

# *Frontal Assault*



*Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be like to have the slightest idea... So much for the philosophy of consciousness. (Fodor 1992)*

One of my clearest memories of early childhood finds me sitting alone in my bedroom at twilight when I was about five, pondering a curious family of questions. Why does the universe exist? What if it didn't? What would be left over if it stopped existing? Wouldn't *something* still exist? What colour would it be? Even now, the questions elicit the same peculiar twisting sensation from my stomach. And now, as then, I find the basic mystery of why anything exists the most unfathomable of all.

Probably it was fortunate for the sake of a happy childhood that the gravity of the second most difficult question didn't impress itself on me until years later: how is it that it feel like it does to be me (or anyone else) if that feeling is done with nothing more than a stringy ball of nerve fibres, glia, and other organic structures made mostly of water? How can it *possibly* feel like anything?

It is unusual that I agree significantly with Fodor, especially in print. But I think his pessimistic prognosis that an unassailable conceptual roadblock separates consciousness and the material world is *almost* right. After all, conscious experience seems entirely different from mundane features of material objects like mass or colour or fuzziness. Such attributes generate comparatively few deep metaphysical quandaries. While the road from fundamental physics and chemistry to their macroscopic appearance is, in practice, a very long and complex one which has yet to be explored in every detail, in principle nothing stands in the way of a complete low level account of, say, the particular bulk, reddish-yellow hue, and deceptively appealing fuzziness of a ripe peach. No one puzzles over the question 'how can a material object like a peach possibly be *fuzzy*?'. But on the face of it, at least, the characteristic features of phenomenal experience stand in nothing at all like the same sort of relationship to low level physics and chemistry that renders peach fuzz so metaphysically unintriguing.

While a fuzzy peach looks to be a straightforward example of material substances arranged in the right way, it isn't easy to see, even in principle, how Nature could have built something like a *conscious mind* out of matter. Taking as a working hypothesis the notion that fundamental particles of matter lack phenomenal experience entirely, it isn't easy to see how any set of such parti-

cles—however they might be organised, energised, shaken, or stirred—could somehow come to be conscious. Nevertheless, each day we find ourselves surrounded by creatures who seemingly enjoy conscious experience without incorporating within themselves any extra ‘secret ingredients’ beyond myriad ordinary material particles. (At least, contemporary science has yet to uncover any such extra ingredients.)

After years of flirting with extravagant but, ultimately, explanatorily bankrupt nonphysical candidates for a solution to this apparent mystery, some significant conclusions have finally begun to emerge which lead me to think the instantiation of conscious minds by purely material structures really may be fathomable after all. This book explores some of those conclusions and the lines of thought behind them, constructing a story within which conscious experience and the material world may plausibly be unified.

Taking the position that understanding the puzzles of conscious experience cannot proceed without some grasp of who or what the *subject* of conscious experience might be, I ultimately defend the view that the conscious subject—the *I*—is a materially instantiated structure of dynamically changing information. On this account, phenomenal experience is ‘what it is like to be’ such a changing structure. While it may be true that it is not identical to any material thing, I suggest that a conscious mind may nonetheless be *implemented* by matter. For now, I will say little more about this central unifying idea, apart from noting that information is understood here in a wholly naturalistic, precise, and objective sense and quantified within a formal framework provided by the field of *algorithmic information theory*. Several chapters of preliminaries—concerning information and representation, problems of perspective, functionalism, supervenience, and mental states—must be completed before much sense can be read into this view of the conscious subject, which doesn’t receive full treatment until the latter half of Chapter 6. Drawing on resources from traditional philosophy, computer science, mathematics, physics, and neuroscience, the job of developing and exploring this account amounts to a full frontal assault on the clutch of confusions which philosophers collectively refer to as ‘the mind/body problem’.

Many themes underlie this campaign (some of which receive more attention below in section 2), but of paramount concern is my aim of embracing the rich ‘ineffable feel’ of the phenomenal world while still operating within the constraints set by the laws of physics. The method I advocate here takes the laws of physics as given and then examines what sort of picture of consciousness and of cognition might be built up within some framework consistent with those laws. In other words, we *assume* for the sake of enquiry that the creatures of interest, such as humans, are instantiated in a purely physical world<sup>1</sup>—and then attempt to construct a picture of how (or whether) those creatures could possibly be conscious in such a world. The main task then is to evaluate whether such a

picture accounts fully for the way our conscious experience really does seem to us in the actual world or whether some significant feature has been left behind. Only if one or more significant features cannot be accommodated should we then consider rejecting both the framework itself and perhaps even the original assumption that consciousness exists in a purely physical world.

This ‘world first’, or ‘physics first’, approach, which begins with the physical and attempts to work toward a phenomenal goal, contrasts starkly with methods traditionally dear to the hearts of philosophers. More often, the preferred starting place includes some set of features of phenomenal experience thought to be ‘manifestly clear’ to a subject’s introspection, and the goal is to reconcile those features with what we observe in the physical world.<sup>2</sup> But frequently these innocuous clear beginnings lead on in very short order to conclusions clearly conflicting with what a physicalist picture can deliver. It shouldn’t be too surprising, however, if starting only with what *seems* manifestly clear from heady introspection sometimes leads us to places which are in fact impossible to accommodate with actual mere physics.

For instance, it is not unusual to include in an initial flotilla of assumptions the notion that human minds can generate arbitrarily many well-formed English sentences, or that they can perform addition over the full domain of natural numbers, or that their conception of a valid proof can be infinitely extended. Likewise, it is not uncommon to suppose that human minds are capable, in principle, of perfect rationality—in the sense of being able to find all deductively valid consequences of a set of beliefs—or even that they are capable of determining whether any particular proposition (say, in first order logic) follows from a given set of assumptions. Despite the fact that not one of the above assumptions finds support from a single scrap of empirical evidence, even many naturalistically minded philosophers happily invite them on board. A popular justification for doing so, it seems, boils down to a belief that these ‘in principle’ ideals are what *really* need explaining, and that actual world deviations from the ideal are little more significant than the noisy boundary conditions which always blur experimental measurements away from the ideal predictions of a good theory. As it happens, however, convincing reasons suggest all the above assumptions are false for any cognizer physically instantiated in the real world—even though the failure of each is consistent with the *appearance*, from the vantage point of the cognizer concerned, that it is true.

Many of the mind/body confusions which this book attempts to clear away find their roots in just such lavish assumptions about what *seems* clear from introspection. For progress to be made on the puzzle Fodor thinks no one should even hope to understand, it is helpful for philosophers to adopt a little more modesty about their own capabilities, acknowledging that—in the absence of some *argument* to the contrary, and mere appearances aside—the default as-

sumption should be that we humans are as subject to the constraints of the laws of physics as any helium atom in the core of a star or any fuzzy peach poised to fall from a tree. This book, I hope, marks significant moves in that direction.

## 1. STRUCTURE AND OVERVIEW

In view of its interdisciplinary nature, the text has been organised in a way which I hope will accommodate readers with a wide variety of backgrounds and personal interests in the material. Below, I outline my attempts to package far-flung topics in an accessible way, describe the book's anticipated audiences, and briefly preview each chapter.

### *1.1 Pathways Through the Book*

Because perhaps only a minority of readers whose interests happen to coincide with my own will want to pursue the arguments set out here linearly from start to finish, I have structured the book with a view to making other reading strategies as painless as possible. Especially technical remarks are typically flagged in advance, and in each such case I suggest an alternative route through the text which will allow readers with less interest in details to skip ahead without missing central points. Around two hundred cross-references to particular chapters, sections, or page numbers, together with an extensive index, should help those taking a nonlinear path through the book to locate other points supporting key lines of thought.<sup>3</sup>

With a few exceptions, the discussions require little in the way of specialist background knowledge, but by no means do I intend the book as a comprehensive introduction to any of the subdisciplines which feature in it, nor to the mind/body problem itself. Although in places the text still reveals roots in a doctoral dissertation, the overhead dedicated to one dissertation favourite, reviews of existing literature, is substantially reduced—particularly in areas (such as representation) where fresh perspectives are in the offing, perspectives which may fit only awkwardly or not at all into the categories provided by that literature. Likewise, I aim mainly to present views positively, largely dispensing with cautionary lists detailing what they *are not* and contrasting them with the many similar cousins for which they might be mistaken. And while I hope the broad collection of about five hundred references provides a helpful start for readers following up particular threads, it remains far from complete. I have tried to avoid the 'my bibliography is bigger than your bibliography' syndrome—the urge to cite the kitchen sink—which seems to be spreading rampantly through populations of undergraduates and seasoned researchers alike (perhaps under the influence of easy to use electronic abstract databases).

As for exceptions to general accessibility, I do assume familiarity with basic logical connectives, plus a few mathematical and set theoretic symbols, and I assume experience with the philosophical notions of *a priori*, *a posteriori*, intension, extension, possible worlds, and the like. Prior acquaintance with Turing Machines is helpful. The ‘worst’ exception occurs in Chapter 7; the summary notes on the quantum formalism beginning on page 145, as well as some subsequent sections, will be most useful for those with at least a passing familiarity with linear algebra. However, as always, those portions may safely be skipped by readers preferring to bypass the particulars, and pointers on where to pick up again are of course included.

Overall, I hope the book will be enjoyable for most people interested in materialist cognitive science and the challenges of understanding consciousness, from advanced undergraduates to senior researchers and those whom publishers’ marketing departments sometimes dub ‘the motivated lay reader’. It may find a place in graduate seminars or as a supplementary text in philosophy, cognitive science, artificial intelligence, or artificial life.

### *1.2 Chapters Summary*

The second chapter is a warm-up exercise. Taking up Dennett’s recent challenge to defend the philosophical relevance of zombies without begging important questions about their capabilities (or lack thereof), the chapter includes an architecturally explicit zombie construction and applies it to motivate later explorations not of how conscious subjects *behave* but of what internal processes bring about that behaviour. Main outcome: nonconscious zombies with external behaviour indistinguishable from that of normal subjects are logically possible, so conscious experience requires more than the right external behaviour.

Those later explorations of internal processes depend largely on the precise and objective notions of information and of representation which take centre stage in Chapter 3. Introducing a purely physical view of information based on Gregory Chaitin’s version of algorithmic information theory, the chapter describes representation in the general case with a formal measure of *mutual information content* between two physical objects. (Note that this appeal is to a concept of information which differs from Shannon’s, used by Dretske 1981.) Along the way, I outline the modern and easily understood information theoretic version of Gödel’s incompleteness results, questioning the foundations of claims occasionally made about incompleteness and its bearing on philosophy of mind. Main outcome: representation is objectively linked to the physical world.

The next chapter shows why, far from underwriting a convincing argument against physicalism, the curiosities of Jackson’s example of Mary, the colour-deprived neuroscientist, arise naturally within a wholly physicalist setting. Tricky puzzles of perspective, or points of view, first arise here—to return in

Chapter 6—and the relationship between logic and the physically instantiated cognizers using it merits brief remarks. Main outcome: the same information may be physically instantiated in many distinct ways and with disparate ramifications for conscious experience; Mary’s ignorance before seeing red for herself is not a matter of information, but of state.

Chapter 5 outlines a new formal framework for understanding functional systems. Problems of mathematical triviality threaten traditional approaches to functionalism based on correlation or correspondence, while teleofunctionalism requires links to facts historically removed from the system in question, rendering it more suitable for questions like ‘why is this component here?’ than those like ‘how does this system work?’. This chapter’s objective method of functional decomposition, using a formal measure of process complexity called *functional logical depth* (inspired by work of Chaitin and Charles Bennett), circumvents both difficulties. Main outcome: the revised functionalism is now objective and better disposed for explanatory work.

Within a context featuring arguments about supervenience, perspectives, and mental states, Chapter 6 outlines the first components of a theory of consciousness based on the *self model*, a materially instantiated dynamic data structure which emerges as a promising candidate for the seat of phenomenal experience. On this view, one motivated by the need for a conceptual link between consciousness and the material world yet grounded empirically and thus falsifiable, conscious experience is ‘what it is like to be’ a particular kind of changing data structure. The self is not identical to any of those physical components underlying it, yet it is implemented by them; *I am a self model*. Main outcome: taking the concept with an appropriate intension, consciousness supervenes logically on the physical world, with the self model as link.

Taking a brief side trip to debunk a competing class of theories, the so-called ‘quantum theories of consciousness’, the next chapter describes the work of Roland Omnès and the mechanisms of *interactive decoherence* which, automatically and extremely efficiently, eliminate the need for a conscious observer in quantum mechanics and guarantee that special quantum effects are virtually nonexistent at the levels of description where they are sometimes imagined to feature in the human brain. As a bonus, interactive decoherence accounts for the automatic emergence of apparently deterministic quasi-classical reality from a quantum substrate. Main outcome: quantum physics doesn’t need consciousness, and consciousness doesn’t need quantum physics.

Where Chapter 6 focused mainly on the conceptual territory between the self model and supervening consciousness, Chapter 8 sets out in the opposite direction, from the self model down toward the sorts of lower level materials—neural tissue, in the case of humans and other terrestrial life forms—which may implement such data structures. After tidying the information theoretic description

of self models given in Chapter 6 and introducing basic tools from neuroscience, the chapter shows how one particular research programme, Stephen Grossberg's adaptive resonance theory, bears especially on the task of implementing self models in real neural systems. Main outcome: the right neural systems can do the representational work which self models require, and real organisms may plausibly have evolved so as to implement them.

Turning from cognitive models to the mathematical presuppositions underlying them, Chapter 9 examines the relationship between models and properties of the physical systems being modelled. In particular, the chapter examines recent work by Hava Siegelmann and Eduardo Sontag suggesting that analogue models (based on the real numbers) may display computational capabilities which exceed those of digital models (based on the rational numbers). To the extent that the two sorts of models differ in their capabilities, an interesting question then arises as to whether one or the other makes for a better match with reality. Main outcome: contrary to popular dogma, the choice of number systems does make a difference for models of cognitive systems, but the significance of that difference for the real world has yet to be established.

Finally, Chapter 10 recaps some of the text's central themes and reflects on the broader significance of the mind/body problem for understanding who we are, as individuals and as a civilisation. It finishes with some remaining open questions and possible directions for future research.

Readers who find my attempts in each of these chapters to situate the discussion within an overall picture a little too obscure might also want to skim through that last chapter's section 2, starting on page 243, which, unlike the above summary, surveys the principal developments of each discussion on the assumption that readers will already have acquired some familiarity with them.

## 2. UNDERLYING THEMES AND METHODS

Although the book treats many distinct topics, all are united both in the sense that they relate to some aspect of the mind/body problem and in terms of underlying themes. Below I begin with a theme I specifically attempt to sidestep and move on to two more positive ones. Readers less interested in comparatively dull methodological issues should skip ahead to section 3 below.

### *2.1 Dynamics and Computation*

One fracas I hope to avoid is the war raging between advocates of dynamical and computational methods in cognitive science. At a recent workshop marking the opening of Sussex University's Centre for Computational Neuroscience and Robotics, for example, I was astonished at the number of participants who appeared quite abruptly to have rediscovered dynamical systems and were vigor-

ously promoting a dynamical approach to artificial life and related fields as the greatest new advance—perhaps ‘revelation’ would be more fitting—in decades. Worse than the hype<sup>4</sup> was the impression that one should be exclusively either dynamicist or computationalist and that for the sake of scientific purity the two must not be mixed—a notion asserted with characteristic vehemence by Tim Smithers, the well known ‘non-representationalist’ roboticist from the University of the Basque Country. I hardly dared raise a voice for a hybrid view for fear of witnessing my own public stoning!

Typically, allegiances to a particular approach run deep, often it seems to such an extent that what appears at a glance a straightforward observation supporting one or the other position becomes utterly invisible to those favouring the opposing viewpoint; all too frequently, arguments are simply ignored or contradicted rather than rebutted. I hope in this book to challenge both sides while incensing neither. For my part, while generally inclined more in the direction of a dynamical approach, I believe tools of *both* types play a valuable rôle in cognitive science: both dynamics and a computationally defined variety of representation feature crucially in my own view. In Chapter 3, I outline a method of quantifying information content (and representation) which, while itself wholeheartedly computational, applies equally well to all physical systems, whether interpreted as computational or dynamical entities. Later, in Chapter 6, I argue against the typical computationalism-friendly notion of an instantaneous mental state in favour of an inherently temporal replacement, advancing a view of consciousness itself which, while described in ‘computational’ information theoretic terms, subsequently receives a dynamical account of neural instantiation in Chapter 8. (For readers harbouring strong feelings on the debate and an irrepressible urge to categorise this book in the language of it, the approach I adopt here might be verbosely labelled, with tongue in cheek, as ‘computationally described, dynamical semi-representationalism’.)

*Contra* Smithers, I emphatically *do not* believe that one should decide in advance to adopt one framework or the other, on pain of tainted science. Whether a good explanation of observed phenomena should be dynamical, computational, or hybrid in nature is a question properly evaluated in light of empirical evidence and the resulting matches between candidate theories and reality; it is not, to my mind, a matter of evaluating empirical evidence in light of a pre-existing conviction that explanation is to be found in one and only one particular form. Likewise, *contra* van Gelder (see note 4 above), I find very little *ontological* mileage in the methodological distinction between dynamical and computational approaches to cognitive science. As far as I can see, whether we speak in computational language, dynamical language, or little green men language, we still talk about the *same actual world*. (‘World first’, not ‘words first’!) In the terminology of Chapter 6, the ‘things’ talked about supervene

logically on the physical world. Depending on the needs and aims of any given situation, we might find it useful sometimes to describe cognitive processes computationally, while at other times dynamical or hybrid descriptions will prove most helpful. There may be a *de facto* rough cultural division between cognitive scientists who generally prefer computational tools and those who generally prefer dynamical tools, just as there is a *de facto* rough division between those who prefer reverse Polish notation calculators and those who prefer algebraic calculators, but such a division needn't necessarily reflect ontologically significant facts about the real world of cognizers which they are studying.

### *2.2 Priorities and the Naturalistic Urge*

Closely related to the business of choosing appropriate languages of explanation are questions about the goals of research itself and the choices of tools which those goals motivate. One opinion popular in the United Kingdom, at least, counts the task of furthering the cause of a particular discipline or even department as, if not quite an end in itself, at least a highly significant component of academic research. On this view, a central aim of the philosophy researcher should be to advance the understanding of philosophy for philosophy's sake; the general field of philosophy is primary, while the particular issues to be explored take a back seat. A diametrically opposed tack shapes my own research in general and this book in particular.

My main interest lies not with furthering a particular discipline (or, even less, a department), but with examining fascinating questions, whatever disciplines may count as their own those questions or their solutions. I make no apologies for importing methods or conclusions from outwith traditional philosophy, particularly from scientific fields, in the course of exploring the mind/body problem.<sup>5</sup> Although many philosophical colleagues in the United Kingdom are quick to dismiss my own work as “science, not philosophy!”—as though the two were mutually exclusive—I am only too happy to recruit for ‘philosophical’ purposes the tools of other fields. (Were I to encounter tomorrow some good reasons for believing the art of pottery held the key to understanding problems of consciousness, I would learn to make pots!) In the spirit of thinkers like Kitcher (1992) and those featured in Kornblith (1994), I take philosophy and the natural sciences as forming a continuum. The eloquent words of E.W. Hobson, Sadleirian Professor of Pure Mathematics and Fellow of Christ's College in Cambridge, speaking at the conclusion of his 1921-22 Gifford Lectures at the University of Aberdeen, express the idea succinctly:

If we were in possession of, and able to grasp, a unified view of the Universe, in which all the elements of existence and valuation were completely synthesized...we should not require to mark out frontiers between Science and Philosophy or Theology... The untrammelled freedom which must be allowed to workers in all depart-

ments of the great cultural work of humanity...should not...involve the erection of rigid impassable barriers which shall mark off domains which hold no communication with one another. On the contrary, workers in one department will often receive the most valuable enlightenment, and most important suggestions, from quarters outside their own special line. (Hobson 1923, p. 501)

Borrowing the contemporary words of Sussex researcher Inman Harvey (personal communication), I also agree wholeheartedly with the notion of “doing philosophy of mind with a screwdriver”—testing philosophical ideas with robotics and artificial life in real world laboratories. In an article entitled ‘Artificial Life as Philosophy’, Dennett (1994) advocates a similar view.

But despite a distinguished tradition of positive and mutually beneficial interaction between philosophy and the sciences, philosophy of mind is, perhaps more than at any time in its history, feeling the heat from scientific areas as those fields ‘encroach upon’ the study of questions still considered by many to be philosophers’ territory. Quite justifiably, methods and strategies from areas like quantum physics, chaos theory, neuroscience, and artificial life have been stirring up the field, eliciting responses from the philosophers’ camp ranging from the sort of dismissiveness recalled above to enthusiastic endorsements of method  $x$  as that which will at last naturalise the study of mind. Although my own tendencies draw me firmly in the naturalistic direction, explicitly arguing the case for the relevance of scientific fields to ‘philosophical’ questions really is not my battle. I hope others will simply evaluate my judgement in choosing the tools I have in light of the lines of argument they support, rather than under the shadow of some pre-existing prejudice as to how best to secure the borders of philosophy against infiltration.

### *2.3 Description Complementarity and Puzzles of Perspective*

After the central goal mentioned at the start of the chapter, that of embracing the ‘ineffable feel’ of phenomenal experience while simultaneously heeding the constraints of physical law, the next most significant thread running through this book concerns the relationships between different levels of description<sup>6</sup> of the same cognitive system and between first and third person perspectives on cognitive systems. I believe both relationships bear crucially on the project of understanding how (or whether) matter can implement a mind.

With respect to the first, a good many problems in the history of philosophy originate, in my opinion, with a failure to appreciate the complementarity between alternative descriptions of the very same thing—a failure which, at times, borders on stubborn-minded silliness. Throughout the explorations of the topic which appear here, I have in mind a background context of working hypotheses inspired by the likes of Davidson (1973) and Hellman and Thompson (1975); I opt for what the latter call ‘ontological determination’—the physical is all there

is, and everything that happens physically is governed entirely by the low level laws of physics—coupled with ‘explanatory anti-reductionism’. In other words, nothing ever happens which is not, at the lowest level, entirely a result of the laws of physics;<sup>7</sup> yet, in giving intelligible *explanations* of processes, we may well have to rely on entities constructed at a higher level of description commensurate with that at which we describe the processes themselves.

For instance, it would be no explanation of how a clock keeps time with its hour and minute hands to describe the interactions, in accordance with the laws of physics, of every single subatomic particle within it. A good explanation would describe instead the interactions of cogs and pendulums or transistors and quartz crystals (depending on the clock) and the relationship of those interactions to the movements of the hands. Yet nothing ever happens to (or between) cogs, pendulums, transistors, quartz crystals, or arms of a clock which doesn’t follow straightforwardly from the lowest level laws of physics. The approach to explanation favoured in this book embraces both observations together. The view contrasts starkly with those of Durkheim (1938) or Radcliffe-Brown (1952), for instance, who maintain that good explanations require not only an appropriate level of description, but a new independent level of *things* whose causal properties, significantly, *do not* follow from those of their constituent parts.<sup>8</sup> In the guise of a discussion of evolution and Conway’s Game of Life (Poundstone 1985), Dennett (1995c, pp. 166-75) offers the tidiest look at levels of description I have encountered to date.

Interest in the relationships between alternative descriptions of the same thing doesn’t end with different levels; a special case of complementary descriptions exists in the distinction between first and third person *perspectives* on cognitive systems. Just as I believe there is a common failure to appreciate the relationships between levels of description, so, too, is there a common failure to appreciate the factors underlying the truism that reasoning *about* a system as a third person differs greatly from *being* that system in the first person. If, for instance, I cannot come to know ‘what it is like’ for *me* to be in a particular state until I have actually been in that state or one relevantly similar to it, why should it be at all puzzling that I cannot come to know what it is like for someone else to be in such a state? And is there any reason to think I *should* be able to grasp what it is like for me to be in a particular state before having actually been in such a state? Such perspectival issues figure in a wide range of questions about the relationship between mind and the physical world. My direct quarrel with a standard view that problems of perspective support a case against physicalism begins in Chapter 4 and returns in Chapter 6, but a reluctance to accept the standard view underlies much of the rest of the book.

### 3. COGNITIVE DISSONANCE

It was in one of the first philosophy texts I encountered as an undergraduate, Richard Taylor's (1963) *Metaphysics*, that I vividly recall reading the tale of a poor fellow called Osmo, who perishes while trying to escape a future foretold for him in a special book. With perfect accuracy, the mysterious volume describes every event in Osmo's entire life, from his birth through to the present and on into the future and his eventual death. Like the protagonist in a Greek tragedy, Osmo cannot escape his Fate: eerily, whatever the book foretells *always* comes to pass, and try as he might, his every failed struggle to forge a new and different future for himself only underscores the book's apparent infallibility.

Taylor's point in articulating the story—apart, perhaps, from terrorising impressionable young philosophers—is to suggest that we all should view the future with the eyes of a fatalist. Fatalism should appeal, Taylor argues, for purely logical reasons: since some complete description of all my life's future events is true *right now*, and since that description could *already* have been written down in a book (say, by an omniscient and prolific author), I lack any sort of freedom to change it, and I ought just to stop worrying about it! Despite Taylor's protests to the contrary, the argument for fatalism seems a textbook example of modal fallacy. But nevertheless, I think I can almost grasp what might have led him to suggest the line of thought, despite his fluency with the logic of modal operators. Although most philosophers reject Taylor's reasoning, the years have not dulled my impression that something remains deeply disturbing about the fantasy scenario he describes; a book accurately foretelling my every experience—perhaps even my every thought—for the rest of my life truly would be a frightening prospect.

That peculiar discomfort, I believe, bears a possibly illuminating likeness to the distaste many express for the idea that human cognition and consciousness might one day be explained within a purely materialist framework. I am *not* suggesting that those unhappy about such a possibility are confused about modal logic! Rather, although the analogy is imperfect with respect to logical structure, there seems to me an appealing kind of symmetry between the two cases. On one hand, it is disturbing to think that some book could in principle expose, for all to see, every single event in my entire life—even though I know perfectly well that the logical possibility of such a volume, in and of itself, in no way diminishes my freedom in bringing about the events described within it.<sup>9</sup> There might be other reasons why I lack freedom, but the possibility of such a book is certainly not one of them. And on the other hand, some may likewise find it disturbing that a complete materialist theory of consciousness could in principle expose, for all to see, the underlying physical foundation of every sin-

gle pain or taste or visual impression (or hope or lust or intuition) throughout a subject's entire life. This might seem disturbing *even though* such a prospect would in no way diminish the painfulness of the pain or the lustfulness of the lust; that rich phenomenal experience might have a physical explanation would render it no less rich phenomenal experience.

A usually unarticulated further worry might grow from the notion that anything which admits of a physical explanation is no different in principle from any other physical thing. The possibility of a book explaining my entire conscious existence in the language of physics threatens to reduce that experience itself to nothing more than *mere physics*, siphoning off its *value* down to the level of some least common denominator appropriate for other physical things like bricks and globules of sludge. Probably it is only natural to feel some discomfort at the idea that the very sciences which our own ingenuity created could drain away our fundamental value in this way. On this view, perhaps our value can only be preserved by finding it a nonphysical refuge categorically separate from (and impossible to unify with) the merely mechanical transactions of bricks and balls of sludge.

Needless to say, these are worries I do not share. On the contrary, I feel that such a naturalistic physical explanation could only *add* to the sheer marvellousness of phenomenal experience. There can be little doubt that it would be amazing if it turned out instead that the subject of my phenomenal experience was really an independent, immaterial, ethereal spirit of some sort—that *I* was such a spirit and my consciousness one of that spirit's properties. But how much more truly amazing it would be to discover that the conscious 'I' is instantiated purely physically—that somehow, despite having nothing but that stringy ball of nerve fibres and other organic matter with which to do it, I still manage to enjoy my full remarkable repertoire of rich conscious experience! That matter simply organised in the right pattern and changing in the right ways could actually instantiate *me*, with all my vivid phenomenal experience intact: *that* is a marvel worthy of the name. And, indeed, if what really counts is its rôle in enabling our conscious lives, why should a *pattern* be of any less value than an immaterial spirit?

In the next nine chapters, I set out what I believe are some useful steps toward understanding the sorts of links between mind and matter which might make such a view attractive in its own right and which may allay Osmo-style discomfort. Do I believe I have given a rigorous, complete, and definitive solution to the mind/body problem(s)? Of course not! But I do think a 'solution'—or, rather, a set of solutions to a cluster of related problems—will eventually be found to share the general form, and perhaps some of the details, of what I outline here. Many times in these pages I deliberately reach far out on a limb while constructing some view or other. In so doing, I aim to lay out a sort of tree of possibilities; not all of its branches will ultimately bear fruit, I am sure. But I'm

convinced that while some limbs will eventually need cutting out and others will benefit from considerable reshaping, the central thematic trunk is healthy and planted just about where it ought to be. I hope that laying out the tree as I have will provide new opportunities for progress through the process of evaluating and snipping back those bits which don't belong, strengthening those which do, and shaping this nascent theory into something robust and comprehensive.

### NOTES

<sup>1</sup> Nowhere do I argue positively for any particular rendition of material monism; I find it hard to grasp what a good argument for material monism would even look like.

<sup>2</sup> Philosophers, understandably concerned to be clear about what it is they're trying to explain before setting out to explain it, frequently adopt the convenient assumption that our first hand, direct experience of consciousness suffices to fix the concept appropriately well. But to paraphrase an example due to Aaron Sloman, the assumption is as unjustified as the claim that even before Einstein's analysis of the concept of simultaneity, we really all knew what it was anyway just through our first hand experience. As for the case of simultaneity, Sloman suggests, it may be that only after constructing some good theories, capable of supporting coherent concepts, will we be able to grasp properly what it was we were trying to explain in the first place!

<sup>3</sup> Since endnotes are easily found at the end of each chapter, when cross-referencing them I usually mention the page number for the discussion which is endnoted rather than giving the page number for the endnote text itself.

<sup>4</sup> Not *all* the excitement is hype. While I am dubious about his tendency to overstate the *ontological* significance of what amount to different dynamical (or non-dynamical) *descriptions* of physical entities—see section 3 of Chapter 9—Tim van Gelder (1995a, b) offers a sober but provocative analysis of each approach. In a forthcoming target article for *Behavioral and Brain Sciences*, he succinctly maps out distinctions between the computational hypothesis—roughly, the view that cognizers are digital computers—and the dynamical hypothesis—roughly, the view that cognizers are dynamical systems.

<sup>5</sup> Very often it is expedient to import science into philosophical discourse—nothing clinches an existence proof like empirical evidence, for instance!—but some of the *really* difficult problems of philosophy might well be those which either cannot be naturalised or which would be very difficult to naturalise. (Ethics comes to mind as a possible example.) I have the utmost respect for these sorts of areas, but my temperament generally tempts me more in the direction of those where some 'easier' progress can be made.

<sup>6</sup> Although I usually refer to 'levels', I mean also to include *parallel* or otherwise complementary descriptions which may not exist in an hierarchical relationship with one another.

<sup>7</sup> In the language of Chapter 6, everything *supervenes* on microphysics; see section 1 of that chapter in particular.

<sup>8</sup> On the last especially, also compare the emergentists such as Alexander (1920), Broad (1925), or Pepper (1926); see McLaughlin (1992) for recent discussion.

<sup>9</sup> Specifically, the purported argument to the contrary requires shifting the scope of a necessity operator from a whole conditional to just the consequent of the conditional. But although it is necessarily true that if such a book correctly proclaims that I will perform action *x*, then I will perform action *x*, from this it simply does not follow that such a correct proclamation entails that I necessarily will perform action *x*.

## CHAPTER TWO

# *Zombies and Their Look-Alikes*



*It is an embarrassment to our discipline that what is widely regarded among philosophers as a major theoretical controversy should come down to whether or not...philosophers' zombies...are possible/conceivable. (Dennett 1995b, p. 325)*

Zombies of analytic philosophy, unlike the voodoo victims of Haitian folklore, are hypothetical creatures entirely bereft of conscious experience who nonetheless *behave* indistinguishably from the rest of us. Philosophers' zombies walk and talk as if they're conscious, they appear to wake up in the morning, and over breakfast they even speculate on the meaning of dreams they claim to have had. They don't *realise* they're zombies, of course—no feeling of peculiarity spoils the pristine emptiness of their barren phenomenological landscapes—and an enterprising philosopher engaging one in conversation about the topic might well hear all manner of insightful commentary about what the notion of zombies reveals about philosophy of mind.

Daniel Dennett says zombies don't exist. Most people, no doubt, would agree. But more importantly, he suggests they are logically impossible: misguided philosophers who claim they are conceivable merely fail to imagine the *full* repertoire of zombie behaviours. All too often, philosophers define zombies as above and then proceed to argue for some behavioural clue or other which would give them away. But of course there can't be any such clue. "The philosophical tradition of zombies would die overnight", Dennett says (1995b, p. 325; Dennett 1995a is similar), "if philosophers ceased to mis-imagine them". It is unusual that I disagree significantly with Dennett, especially in print (Mulhauser 1997a). But I think his pessimistic prognosis can't be quite right. Dennett's challenge offers a tailor-made warm-up exercise for the rest of this book:

Show me, please, why the zombie hypothesis deserves to be taken seriously, and I will apologize handsomely for having ridiculed those who think so... If the philosophical concept of zombies is so important, so useful, some philosopher ought to be able to say why in non-question begging terms. I'll be curious to see if anybody can mount such a defence, but I won't be holding my breath. (Dennett 1995b, p. 326)