulated interaction with the social environment.[54] Similarly, just as the simulation hypothesis predicts that the same or similar neural mechanisms are involved in both simulated and actual interactions with the environment, so it predicts that the same or similar neural mechanisms are involved in inner speech as in overt speech.[55]

When the simulation hypothesis was introduced at the start of this chapter, it was emphasized that the neurodynamics of simulated interaction is unlikely to be a faithful imitation of the neurodynamics of real interaction, and that carefully framed verbal reports of the inner life will reflect this. Similarly, the neurodynamics of inner speech is likely to differ in a variety of ways from the neurodynamics of actual conversation, and attempts to recapture the accompanying phenomenology will tend towards the literary. One source of such differences is the absence of a real audience or interlocutor. But if, as Vygotsky proposed, soliloquy—the habit of speaking out loud to oneself—precedes the development of inner speech in the child, then this difference is diminished.[56]

[54] Hesslow (1994; 2002).

[55] Results obtained in imaging studies using an inner speech paradigm are consistent with this prediction, and demonstrate elevated activation during inner speech in Broca's and Wernicke's areas, which are associated with overt language production and comprehension (Shergill, *et al*., 2001; Jones & Fernyhough, 2007).

[56] Vygotsky (1934/1986), especially Chapter 4. Vygotsky's translators use the term 'egocentric speech' rather than 'soliloquy'. The preferred coinage in contemporary psychology is 'private speech'. (Private speech, of course, has nothing to do with 'private language' in Wittgenstein's sense.) Vygotsky's ideas remain influential in the field of child development, and the stages he proposed are widely accepted (Krafft & Berk, 1998; Winsler, *et al*., 2000;

Inner speech is then internalized soliloquy, which remains a form of simulated interaction with the environment, albeit one where the environment offers little in the way of response. Other alleged differences between inner speech and actual conversation can then be comfortably accommodated, although there is no need here to specify what those differences might be.

However, a few remarks are in order on the transition from social speech to soliloquy (or 'private speech', as it's called in the field of child development), and on the significant role imputed to pretend play in effecting this transition. Studies on pre-school children have shown that pretend play develops at around the same time as private speech. Moreover, private speech tends to accompany certain kinds of play more abundantly than others, namely fantasy scenarios involving make–believe characters (who often have imagined voices of their own).[57] Indeed, in an important sense, private speech is itself a form of pretend play, wherein the child pretends that an interlocutor (a carer, teacher, or peer) is present where there is none, and supplies the imaginary interlocutor's voice herself, overlaying it on the real situation just as imaginary tea can be envisioned in a real plastic cup and duly drunk.

Child psychologists have accumulated evidence for the claim, again due to Vygotsky, that private speech helps children to *self-regulate* their behaviour.[58] Suppose a young child is playing with a shape-fitting puzzle. She is struggling to make the yellow piece fit when an older friend comes along and says, 'You've got to start with the green piece.' Later, when the younger child returns to the puzzle on her own, she picks up the yellow piece again. But this time she says to herself, 'Not the yellow piece, silly! Where's the green one?' and puts the yellow piece down without attempting to fit it. By recapitulating the older child's verbal instruction, she manages to guide her own behaviour and solve the puzzle more quickly. In a few years, a similar strategy might be beneficial for learning, say, an arithmetic procedure, or rationally to weigh the arguments for and against a position. In all such cases, the voice of the imaginary interlocutor is eventually appropriated and becomes the child's own, offering commentary and guidance for life.[59]

Winsler, *et al.*, 2003; Winsler, 2009). The present discussion is not premised on Vygotsky's views on language in any other respect.

[57] Krafft & Berk (1998); Bergen (2002).

[58] See the review by Winsler (2009).

[59] Wittgenstein (1958) discusses the relationship between thought and language (§327–§342). His strategy, as always, is critical. In affirming that 'when [we] think in language, there aren't "meanings" going through [the] mind in addition to the verbal expressions …' (§329), he is repudiating a particular philosophical view of meaning (as something hidden and private). His discussion has no bearing on the topic of brain processes.

Much of the power of a child's imagination in pretend play derives from the ability to see one thing as something else—a stick as a sword, a box as a house, or a teddy bear as a (talking) person. This ability is accounted for in the present theoretical framework on the assumption that the brain can run an internally closed sensorimotor loop *side by side* with the sensorimotor loop that is closed through interaction with the real world, thus bringing into being a form of augmented reality, the augmentation in this case being the interlocutor's commentary. This ability in turn relies on an integrative facility that enables brain processes that are not ordinarily coupled with each other to form a coalition. When a child tying his shoelaces says to himself, 'The rabbit runs round the tree and pops his head through the hole,' a coalition of brain processes has formed that blends together manual expertise with string-like materials, verbal recitation from episodic memory, and knowledge of the natural world (or a fairy-tale version of it). In short, the recommended combination of an internal sensorimotor loop with a global workspace is a plausible substrate for the functionality required for both pretend play and private speech.

Further reflection on language reinforces the point. Language use in a social setting is no different from any other sort of behaviour insofar as it issues from a succession of coalitions of coupled brain processes. These process coalitions in turn are coupled with the environment, where the environment in this instance includes other language users who can be similarly described. The dynamics of coalition formation and break-up is the same here as described in Chapter 5. The combinatorial productivity of language, and the systematicity of thought it reflects, result from the open-endedness of the repertoire of coalitions of brain processes that can form. As with mental time travel, where the elements of past experience can be combined and recombined in endless ways, so the ways to assemble a sentence and put it to use in everyday life are endless. But with the advent of human language, thanks to its capacity for symbolic compression and abstraction, the potential for opening up new vistas of affordance is dramatically amplified. The integrative capabilities of a global neuronal workspace are once again called upon to explain how all this is possible.

Talk in the absence of an audience, whether externalized as soliloquy or internalized as inner speech, similarly exploits an open-ended repertoire of combinatorial possibilities thanks to the global neuronal workspace. Moreover, just as real sensorimotor interaction with the environment (including conversation) can leave a trace in episodic memory and thereby exercise deferred influence on the system, so can internally simulated interaction, including inner speech as well as visual and tactile imagery. Conversely, the 'subject matter' of speech is not confined to the here and now. Just as other aspects of

behaviour are, in an open-ended way, sensitive to our recollections of the past and our plans for the future, so are the things we say, both to each other and to ourselves. Consequently, insofar as episodic memory enables us to speak about events in the past, it also enables us to speak of our own speech, and to speak inwardly of our own inner speech, drawing on all the powers of compression and abstraction inherent in language as we do so.[60] In this way, our inner life is given the freedom to imprison itself in language, and from here the 'I think' of Descartes and the inauguration of philosophy is just a short walk away.[61]

In the shadow of the theoretical edifice we have erected, for example, it's tempting to speak of the self as a kind of 'confluence of unities'.[62] The past, the present, and the future of an animal, it seems, are brought together within a singular global workspace. Like the hub of a wheel which holds all the spokes in place, the global workspace enables a multiplicity of parallel processes to act as an integrated whole and fulfil a common remit. This common remit is to direct the actions of a spatially localized body, using information from the sensory apparatus fastened to that body, for the benefit of the animal and its progeny.[63] So we seem to find, in our selves, unity over time, unity of function, and unity in space. Yet each of these unities is contingent. In pathology or in thought experiment, the integrity of our memories can be compromised, the connectivity of our brains can be split or fragmented, and our bodies can be broken apart and rebuilt. What then am I? In the end, all we have is simply what we find, and what we can usefully say to each other about what we find is all that needs to be said. And perhaps, in the end, it's best just to sit quietly and let go of that thought too.

[60] This, in the present view, is what it means to have a *higher-order* or *reflexive* thought, that is to say a thought about a thought (Rosenthal, 1986).

[61] See Lakoff & Johnson (1999), especially Chapter 19.

[62] These remarks are evocative of Kant's discussion of the unity of consciousness in the *Critique of Pure Reason*. (Kant, 1781/1929, pp.129–140) (See the commentary by Strawson (1966), pp.72–117.) But the present project is quite different from Kant's. We are not seeking *a priori* grounds for the possibility of experience.

[63] For a discussion of this point, see Shanahan (2005).

This page intentionally left blank

Bibliography

Abarbanel, H.D.I., Brown, R., & Kennel, M.B. (1991). Variation of Lyapunov exponents on a strange attractor. *Journal of Nonlinear Science* **1**, 175–199.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., & Bullmore, E. (2006). A resilient, low-frequency, small-world, human brain functional network with highly-connected association cortical hubs. *Journal of Neuroscience* **26**(1), 63–72.

Aitken, R. (1991). *The Gateless Barrier: The Wu-Men Kuan (Mumonkan)*. North Point Press.

Aleksander, I. (2005). *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in Humans, Animals and Machines*. Imprint Academic.

Amit, D.J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press.

Andersen, H.K. & Grush, R. (2009). A brief history of time-consciousness: historical precursors to James and Husserl. *Journal of the History of Philosophy* **47**(2), 277–307.

Aristotle (350 B.C./1924). *Aristotle's Metaphysics*. Trans. W.D. Ross. Clarendon Press.

Aristotle (350 B.C./1941). De Interpretatione (On Interpretation). Trans. E.M. Edghill. In R. McKeon (ed.), *The Basic Works of Aristotle*, Random House, pp. 38–61.

Augustine. (398/1961). *Confessions*. Trans. R.S. Pine-Coffin. Penguin.

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

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

Baars, B.J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences* **6**(1), 47–52.

Baars, B.J. & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences* **7**(4), 166–172.

Baddeley, A.D. & Hitch, G.J. (1974). Working memory. In G.A. Bower (ed.), *Recent Advances in Learning and Motivation* (Vol. 8), Academic Press, pp. 47–90.

Baddeley, A.D. (2007). *Working Memory, Thought, and Action*. Oxford University Press.

Baillargeon, R. (1993). The object concept revisited: new directions in the investigation of infants' physical knowledge. In C.E. Granrun (ed.), *Visual Perception and Cognition in Infancy*, Lawrence Erlbaum Associates, pp. 265–315.

Bak, P. (1997). *How Nature Works*. Oxford University Press.

Balduzzi, D. & Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology* **4**(6), e1000091.

Bargh, J.A. & Chartrand, T.L. (1999). The unbearable automaticity of being. *American Psychologist* **54**(7), 462–479.

Bargh, J.A. & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science* **3**(1), 73–79.

Barrie, J.M., Freeman, W.J., & Lenhart, M. (1996). Modulation by discriminative training of spatial patterns of gamma EEG amplitude and phase in neocortex of rabbits. *Journal of Neurophysiology* **76**, 520–539.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences* **22**, 577–609.

Barsalou, L.W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B* **364**, 1281–1289.

Bartolomeo, P. (2008). The neural correlates of visual imagery: an ongoing debate. *Cortex* **44**, 107–108.

Bassett, D.S. & Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist* **12**(6), 512–523.

Bateson, G. (1972/2000). *Steps to an Ecology of Mind*. University of Chicago Press.

Bayne, T. & Chalmers, D. (2003). What is unity of consciousness? In A. Cleermans (ed.), *The Unity of Consciousness: Binding, Integration, and Dissociation*, Oxford University Press, pp. 23–58.

Bennett, M.R. & Hacker, P.M.S. (2003). *Philosophical Foundations of Neuroscience*. Blackwell.

Berg, E.A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology* **39**, 15–22.

Bergen, D. (2002). The role of pretend play in children's cognitive development. *Early Childhood Research and Practice* **4**(1).

Bermúdez, J.L. (2003). *Thinking Without Words*. Oxford University Press.

Bird, C.D. & Emery, N.J. (2009). Insightful problem solving and creative tool modification by captive non-tool-using rooks. *Proceedings of the National Academy of Sciences* **106**, 10370–10375.

Blackmore, S. (2009). *Ten Zen Questions*. Oneworld Publications.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**(2), 227–287.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* **30**, 481–548.

Borst, G. & Kosslyn, S.M. (2008). Visual mental imagery and visual perception: structural equivalence revealed by scanning processes. *Memory and Cognition* **36**(4), 849–862.

Braitenberg, V. & Schüz, A. (1998). *Anatomy of the Cortex: Statistics and Geometry of Neural Connectivity*. Springer.

Braithwaite, V. (2010). *Do Fish Feel Pain*. Oxford University Press.

Breitmeyer, B. & Öğmen, H. (2006). *Visual Masking: Time Slices Through Conscious and Unconscious Vision*. Oxford University Press.

Bressler, S.L. & Kelso, J.A.S. (2001). Cortical coordination dynamics and cognition. *Trends in Cognitive Sciences* **5**(1), 26–36.

Bringsjord, S. (2007). Offer: one billion dollars for a conscious robot; if you're honest, you must decline. *Journal of Consciousness Studies* **14**(7), 28–43.

Brooks, R.A. (1991). Intelligence without reason. In *Proceedings 1991 International Joint Conference on Artificial Intelligence (IJCAI 91)*, pp. 569–595.

Buck, J.B. (1938). Synchronous rhythmic flashing of fireflies. *Quarterly Review of Biology* **13**(3), 301–314.

Buckner, R.L. & Carroll, D.C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences* **11**(2), 49–57.

Bullmore, E. & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10**, 186–198.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Canfield, J. (1975). Wittgenstein and Zen. *Philosophy* **50**, 383–408.

Caporale, N. & Yang, D. (2008). Spike-timing dependent plasticity: a Hebbian learning rule. *Annual Review of Neuroscience* **31**, 25–46.

Caputo, J. (1983). The thought of being and the conversation of mankind: the case of Heidegger and Rorty. *Review of Metaphysics* **36**, 661–685.

Carman, T. (2007). Dennett on seeming. *Phenomenology and the Cognitive Sciences* **6**, 99–106.

Carnell, A. & Richardson, D. (2007). Parallel computation in spiking neural nets. *Theoretical Computer Science* **386**, 57–72.

Carruthers, P. (2003). On Fodor's problem. *Mind and Language* **18**(5), 502–523.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford University Press.

Cavanna, A.E. & Trimble, M.R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* **129**, 564–583.

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

Chen, Z.J., He, Y., Rosa-Neto, P., Germann, J., & Evans, A.C. (2008). Revealing modular architecture of human brain structural networks by using cortical thickness from MRI. *Cerebral Cortex* **18**, 2374–2381.

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America* **25**, 975–979.

Chialvo, D.R. (2004). Critical brain networks. *Physica A* **340**, 756–765.

Chialvo, D.R. (2008). Emergent complexity: what uphill analysis or downhill invention cannot do. *New Ideas in Psychology* **26**, 158–173.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Churchland, P.M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**, 67–90.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B* **362**, 1585–1599.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Clayton, N.S. & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature* **395**, 272–274.

Cole, M.W., Pathak, S., & Schneider, W. (2010). Identifying the brain's most globally connected regions. *NeuroImage* **49**, 3132–3148.

Cole, M.W. & Schneider, W. (2007). The cognitive control network: integrated cortical regions with dissociable functions. *NeuroImage* **37**, 343–360.

Coolidge, F.L. & Wynn, T. (2009). *The Rise of* Homo sapiens*: The Evolution of Modern Thinking*. Wiley-Blackwell.

Costall, A. (2006). 'Introspectionism' and the mythical origins of scientific psychology. *Consciousness and Cognition* **15**, 634–654.

Cotterill, R. (1998). *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.

Craik, K.J.W. (1943). *The Nature of Explanation*. Cambridge University Press.

Curtis, C.E. & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences* **17**(9), 415–423.

Damasio, A. R. (1995). *Descartes' Error: Emotion, Reason, and the Human Brain*. Picador.

Damoiseaux, J.S. & Greicius, M.D. (2009). Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Structure and Function* **213**, 525–533.

Davidson, D. (1982). Rational animals. *Dialectica* **36**(4), 317–327.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Decety, J. & Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Research* **1079**, 4–14.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press.

Dehaene, S. & Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *Public Library of Science Biology* **3**(5), e141.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences* **10**(5), 204–211.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Science* **95**, 14529–14534.

Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science* **100**(14), 8520–8525.

Dennett, D. (1984). Cognitive wheels: the frame problem in artificial intelligence. In C. Hookway (ed.), *Minds, Machines and Evolution*, Cambridge University Press, pp. 129–151.

Dennett, D. (1991). *Consciousness Explained*. Penguin.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition* **79**, 221–237.

Derrida, J. (1967/1973). *Speech and Phenomena and Other Essays on Husserl's Theory of Signs*. Trans. D.B. Allison. Northwestern University Press.

Derrida, J. (1982). *Margins of Philosophy*. Trans. A.Bass. Harvester.

Descartes, R. (1637/1968). Discourse on method. In *Discourse on Method and The Meditations*, trans. F.E. Sutcliffe, Penguin, pp. 25–91.

Descartes, R. (1641/1968). Meditations. In *Discourse on Method and The Meditations*, trans. F.E. Sutcliffe, Penguin, pp. 93–169.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B* **362**, 761–772.

Dias, J.R., Oliveira, R.F., & Kinouchi, O. (2008). Chaotic itinerancy, temporal segmentation and spatio-temporal combinatorial codes. *Physica D* **237**, 1–5.

Doesburg, S.M., Green, J.J., McDonald, J.J., & Ward, L.M. (2009). Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual attention. *PLoS ONE* **4**(7), e6142.

Doesburg, S.M., Roggeveen, A.B., Kitajo, K., & Ward, L.M. (2008). Large-scale gamma-band phase synchronization and selective attention. *Cerebral Cortex* **18**(2), 386–396.

Dreyfus, H.L. (1991). *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. MIT Press.

Dreyfus, H.L. (2008). Why Heideggerian A failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland & M. Wheeler (eds.), *The Mechanical Mind in History*, MIT Press, pp. 331–371.

Dupret, D., Pleydell-Bouvarie, B., & Csicsvari, J. (2008). Inhibitory interneurons and network oscillations. *Proceedings of the National Academy of Science* **105**(47), 18079–18080.

Edelman, D.B., Baars, B.J., & Seth, A.K. (2005). Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition* **14**, 169–187.

Edelman, D.B. & Seth, A.K. (2009). Animal consciousness: a synthetic approach. *Trends in Neurosciences* **32**(9), 476–484.

Edelman, G.M. & Tononi, G. (2000). *A Universe of Consciousness: How Matter Becomes Imagination*. Basic Books.

Eguíluz, V.M., Chialvo, D.R., Cecchi, G.A., Baliki, M., & Apkarian, A.V. (2005). Scale-free brain functional networks. *Physical Review Letters* **94**, 018102.

Eilan, N. (2002). The reality of consciousness. In D. Charles & W. Child (eds.), *Wittgensteinian Themes: Essays in Honour of David Pears*, Oxford University Press, pp. 163–194.

Emery, N.J. & Clayton, N.S. (2005). Evolution of the avian brain and intelligence. *Current Biology* **15**(23), R946–R950.

Eredyi, M.H. (2004). Subliminal perception and its cognates: theory, indeterminacy, and time. *Consciousness and Cognition* **13**, 73–91.

Ericsson, K.A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during expert performance on representative tasks. In K.A. Ericsson, N. Charness, R.R. Hoffman, & P.J. Feltovich (eds.), *The Cambridge Handbook of Expertise and Performance*, Cambridge University Press, pp. 223–242.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data*. MIT Press.

Fann, K.T. (1969). *Wittgenstein's Conception of Philosophy*. Basil Blackwell.

Fauconnier, G. & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Felleman, D.J. & Van Essen, D.C. (1991). Distributed hierarchical processing in the primate visual cortex. *Cerebral Cortex* **1**, 1–47.

Ferkin, M.H., Combs, A., delBarco-Trillo, J., Pierce, A.A., & Franklin, S. (2008). Meadow voles, *Microtus Pennsylvanicus*, have the capacity to recall the 'what', 'where', and 'when' of a single past event. *Animal Cognition* **11**, 147–159.

Ferrarini, L., Veer, I.M., Baerends, E., van Tol, M.-J., Renken, R.J., van der Wee, D.J., Veltman, N.J.A., Aleman, A., Zitman, F.G., Penninx, B.W.J.H., van Buchem, M.A., Reiber, J.H.C., Rombouts, S.A.R.B., & Milles, J. (2009). Hierarchical functional modularity in the resting state human brain. *Human Brain Mapping* **30**, 2220–2231.

Fields, R.D. (2008). White matter in learning, cognition, and psychiatric disorders. *Trends in Neurosciences* **31**(7), 361–370.

Fodor, J.A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fodor, J.A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., & Raichle, M.E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Science* **102**(27), 9673–9678.

Fraiman, D., Balenzuela, P., Foss, J., & Chialvo, D.R. (2009). Ising-like dynamics in large-scale functional brain networks. *Physical Review E* **79**, 061922.

Franklin, S. & Graesser, A. (1999). A software agent model of consciousness. *Consciousness and Cognition* **8**, 285–301.

Freeman, W.J. (1999). *How Brains Make Up Their Minds*. Weidenfeld & Nicolson.

Freeman, W.J. (2003). Evidence from human scalp electroencephalograms of global chaotic itinerancy. *Chaos* **13**(3), 1067–1077.

Freeman, W.J. (2004a). Origin, structure, and role of background EEG activity. Part 1. Analytic amplitude. *Clinical Neurophysiology* **115**(9), 2077–2088.

Freeman, W.J. (2004b). Origin, structure, and role of background EEG activity. Part 2. Analytic phase. *Clinical Neurophysiology* **115**(9), 2089–2107.

Freeman, W.J. (2006). A cinematographic hypothesis of cortical dynamics in perception. *International Journal of Psychophysiology* **60**, 149–161.

Freeman, W.J. & Rogers, L.J. (2003). A neurobiological theory of meaning in perception Part V: multicortical patterns of phase modulation in gamma EEG. *International Journal of Bifurcation and Chaos* **13**(10), 2867–2887.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences* **9**(10), 474–480.

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical communication. *Annual Review of Neuroscience* **32**, 209–224.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks* **16**, 1325–1352.

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., & Naccache, L. (2009). Converging intracranial markers of conscious access. *PLoS Biology* **7**(3), e1000061.

Galesse, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B* **362**, 659–669.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition* **17**(3), 887–910.

Garfield, J.L. & Priest, G. (2003). Nāgārjuna and the limits of thought. *Philosophy East and West* **53**(1), 1–21.

Geldard, F.A. & Sherrick, C.E. (1972). The cutaneous 'rabbit': a perceptual illusion. *Science* **13**(178), 178–179.

Ganis, G., Thompson, W.L., & Kosslyn, S.M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research* **20**, 226–241.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science* **99**, 7821–7826.

Glazebrook, J.F. & Wallace, R. (2009). Small worlds and red queens in the global workspace: an information-theoretic approach. *Cognitive Systems Research* **10**, 333–365.

Gong, G., He, Y., Concha, L., Lebel, C., Gross, D.W., Evans, A.C., & Beaulieu, C. (2009). Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography. *Cerebral Cortex* **19**, 524–536.

Goodale, M.A. & Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences* **15**(1), 20–25.

Gordon, R.M. (1986). Folk psychology as simulation. *Mind and Language* **1**, 158–171.

Gray, C.M., König, P., Engel, A.K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337.

Greicius, M.D., Krasnow, B., Reiss, A.L., & Menon, V. (2003). Functional connectivity in the resting brain: a networks analysis of the default mode hypothesis. *Proceedings of the National Academy of Science* **100**(1), 253–258.

Grèzes, J. & Decety, J. (2001). Does perception of object afford action? Evidence from a neuroimaging study. *Neuropsychologia* **40**, 212–222.

Griffin, D.R. (2001). *Animal Minds: Beyond Cognition to Consciousness*. University of Chicago Press.

Gros, C. (2009). Cognitive computation with autonomously active neural networks: an emerging field. *Cognitive Computation* **1**, 77–90.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* **27**, 344–442.

Güntürkün, O. (2005). The avian 'prefrontal cortex' and cognition. *Current Opinion in Neurobiology* **15**, 686–693.

Güzeldere, G. (1997). The many faces of consciousness: field guide. In N. Block, O. J. Flanagan & G. Güzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*, MIT Press, pp. 1–67.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, C.J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology* **6**(7), e159.

Haikonen, P.O. (2003). *The Cognitive Approach to Conscious Machines*. Imprint Academic.

Haken, H. (2006). Synergetics of brain function. *International Journal of Psychophysiology* **60**, 110–124.

Hansell, M. (2000). *Bird Nests and Construction Behaviour*. Cambridge University Press.

Hansell, M. (2007). *Built by Animals: The Natural History of Animal Architecture*. Oxford University Press.

Hare, B., Call, J. & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour* **61**, 139–151.

Hasentraub, A., Shu, Y., Halder, B., Kraushaar, U., & Duque, A. (2005). Inhibitory postsynaptic potentials carry synchronized frequency information in active cortical networks. *Neuron* **47**, 423–435.

Hassabis, D. & Maguire, E.A. (2008). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences* **11**(7), 299–306.

He, Y., Chen, Z.J., & Evans, A.C. (2007). Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cerebral Cortex* **17**, 2407–2419.

Hedwig, B. (2006). Pulses, patterns, and paths: neurobiology of acoustic behaviour in crickets. *Journal of Comparative Physiology A* **192**, 677–689.

Heidegger, M. (1926/1962). *Being and Time*. Trans. J. Macquarrie & E. Robinson. Blackwell.

Hellwig, B. (2000). A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological Cybernetics* **82**, 111–121.

Hesslow, G. (1994). Will neuroscience explain consciousness? *Journal of Theoretical Biology* **171**, 29–39.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences* **6**(6), 242–247.

Hidalgo, C.A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science* **317**, 482–487.

Hoare, C.A.R. (1985). *Communicating Sequential Processes*. Prentice Hall.

Holender, D. & Duscherer, K. (2004). Unconscious perception: the need for a paradigm shift. *Perception and Psychophysics* **66**(5), 872–881.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press.

Holland, O. (ed.) (2003). *Machine Consciousness: Special Issue of the Journal of Consciousness Studies*. Imprint Academic.

Houk, J.C., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P.J., Roy, S.A., & Simo, L.S. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Philosophical Transactions of the Royal Society B* **362**, 1573–1583.

Hughes, R.N. & Blight, C.M. (1999). Algorithmic behaviour and spatial memory are used by two intertidal fish species to solve the radial maze. *Animal Behaviour* **58**, 601–613.

Humphries, M.D., Gurney, K., & Prescott, T.J. (2007). Is there a brainstem substrate for action selection? *Philosophical Transactions of the Royal Society B* **362**, 1627–1639.

Hurlburt, R.T. & Akhter, S.A. (2006). The descriptive experience sampling method. *Phenomenology and the Cognitive Sciences* **5**, 271–301.

Hurley, S. (2006). Making sense of animals. In S. Hurley & M. Nudds (eds.), *Rational Animals?* Oxford University Press, pp. 139–171.

Husserl, E. (1911/1991). On the phenomenology of the consciousness of internal time (1893–1917). Trans. J.B. Brough. Kluwer.

Ijspeert, A.J. (2008). Central pattern generators for locomotion and control in animals and robots: a review. *Neural Networks* **21**, 642–653.

Ikegami, T. (2007). Simulating active perception and mental imagery with embodied chaotic itinerancy. *Journal of Consciousness Studies* **14**(7), 111–125.

Iturria-Medina, Y., Sotero, R.C., Canales-Rodríguez, E.J., Alemán-Gómez, Y., & Melie-García, L. (2008). Studying the human brain anatomical network via diffusion-weighted MRI and graph theory. *NeuroImage* **40**, 1064–1076.

Izhikevich, E.M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly* **32**, 127–136.

James, W. (1890/1950). *The Principles of Psychology*. Dover.

James, W. (1902/1985). *The Varieties of Religious Experience*. Penguin.

Jonas, H. (1966). *The Phenomenon of Life*. Harper and Row.

Jones, S.R. & Fernyhough, C. (2007). Neural correlates of inner speech and auditory verbal hallucinations: a critical study and theoretical integration. *Clinical Psychology Review* **27**, 140–154.

Kaneko, K. & Tsuda, I. (2003). Chaotic itinerancy. *Chaos* **13**(3), 926–936.

Kant, I. (1781/1929). *Critique of Pure Reason*. Trans. N. Kemp Smith. Macmillan.

Kasulis, T.P. (1981). *Zen Action Zen Person*. University of Hawaii Press.

Kelso, J.A.S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.

Kim, C.-Y. & Blake, R. (2005). Psychophysical magic: rendering the visible 'invisible'. *Trends in Cognitive Sciences* **9**(8), 381–388.

Kirk, R. (1974). Sentience and behaviour. *Mind* **83**, 43–60.

Kirk, R. (2005). *Zombies and Consciousness*. Oxford University Press.

Kitzbichler, M.G., Smith, M.L., Christensen, S.R., & Bullmore, E. (2009). Broadband criticality of human brain network synchronization. *PLoS Computational Biology* **5**(3), e1000314.

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Roberts & Company.

Kolers, P.A. & von Grünau, M. (1976). Shape and color in apparent motion. *Vision Research* **16**(4), 329–335.

Korn, H. & Faure, P. (2003). Is there chaos in the brain? II: experimental evidence and related models. *Comptes Rendus Biologies* **326**, 787–840.

Kosslyn, S.M. (1981). The medium and the message in mental imagery: a theory. *Psychological Review* **88**(1), 46–66.

Kosslyn, S.M., Ganis, G., & Thompson, W.L. (2003). Mental imagery: against the nihilistic hypothesis. *Trends in Cognitive Sciences* **7**(3), 109–111.

Kozma, R. & Freeman, W.J. (2009). The KIV model of intentional dynamics and decision making. *Neural Networks* **22**, 277–285.

Krafft, K.C. & Berk, L.E. (1998). Private speech in two preschools: significance of open-ended activities and make-believe play for verbal self-regulation. *Early Childhood Research Quarterly* **13**(4), 637–658.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By*. The University of Chicago Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books.

Lakoff, G. & Núñez, R.E. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. Basic Books.

Lambie, J.A. & Marcel, A.J. (2002). Consciousness and the varieties of emotional experience. *Psychological Review* **109**(2), 219–259.

Lamme, V.A.F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences* **7**(1), 12–18.

Lamme, V.A.F. & Roelfsema, P.R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences* **23**, 571–579.

Landman, R., Sprekreijse, H., & Lamme, V.A.F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research* **43**(2), 149–164.

Langton, C. (1989). Artificial life. In C. Langton (ed.), *Artificial Life*, Addison Wesley, pp. 1–47.

Lien, M.-C., Ruthruff, E., & Johnston, J.C. (2006). Attentional limitations in doing two tasks at once. *Current Directions in Psychological Science* **15**(2), 89–93.

Llinás, R., Ribary, U., Contreras, D., & Pedroarena, C. (1998). The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society B* **353**, 1841–1849.

Loy, D.R. (1992). The deconstruction of Buddhism. In H. Coward & T. Foshay (eds.), *Derrida and Negative Theology*, State University of New York Press, pp. 227–253.

Ludlow, P., Nagasawa, Y. & Stoljar, D. (eds.) (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. MIT Press.

Ma, R. & Kaber, D.B. (2005). Situation awareness and workload in driving while using adaptive cruise control and a cell phone. *International Journal of Industrial Ergonomics* **35**, 939–953.

Maass, W. (1996). Lower bounds for the computational power of networks of spiking neurons. *Neural Computation* **8**, 1–40.

MacLeod, C.M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin* **109**(2), 163–203.

McCarthy, J. (1959a). Programs with common sense. In *Proceedings of the Symposium on the Mechanization of Thought Processes*, H.M.S.O. London, pp. 75–91.

McCarthy, J. (1959b). Comment in Selfridge (1959), p. 527.

McCarthy, J. & Hayes, P.J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie & B. Meltzer (eds.), *Machine Intelligence 4*, Edinburgh University Press, pp. 463–502.

McDowell, J. (1994). *Mind and World*. Harvard University Press.

McFarland, D. (2008). *Guilty Robots, Happy Dogs: The Question of Alien Minds*. Oxford University Press.

McFarland, D. & Bösser, T. (1993). *Intelligent Behavior in Animals and Robots*. MIT Press.

Mack, A. & Rock, I. (1998). *Inattentional Blindness*. MIT Press.

Maravita, A. & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences* **8**(2), 79–86.

Margulies, D.S., Vincent, J.L., Kelly, C., Lohmann, G., Uddin, L.Q., Biswal, B.B., Villringer, A., Castellanos, F.X., Milham, M.P., & Petrides, M. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences* **106**(47), 20069–20074.

Martin-Ordas, G., Call, J., & Colmenares, F. (2008). Tubes, tables, and traps: great apes solve functionally equivalent tasks but show no evidence of transfer across tasks. *Animal Cognition* **11**, 423–430.

Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T., & Macrae, C.N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science* **315**, 393–395.

Maturana, H.R. & Varela, F.J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Kluwer.

Meister, I.G., Krings, T., Foltys, H., Boroojerdi, B., Müller, M., Töpper, R., & Thron, A. (2004). Playing piano in the mind – an fMRI study on music imagery and performance in pianists. *Cognitive Brain Research* **19**, 219–228.

Merikle, P.M., Smilek, D., & Eastwood, J.D. (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition* **79**, 115–134.

Merleau-Ponty, M. (1945/1962). *Phenomenology of Perception*. Routledge & Kegan Paul.

Metzinger, T. (2003). *Being No-One: The Self-Model Theory of Subjectivity*. MIT Press.

Milner, A.D. & Goodale, M.A. (2006). *The Visual Brain in Action* (2nd Edition). Oxford University Press.

Milner, A.D. & Goodale, M.A. (2008). Two visual streams re-viewed. *Neuropsychologia* **46**, 774–785.

Milner, R. (1980). *A Calculus of Communicating Systems*. Springer (Lecture Notes in Computer Science No.92).

Mithen, S. (1996). *The Prehistory of the Mind*. Thames & Hudson.

Monaghan, J. (2001). Young peoples' ideas of infinity. *Educational Studies in Mathematics* **48**(2–3), 239–257.

Moore, A.W. (2001). Arguing with Derrida. In S. Glendinning (ed.), *Arguing with Derrida*, Blackwell, pp. 57–88.

Moors, A. & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin* **132**(2), 297–326.

Müller-Linow, M., Hilgetag, C.C., & Hütt, M.-T. (2008). Organization of excitable dynamics in hierarchical biological networks. *PLoS Computational Biology* **4**(9), e1000190.

Nāgārjuna. (2nd century/1995). Mūlamadhyamakakārikā. In J.L. Garfield (ed./trans.), *The Fundamental Wisdom of the Middle Way: Nāgārjuna's Mūlamadhyamakakārikā,* Oxford University Press, pp. 1–83.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review* **83**(4), 435–450.

Nara, S. & Davis, P. (1992). Chaotic wandering and search in a cycle-memory neural network. *Progress of Theoretical Physics* **88**(5), 845–855.

Newman, M.E.J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**(23), 8577–8582.

Nichols, S. & Stich, S. (2000). A cognitive theory of pretense. *Cognition* **74**, 115–147.

Nii, H.P. (1986). The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine* **7**(2), 38–53.

Nilsson, N.J. (ed.) (1984). *Shakey the Robot*. SRI Technical Note no. 323. SRI, Menlo Park, California.

Olsson, I.A. & Keeling, L.J. (2005). Why in earth? Dustbathing behaviour in jungle and domestic fowl reviewed from a Tinbergian and animal welfare perspective. *Applied Animal Behaviour Science* **93**(3–4), 259–282.

O'Regan, J.K. & Nöe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* **24**, 939–1031.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance* **10**(3), 358–377.

Peirce, C.S. (1878). How to make our ideas clear. *Popular Science Monthly* **12**, 286–302.

Penn, D.C. & Povinelli, D.J. (2007). Causal cognition in human and nonhuman animals: a comparative, critical review. *Annual Review of Psychology* **58**, 97–118.

Penrose, R. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Phillips, D.Z. (1993). *Wittgenstein and Religion*. Macmillan.

Piaget, J. (1954). *The Construction of Reality in the Child*. Routledge and Kegan Paul.

Pikovsky, A., Rosenblum, M. & Kurths, J. (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press.

Povinelli, D.J. (2000). *Folk Physics for Apes*. Oxford University Press.

Pylyshyn, Z.W. (1981). The imagery debate: analog media versus tacit knowledge. *Psychological Review* **88**(1), 16–45.

Pylyshyn, Z.W. (2003). Return of the mental image: are there really pictures in the brain? *Trends in Cognitive Sciences* **7**(3), 113–118.

Rabinovich, M.I. & Abarbanel, D.I. (1998). The role of chaos in neural systems. *Neuroscience* **87**(1), 5–14.

Raby, C.R., Alexis, D.M., Dickinson, A., & Clayton, N.S. (2007). Planning for the future by western scrub-jays. *Nature* **445**, 919–921.

Raichle, M.E. (2010). Two views of brain function. *Trends in Cognitive Sciences* **14**(4), 180–190.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., & Shulman, G.L. (2001). A default mode of the brain. *Proceedings of the National Academy of Science* **98**(2), 676–682.

Rubinov, M., Sporns, O., van Leeuwen, C., & Breakspear, M. (2009). Symbiotic relationship between brain structure and dynamics. *BMC Neuroscience* 10:55.

Quine, W.V.O. (1960). *Word and Object*. MIT Press.

Quine, W.V.O. (1969). *Ontological Relativity and Other Essays*. Columbia University Press.

Redgrave, P., Prescott, T.J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem. *Neuroscience* **89**(4), 1009–1023.

Reeve, R. & Webb, B. (2002). New neural circuits for robot phonotaxis. *Proceedings of the Royal Society A* **361**, 2245–2266.

Reigl, M., Alon, U., & Chklovskii, D.B. (2004). Search for computational modules in the *C. Elegans* brain. *BMC Biology* **2**, 25–37.

Reingold, E.M. & Merikle, P.M. (1988). Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* **44**(6), 563–575.

Reingold, E.M. (2004). Unconscious perception and the classic dissociation paradigm: a new angle. *Perception and Psychophysics* **66**(5), 882–887.

Rodriguez, E., George, N., Lachaux, J.-P., Martinerie, J., Renault, B., & Varela, F. (1999). Perception's shadow: long-distance synchronization of human brain activity. *Nature* **397**, 430–433.

Rolls, E.T. (2005). *Emotion Explained*. Oxford University Press.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton University Press.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies* **49**, 329–359.

Ruthruff, E., Pashler, H.E., & Klaasen, A. (2001). Processing bottlenecks in dual-task performance: structural limitation or strategic postponement. *Psychonomic Bulletin and Review* **8**(1), 73–80.

Sackur, J. & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. *Cognition* **111**, 187–211.

Santos, L.R, Pearson, H.M., Spaepen, G.M., Tsao, F., & Hauser, M.D. (2006). Probing the limits of tool competence: experiments with two non-tool-using species (*Cercopithecus aethiops* and *Saguinus oedipus*). *Animal Cognition* **9**, 94–109.

Sauer, T. (2003). Chaotic itinerancy based on attractors of one-dimensional maps. *Chaos* **13**(3), 947–952.

Scannell, J.W. & Young, M.P. (1993). The connectional organization of neural systems in the cat cerebral cortex. *Current Biology* **3**, 191–200.

Schacter, D.L. & Addis, D.R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B* **362**, 773–786.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2008). Episodic simulation of future events. *Annals of the New York Academy of Sciences* **1124**, 39–60.

Schmahmann, J.D. & Pandya, D.N. (2006). *Fiber Pathways of the Brain*. Oxford University Press.

Schneider, W. (2009). Automaticity and consciousness. In W. Banks (ed.), *The Elsevier Encyclopedia of Consciousness*, pp. 83–92.

Schneider, W. & Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* **84**(1), 1–66.

Searle, J.R. (1992). *The Rediscovery of the Mind*. MIT Press.

Seed, A.M., Tebbich, S., Emery, N.J., & Clayton, N.S. (2006). Investigating physical cognition in rooks, *Corvus frugilegus*. *Current Biology* **16,** 697–701.

Selfridge, O.G. (1959). Pandemonium: a paradigm for learning. In *Proceedings of the Symposium on the Mechanization of Thought Processes*, H.M.S.O. London, pp. 511–529.

Seth, A.K. & Baars, B.J. (2005). Neural Darwinism and consciousness. *Consciousness and Cognition* **14**, 140–168.

Seth, A.K., Dienes, Z. Cleermans, A., Overgaard, M., & Pessoa, M. (2008). Measuring consciousness: relating behavioral and neurophysiological approaches. *Trends in Cognitive Sciences* **12**(8), 314–321.

Shanahan, M.P. (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press.

Shanahan, M.P. (2005). Global access, embodiment, and the conscious subject. *Journal of Consciousness Studies* **12**(12), 46–66.

Shanahan, M. P. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition* **15**, 433–449.

Shanahan, M.P. (2008a). A spiking neuron model of cortical broadcast and competition. *Consciousness and Cognition* **17**, 288–303.

Shanahan, M.P. (2008b). Dynamical complexity in small-world networks of spiking neurons. *Physical Review E* **78**, 041924.

Shanahan, M.P. (2010). Metastable chimera states in community-structured oscillator networks. *Chaos* **20**, 013108.

Shanahan, M.P. & Baars, B.J. (2005). Applying global workspace theory to the frame problem. *Cognition* **98**(2), 157–176.

Shannon, C.E. & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Shefi, O., Golding, I., Segev, R., Ben-Jacob, E., & Ayali, A. (2002). Morphological characterization of *in vitro* neuronal networks. *Physical Review E* **66**, 021905.

Shepard, R.N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science* **171**(3972), 701–703.

Shergill, S.S., Bullmore, E.T., Brammer, M.J., Williams, S.C.R., Murray, R.M., & McGuire, P.K. (2001). A functional study of auditory verbal imagery. *Psychological Medicine* **31**, 241–253.

Sherman, S.M. & Guillery, R. (2005). *Exploring the Thalamus and its Role in Cortical Function*. Second edition. MIT Press.

Siegelmann, H.T. (2003). Neural and super-Turing computing. *Minds and Machines* **13**, 103–114.

Siegelmann, H.T. & Sontag, E.D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences* **50**, 132–150.

Simons, D.J. & Chabris, C.F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* **28**, 1059–1074.

Singer, W. & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**, 555–586.

Skarda, C.A. & Freeman, W.J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* **10**, 161–195.

Sloman, A. (1984). The structure of the space of possible minds. In S. Torrance (ed.), *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*, Ellis Horwood, pp. 35–42.

Sloman, A. & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies* **10**(4–5), 133–172.

Smallwood, J. & Schooler, J. (2007). The restless mind. *Psychological Bulletin* **132**(6), 946–958.

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception: a model-based approach to method and evidence. *Perception and Psychophysics* **66**(5), 846–867.

Song, S., Miller, K.D., & Abbott, L.F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience* **3**(9), 919–926.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied* **74**(11), 1–29.

Sporns, O. (2010). *Networks of the Brain*. MIT Press.

Sporns, O., Tononi, G., & Edelman, G.M. (2000). Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex* **10**, 127–141.

Sporns, O. & Zwi, J.D. (2004). The small world of the cerebral cortex. *Neuroinformatics* **2**, 145–162.

Spreng, R.N., Mar, R.A., & Kim, A.S.N. (2008). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience* **21**(3), 489–510.

Strayer, D.L. & Drews, F.A. (2007). Cell-phone-induced driver distraction. *Current Directions in Psychological Science* **16**(3), 128–131.

Stone, A.A. & Shiffman, S. (2002). Capturing momentary, self-report data: a proposal for reporting guidelines. *Annals of Behavioral Medicine* **24**(3), 236–243.

Stout, D. & Chaminade, T. (2007). The evolutionary neuroscience of tool making. *Neuropsychologia* **45**, 1091–1100.

Strawson, P.F. (1966). *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. Methuen & Co.

Striedter, G.F. (2005). *Principles of Brain Evolution*. Sinauer Associates.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* **18**, 643–662.

Suddendorf, T. & Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs* **123**(2), 133–167.

Suddendorf, T. & Corballis, M.C. (2007). The evolution of foresight: what is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* **30**, 299–351.

Suddendorf, T., Addis, D.R., & Corballis, M.C. (2009). Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society B* **364**, 1317–1324.

Taga, G, Yamaguchi, Y., & Shimizu, H. (1991). Self-organised control of bipedal locomotion by neural oscillators in unpredictable environments. *Biological Cybernetics* **65**, 147–159.

Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Transactions on Systems, Man, and Cybernetics B* **26**(3), 421–436.

Taylor, A.H., Hunt, G.R., Medina, F.S., & Gray, R.D. (2009). Do new caledonian crows solve physical problems through causal reasoning? *Proceedings of the Royal Society B* **276**, 247–254.

Tebbich, S., Seed, A.M., Emery, N.J., & Clayton, N.S. (2007). Non-tool-using rooks, *Corvus frugilegus*, solve the trap-tube problem. *Animal Cognition* **10**, 225–231.

Thomas, N.J.T. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science* **23**(2), 207–245.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biological Bulletin* **215**, 216–242.

Tononi, G., Edelman, G.M., & Sporns, O. (1998). Complexity and coherency: integrating information in the brain. *Trends in Cognitive Sciences* **2**(12), 474–484.

Torres, J.J., Marro, J., Cortes, J.M., & Wemmenhove, B. (2008). Instabilities in attractor networks with fast synaptic fluctuations and partial updating of the neurons activity. *Neural Networks* **21**, 1272–1277.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology* **12**, 97–136.

Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences* **24**, 793–847.

Tsuda, I. (2009). Hypotheses on the functional roles of chaotic transitory dynamics. *Chaos* **19** (1), 015113.

Tsuda, I. & Umemura, T. (2003). Chaotic itinerancy generated by coupling of milnor attractors. *Chaos* **13** (3), 937–946.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (eds.), *Organization of Memory*, Academic Press, pp. 381–403.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford University Press.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology* **53**, 1–25.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind* **59**, 433–460.

Turner, M. (1996). *The Literary Mind: The Origins of Thought and Language*. Oxford University Press.

Ungerleider, L.G. & Mishkin, M. (1982). Two cortical visual systems. In D.J. Ingle, M.A. Goodale & R.J.W. Mansfield (eds.), *Analysis of Visual Behavior*, MIT Press, pp. 549–586.

Uttal, W.R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. MIT Press.

Van den Heuvel, M.P., Mandl, R.C.W, Kahn, R.S., & Pol, H.E.H. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Human Brain Mapping* **30**, 3127–3141.

Vanhaudenhuyse, A., Noirhomme, Q., Tshibanda, L. J.-F., Bruno, M.-A., Boveroux, P., Schnakers, C., Soddu, A., Perlbarg, V., Ledoux, D., Brichant, J.-F., Moonen, G., Maquet, P., Greicius, M.D., Laureys, S. & Boly, M. (2010). Default network connectivity reflects the level of consciousness in non-communicative brain-damaged patients. *Brain* **133**, 161–171.

Varela, F., Lachaux, J.P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: phase synchronization and large scale integration. *Nature Reviews Neuroscience* **2**, 229–239.

Von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology* **5**, 520–526.

Von Uexküll, J. (1957). A stroll through the worlds of animals and men. In C.H. Schiller (ed.) *Instinctive Behaviour: The Development of a Modern Concept*. International Universities Press, pp. 5–80.

Vygotsky, L. (1934/1986). *Thought and Language*. Trans. A. Kozulin. MIT Press.

Wakana, S., Jiang, H., Nagae-Poetscher, L., van Zijl, P.C.M., & Mori, S. (2004). Fiber tract-based atlas of human white matter anatomy. *Radiology* **230**, 77–87.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences* **24** (8), 455–463.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamic of 'small-world' networks. *Nature* **393**, 440–442.

Werner, G. (2007). Metastability, criticality, and phase transitions in brain and its models. *BioSystems* **90**, 496–508.

Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press.

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies* **16**(3), 323–349.

Williams, B. (1978). *Descartes: The Project of Pure Enquiry*. Penguin.

Winsler, A. (2009). Still talking to ourselves after all these years: a review of current research on private speech. In A. Winsler, C. Fernyhough, & I. Montero (eds.), *Private Speech, Executive Functioning, and the Development of Verbal Self-Regulation*, Cambridge University Press, pp. 3–41.

Winsler, A., Carlton, M.P., & Barry, M.J. (2000). Age-related changes in preschool children's systematic use of private speech in a natural setting. *Journal of Child Language* **27**, 665–687.

Winsler, A., De León, J.R., Wallace, B.A., Carlton, M.P., & Willson-Quayle, A. (2003). Private speech in preschool children: developmental stability and change, across-task consistency, and relations with classroom behaviour. *Journal of Child Language* **30**, 583–608.

Wittgenstein, L. (1921/1961). *Tractatus Logico-Philosophicus*. Trans. D.F. Pears & B.F. McGuinness, Routledge & Kegan Paul.

Wittgenstein, L. (1958). *Philosophical Investigations*. Trans. G.E.M. Anscombe. Blackwell.

Wittgenstein, L. (1969). *On Certainty*. Trans. D. Paul & G.E.M. Anscombe. Blackwell.

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A.K., & Fries, P. (2007). Modulation of neuronal interaction through neuronal synchronization. *Science* **316**, 1609–1612.

Wood, D., Bruner, J.S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* **17**, 89–100.

Wynn, T. (2002). Archaeology and cognitive evolution. *Behavioral and Brain Sciences* **25**, 389–438.

Yuste, R., MacLean, J.N., Smith, J. & Lansner, A. (2005). The cortex as a central pattern generator. *Nature Reviews Neuroscience* **6**, 477–483.

Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. MIT Press.

Zeigler, H.P. (1993). Visuomotor mechanisms. In H.P. Zeigler & H.-J. Bischof (eds.), *Vision, Brain, and Behavior in Birds*, MIT Press, pp. 243–264.

This page intentionally left blank

Index

Note: page numbers in *italics* refer to Figures.